Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 05, May-2019

A study on T-Closeness over K-anonymization Technique for Privacy Preserving in Big Data

Kajol Patel[1]
[1]Lecturar, Department of Computer Enginerering,
Parul Institute of Engineering & Technology- (Diploma Studies),
Vadodaara, Gujarat.

***Abstract*— Big data has continues a rising in a word of data analytics. It contains very large set and complex data structure. In big data there will be three methods to protect data: k-anonymization, L-diversity and T-closeness. Here, we discuss T-closeness method which requires the distribution of a sensitive attribute in any equivalence class is close to the distribution of attribute in the overall table.**

Keywords*— Data privacy, privacy preserving methods, T-closeness, k- anonymization

I. INTRODUCTION

With the rapid development of computer technology, the importance of data sharing emerge gradually which based on scientific research, business application and knowledge discovery. However, shared data contains some individual sensitive information such as medical records, therefore privacy is at risk of experts and scholars and the information owner. The k-anonymization methods and the l-diversity methods are the common privacy protection methods, and they are simple and practical. Compared with the former two methods, t- closeness model enhanced ability to prevent privacy leaks, and has been widely used and research [5].

Data privacy has always been important. Because most of our data becomes digitized, and we share more information online, data privacy is taking on greater importance.

A single company may possess the personal information of millions of customers data that it needs to keep private so that customer's identities stay as safe and protected as possible, and the company reputation remains immaculate, but data privacy is not just a business concern.

- What is data privacy?

Data privacy relates to how a piece of data should be handling based on its relative importance. For illustration you likely would not mind sharing your name with a stranger in the process of introducing yourself, but there is other information you would not share, in any case not until you become more aware with that person. For example, open a new bank account and you will might be asked to share a incredible amount of personal information, well beyond your name.

In the digital era, we typically apply the concept of data privacy to important personal information and also known as personally identifiable information and personal health information. This can include Social Security numbers, health and medical records, financial data, including bank account and credit card numbers, and even basic, but still perceptive, information, such as full names, addresses and birthdates.

- Why is data privacy important?

When data that should be kept private gets in the wrong hands, bad things can happen. A data breach at a government

Agency can, for example, put top secret information in the hands of an enemy state. A breach at a corporation can put proprietary data in the hands of a competitor. A breach at a school could put students' PII in the hands of criminals who could commit identity theft. A breach at a hospital or doctor's office can put PHI in the hands of those who might misuse it.

Since data privacy is such a prevalent issue, many government organizations and corporations spend millions of dollars each year to help protect their data—which could include your PII—from exposure. The average consumer probably doesn't have that kind of money to spend. But there are inexpensive steps you can take to help protect your data. Here are a few suggestions:

1. At home, use a mail slot or locking mailbox, so that thieves can't steal your mail.

2. Before discarding, shred documents, including receipts and bank and credit card statements, that contain personal information.

3. Make sure to secure your home Wi-Fi network and other devices so that criminals can't "eavesdrop" on your online activity.

4. Don't automatically provide your Social Security number just because someone asks for it. Determine if they really need it and, if so, ask how they'll help protect it.

5. Use strong, unique passwords for all of your online accounts.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 05, May-2019

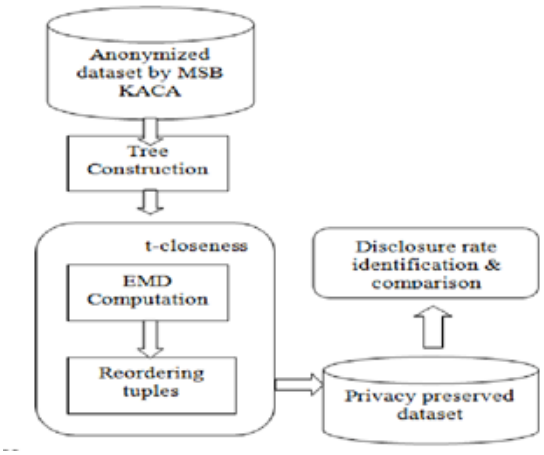

Fig. 1 Utility and privacy using T-closeness

II. METHODOLOGY

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 relationship between the quasi-identifier attributes and the sensitive attributes. The distance between the distributions is measured using Earth Mover's Distance (EMD). For a categorical attribute, EMD is used to measure the distance between the values in it according to the minimum level of generalization of these values in the domain hierarchy [11].

A. k-Anonymity

Most of the data holder including the government agencies and hospitals misunderstand that the data, e.g. medical records, will remain anonymous if the explicit details such as name, address and phone number are concealed before disclosing the rest of the records. Nevertheless, re-identification of individual by linking the data with other published data, e.g. voter's list, will result in loss of anonymity. Though adding noise to the dataset such as false values and scrambling might provide anonymity, but this will give inaccurate statistical results within tuples when performing data mining and analysis. To address these problems, Samarati and Sweeney formalized a technique called k-anonymity in 1998, which use generalization and suppression methods to allow data revelation in a controlled manner while securing the value integrity of tuples. Quasi-identifiers are unique attributes that recognize an individual such as birth date and gender. A table containing these quasi-identifiers is said to meet k- anonymity if each tuple value of the quasi-identifiers recurs at least 'k' times, thus making the tuple distinguishable from one another [14] [16].

- *Principle of k-Anonymity*

If each value in a given dataset is indistinct from a minimum of (k-1) records from the same table, then the table is said to be k-anonymous. The greater value of k-value is higher than the privacy protection [2].

- *Generalization*

Generalization is a technique used to represent the attribute values in a table to make the identification of tuples less discrete. In this method, the original attribute is represented as a ground domain and the domain value increases with increasing generalization.

The limitation of this method is that there will be a need for high level of generalization when there are lesser outliers, i.e. tuples that occur less than k-times [8].

- *Suppression*

To complement k-anonymity, suppression is used with generalization. Suppression is a technique that is used to mask certain values in the quasi-identifiers [2]. The suppressed value is represented with an asterisk (*) and this can be applied to both domain and value generalization hierarchies.

- *Pros of k-Anonymity*

It preserves against identity disclosure by inhibiting the links to a dataset with less than 'k' values. This prevents the adversary from connecting a sensitive data with an external data [8] [15].

The cost of incurred in establishing this method is considerably lesser compared to the cost of another anonymity method such as cryptographic solution [5].

Algorithms of k-anonymity such as Datafly, Incognito, and Mondrian are used extensively, especially in public data. It is also mentioned that clustering is incorporated in k-anonymity to enhance privacy preservation [4].

- *Cons of k-Anonymity*

There are many limitations that have been identified in this technique, mainly attacks such as unsorted matching, complementary release, discreet and temporal attacks [8][9] [16]. Other disadvantages include this technique can cause high utility loss if it is employed in high- dimensional data and exceptional measures are needed if the released data has already undergone anonymization more than once [15]. However, in this research two of the well-known attacks on k-anonymity will be briefed below.

1. Homogeneity attack:

When there is inadequate heterogeneity in the sensitive attributes, this can generate clusters that expose information. Suppose A and B are opponents and A knows that B lives in a particular zip code and is of a particular age, and wants to know B's medical status. So, with A's insight on B, A can identify that the information matches with a number of medical records and all these records have the same medical condition (sensitive attribute), i.e. cancer. Thus, the k-anonymous table should be further sanitized by diversifying the sensitive values within the tuples that share similar values of their quasi-identifiers [8] [12].

2. BACKGROUND KNOWLEDGE ATTACK:

In this type of attack, the adversary has a known knowledge about the individual and with additional logical reasoning;

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 05, May-2019

individual's sensitive attributes can be leaked. Consider A and C are acquaintances and A would like to infer C's personal data which is found in the same patient record as B. As we knows that C is a 45-year old Asian female living at a particular zip code. Nevertheless, the record shows that C can have any of the three diseases - cancer, heart disease and viral infection. Based on A's background information that C prevents high-calorie meals and has low blood pressure, A infers that C has heart disease. Hence, k-anonymity is prone to background knowledge attack [8] [12].

B. T-Closeness

A betterment of *k-anonymity* 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. The distance between the distributions is measured using Earth Mover's Distance (EMD). For a categorical attribute, EMD is used to measure the distance between the values in it according to the minimum level of generalization of these values in the domain hierarchy [11].

- *Principle of t-Closeness*

t-closeness of an equivalence class is attained when the sensitive attribute distance in this class is not greater than the threshold, *t* with the attribute distance in the whole table. The table is acknowledged to have t-closeness if all equivalence classes have t-closeness [11].

- *Pros of t-Closeness*
  - It interrupts attribute disclosure that protects data privacy.
  - Protects against homogeneity and background knowledge attacks mentioned in *k*-anonymity.
  - It identifies the semantic closeness of attributes, a limitation of *l*-diversity.
- *Cons of t-Closeness*
  - Using Earth Mover's Distance (EMD) measure in *t*-closeness, it is hard to identify the closeness between *t*-value and the knowledge gained.
  - Necessitates that sensitive attribute spread in the equivalence class to be close to that in the overall table [9] [11].

CONCLUSION

Preservation of data privacy has transpired as a definite prerequisite in privacy preserving data. The increase in cyber crimes has caused severe risk of privacy breach. This has driven the manifestation of various anonymization techniques. This paper has discussed on these rising concerns in data security, which are converge into the healthcare domain, which poses greater chances of disclosure of personal and sensitive data. To avoid this, a range of anonymization methods applied on medical data was summarized here based on the academic literature dedicated to data use. Moreover, the scope of this research is limited to the k-anonymity technique with its extended modifications, which is t-closeness. Each of these techniques was illuminated in detail with principles and related references. The comparison of the advantages and disadvantages of these two methods were also rationalized. Overall, this research is committed to providing a brief on the existing trends of anonymization techniques orientating medical data in achieving privacy preservation under data security.

ACKNOWLEDGMENT

I would like to offer my special thanks to Mr. G B Jethava. Advice given by him has been a great help in my seminar. Many people have helped, provided direction as well as technical information and I take pleasure in acknowledging the role of people involved, directly or indirectly in the development of this seminar. I also express my gratitude to Prof. Ruchi Shrivastav Principle of the institute, PIET-DS for providing me with adequate facilities, ways and means by which I was able to complete this seminar. Last but not the least I thank all others and especially my classmates and my family members who in one way or another helped me in the successful completion of this work.

REFERENCES

[1] Keerthana rajendran, Manoj Jayabalan, Muhammad Ehsan rana," A study on k-anonymity, l- diversity and t- closeness Techniques focusing Medical Data", Springer, december 2017

[2] Aldeen, Y.A.A.S., Salleh, M. and Razzaque, M.A, "A comprehensive review on privacy preserving data Mining", SpringerPlus, december 2015

[3] Belsis, P. and Pantziou, G., "A k-anonymity privacypreserving approach in wireless medical monitoring environments" , Personal and Ubiquitous Computing, octomber 2013, pp. 61–74.

[4] Li, N., Li, T. and Venkatasubramanian, S. "Tcloseness: Privacy beyond k-anonymity and l-diversity", ICDE 2007 IEEE 23rd International Conference on Data Engineering

[5] Samarati, P. and Sweeney, L ."Protecting privacy when disclosing information: K-anonymity and its enforcement through generalization and suppression" , February 2017.

[6] Singh, A.P. and Parihar,"'A review of privacy preserving data publishing technique", International Journal of Emerging Research in Management &Technology, december 2013. pp. 32–38

[7] Sweeney, L., " Achieving k-Anonymity Privacy Protection Using Generalization And Suppression", International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, Volume 10, Issue 05,October 2002, pp. 571–588.

[8] Hussien, A.A., Hamza, N. and Hefny, H.A., "Attacks on Anonymization-Based privacy-preserving: A survey for data mining and data publishing", Journal of Information Security, volume 4, issue 2, 2013,pp.101–112.

[9] Machanavajjhala, A., Kifer, D., Gehrke, J. and Venkitasubramaniam, M. "L -diversity: privacy beyond k-anonymity", ACM Transactions on Knowledge Discovery from Data, volume 1, Issue 1, 2007.

[10] Li, N., Li, T. and Venkatasubramanian, S., "Tcloseness: Privacy beyond k-anonymity and l-diversity", ICDE IEEE 23rd International Conference on Data Engineering, 2007.

[11] Ayala-Rivera, V., Mcdonagh, P., Cerqueus, T. and Murphy, L., "A systematic comparison and evaluation of kanonymization Algorithms for practitioners", TRANSACTIONS ON DATA PRIVACY, volume 7, issue 3, pp. 337–370, 2014.

[12] Jain, P., Gyanchandani, M. and Khare, N. , "Big data privacy: A technological perspective and review", Journal of Big Data, 3(1), 2016.

[13] Casas-Roma, J., Herrera-Joancomartí, J. and Torra, V., "A survey of graph-modification techniques for privacy-preserving on networks', Artificial Intelligence Review, 2016.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)