Observations on Anonymization Based Privacy Preserving Data Publishing

[1]Dr. Mehul Barot, [2]Nirzari Patel
[1] Professor, [2] Research Scholar
[1] LDRP Institute of Technology & Research, Gujarat, India
[2] Computer Engineering Department, LDRP Institute of Technology & Research, Gujarat, India

Abstract: Anonymization is a process of hiding the information such that an unauthorized user could not gathered anything from the records, on the other hand an analyser will get necessary information[4].The word Data Privacy is related with data collection and distribution of data. Privacy issues appears in different area such as Bank sector, health care, social media data,etc.It is one of the challenging task when publishing or sharing the data between one to many sources for research purpose and data analysis[2].Many organizations also release huge micro data. It excludes an individual's direct identity marks like name, address and consist of specific information like gender, marital status, DOB, Pin-code, which can be combined with other public data to recognize a person[3]. This inference attack can be worked to obtain any sensitive information from social network platform, by that putting the privacy of a person in danger. To stop such attacks by changing micro data, anonymization is used. In this paper, we provide a additional disclosure technique for releasing information from a private table such that the identity of any individual to whom the released data refer cannot be recognized. It is based on the topic of generalization and suppression, from which stored values can be replaced with trustworthy but less specific alternatives, and of k-anonymity.

Index Terms **– privacy preserving, Data publishing, k-anonymization.**

I. INTRODUCTION

In a globally-network society, there is greater demand by society for individual-specific data, yet the widespread availability of information makes it extremely difficult to release any information about individuals without breaching privacy[1].Even when released information has no explicit identifiers, such as name and phone number, other characteristic data, such as birth date and ZIP code, often combine uniquely and can be linked to publicly available information to re-identify individuals[5].Typically, such information is stored in table format(T). Adversaries (attackers) link more than two dataset and use their background knowledge for deducing the sensitive information. Certain attributes are linked with external knowledge to identify the individual's records indirectly[2]. Anonymization techniques are used to convert the micro data D to D'[2].

A. What is the difference between the Security and Privacy?

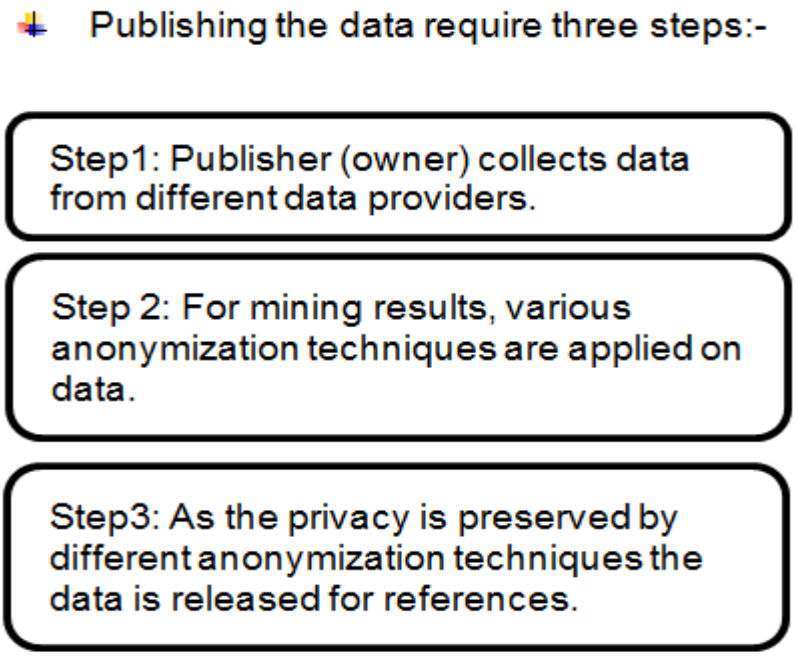

Figure 1: Three steps for publishing the data

In order to secure the data which is stored in the computer, needs to be secured by providing some data encryption, password and decryption algorithms, but the most essential thing is that only authorized person has a ability to deal with data. When privacy is considered, only the authorized person can decide the level to which information can be revealed to the outside world[6]. Objective of Privacy Preserving Data Mining (PPDM) is to publish assertion of privacy preserved dataset and preserve sensitive information in the table, so that researchers can go ahead with the proposal by uncompromising privacy of any individual.Main aim of privacy preservation is to protect oneself from being revealed to unauthorized people.

B. Challenges in privacy preserving data publishing

1) sequential data publishing causes the linking attack of published datasets and infarcts the user's sensitive information.
2) published anonymization techniques for data publishing brings down the data utility.

II. BACKGROUND THEORY AND RELATED WORK

In this section, we evaluate the existing anonymization techniques focusing on data publishing and talk about background knowledge and also problems of privacy preserving data publishing.

A. Background Knowledge

Background knowledge can be explained as the experience that already has, come across formally from the prior rules of the published datasets of various data publisher or as informally from the life experiences. An opponent could have the earlier published datasets and other publicly available datasets. These datasets could help the opponent to acquire the background knowledge for combining with the target sensitive values from the newly published datasets. Data publisher cannot define the background knowledge for the opponent. Therefore it is necessary to prepare a general framework which can deal with all background knowledge attacks[7].

B. Problems of Sequential Data Publishing

In the data publishing framework, the data publisher will publish their data on a regular basis. For example, hospital X(Table 1) publishes their data after every 3 months and user U visits the hospital X in March for the disease D. Later in June user U visits the hospital X for the same disease D. Hospital X publishes their dataset in April and later in August. Now, the user U exists in the all published datasets with the similar QI values. An opponent may use these published datasets to assume the user U and the sensitive values in 100 percent confidence. There is various works have done to handle the data publishing privacy issues. Additionally, these published works decrease the data utility to ensure the personal privacy[7].

C. Anonymization Techniques

There are various privacy preserving data publishing techniques have been published in the last many years. This is based on based on partitioning and randomization. In the partitioning method, the data values of quasi-identifiers QI (e.g., gender, age, and ZIP code) are labeled to construct an similarity class. Therefore, an individual cannot be identified with their sensitive values in the similarity class. By contrast, in a randomization anonymization techniques, the original values have been replaced by attaching some noise therefore it is difficult to point a person in a published data set. Some popular anonymization techniques, have been published for one-time data publishing for information revelation risks. K-anonymity, l-diversity, t-closeness approaches are vulnerable to the linking attack[7].

D. k-anonymity and its variants

A variant of k-anonymity known as l-diversity was introduced by Machanavajjhala et al[8].It gives privacy in some situations where k-anonymity does not, such as when there is little diversity in the sensitive attributes or when the opponent has some background information. The t-closeness model is a more enhancement on the concept k-anonymity and l-diversity. One characteristic of the l-diversity model is that it serves all values of a given attribute in a similar way whatever is its distribution in the data. This is rarely the case for real data sets, since the attribute values may be much twisted. This may make it more difficult to create practical l-diverse representations. Usually, an opponent may use background knowledge of the overall distribution in order to make guessing about sensitive values in the data. Further, not all values of an attribute are equally sensitive. For example, an attribute related to a disease may be more sensitive when the value is positive, comparatively than when it is negative. t- closeness requires that the distribution of a sensitive attribute in any similarity class is close to the distribution of the attribute in the overall data set[9].

III. GENERAL FRAMEWORK OF EXISTING SYSTEM

In existing system architecture , there is an input dataset(file) which is not in appropriate format and then for proper dataset apply some pre-processing techniques(data cleaning, data reduction, data transformation) on it. On that pre-processed dataset apply k-anonymization and that anonymized data is used in simulation tools and identify the different classifier algorithm results. This general framework or architecture is as under:

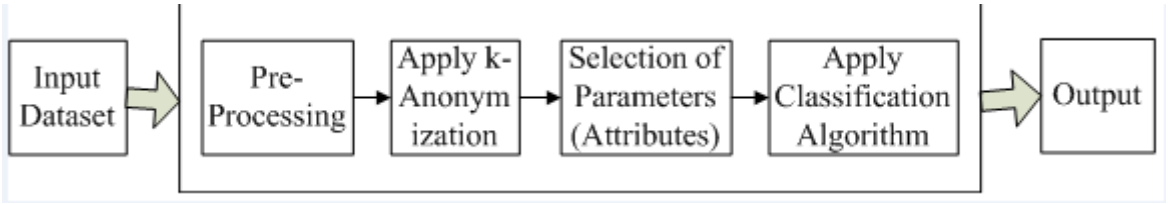

Figure 2: General architecture of existing algorithms

IV. ANONYMIZATION ALGORITHMS

There are number of algorithms based on various models of k-anonymity to achieve k-anonymity. In our relative study, we have chosen some k-anonymization algorithms. In the below section, we explain the algorithms applicable to the scope of this work, we likewise show a simplified representational so, a case for each of the algorithms, with the target of making them effortlessly possible for specialists[1].

(a) Samarati's Algorithm (b)Incognito Algorithm (c) Sweeney's Algorithm.

A. Samarati's Algorithm

This algorithm scans for the possible k- anonymous solutions by grasping different levels in Domain Generalization Hierarchy. It uses the binary search to gain the solution in less time. [11] Samarati makes the hypothesis that great solutions are the ones where end results in a table have minimum generalizations. Thus, her algorithm is planned to look at the generalizations that satisfy k- anonymity with minimal suppression. This algorithm accomplish the AGTS model, generalization is applied on column and suppression is applied on row. MaxSup is the greatest number of tuples that are allowed to be suppressed to achieve k anonymity.

B. Incognito Algorithm

Incognito algorithm [10] produces the set of all conceivable k-anonymous full-domain generalizations of relation T, with an optional tuple suppression threshold. In the algorithm each iteration consists of two parts. It starts by checking single attribute subsets of the quasi-identifier, and afterward repeats, checking k-anonymity with respect to larger subsets of quasi-identifiers.

C. Sweeney's Algorithm

Datafly Data fly algorithm is an algorithm for providing anonymity of Electronic Health Records [12].Anonymization is achieved by means of mechanically generalizing, substituting, inserting and removing statistics without losing details for research.

V. COMPARISION OF EXISTING ALGORITHM

	Algorithm	Pros	Cons
1	**Samarati's [11]**	**1.** Uses the binary search to acquire the solution in minimum time. **2.** Looks for the solution with the least generalization. **3.**samarati's outcome dependably has an chance to be an optimal solution **4.** Great result when compared to Datafly	**1.**The chance to get an optimal solution practically varies with k, MaxSup lattice size.
2	**Incognito [10]**	**1.**The algorithm finds all the k-anonymous generalizations **2.** Optimal solution can be selected according to various criteria	**1.**The algorithm uses breadth first search method which takes a lot of time to pass over the solution space
3	**Sweeney-Datafly [12]**	**1.**The algorithm checks very less nodes for k-anonymity due to which it is capable to give results very fast **2.**It is a greedy approach that creates frequency lists and repeatedly generalizes those composition with less than k occurrences **3.**Practically implementable	**1.** The algorithm skips many nodes, thus, resulting data is much generalized and sometimes this released data may not be useful for research purpose as it gives very less information. 2.Suppressing all values within the tuple

Figure 3: Comparison of existing algorithm

Comparison of Samarati's Algorithm, Incognito Algorithm and Sweeney's Algorithm- Datafly for anonymization is given in the table with advantages and disadvantages of each algorithm.

VI. FUTURE WORK

From this survey we understand that the more research is in work to include different extended data publishing scenarios such as Anonymizing sequential release with new attributes, multiple view publishing and incrementally update data records as well as non-numeric quasi identifiers. Other is to study on data in more detail and design various anonymization techniques which provide more accurate privacy preservation, and work on, semantic anonymization algorithm for decreasing the information loss and the dynamic version is provided based with a acceptable relation between privacy level and the utility.

VII. CONCLUSION

From above survey we can realize that anonymization is proportional to number of records, the value of k has to be chosen in a way it brings down the difference between the released micro data and the privacy. The number of k value enlarges the time taken for anonymization is increase, because when k increases, the time needing for anonymization is also increases. In the case of different size of data the anonymization time is incremented. In Sweeney's algorithm there is large variation of execution time. In Incognito algorithm execution time has less variation with the k value and data size. Execution time is comparatively low in Samarati's algorithm. When the data size is more, there is not any identifiable impact in the execution time. So from this analysis we can conclude that from between these three algorithms of anonymization Samarati's algorithm is the best algorithm for anonymization.

REFERENCES

[1] Pierangela Samarati, Latanya Sweeney "Generalizing Data to Provide Anonymity when Disclosing Information".

[2] R. Mahesh,T. Meyyappan "Anonymization Technique through Record Elimination to Preserve Privacy of Published Data"Proceedings of the 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering, February 21-22

[3] Ms. Simi M S, Mrs. SankaraNayaki K, Dr.M.Sudheep Elayidom "An Extensive Study on Data Anonymization Algorithms Based on KAnonymity" IOP Conf. Series: Materials Science and Engineering 225 (2017) 012279 doi:10.1088/1757-899X/225/1/012279.

[4] Athiramol. S,Sarju. S "A Scalable Approach for Anonymization Using Top Down Specialization and Randomization for Security" 2017 International Conference on Intelligent Computing,Instrumentation and Control Technologies (ICICICT)

[5] Latanya Sweeney"Weaving technology and policy together to maintain confidentiality" Journal of Law, Medicine, Ethics, 25(2(3):98110, 1997.

[6] Preet Chandan Kaur, Tushar Ghorpade, Vanita Mane "Analysis of Data Security by using Anonymization Techniques" 978-1-4673-8203-8/16/in 2016 IEEE.

[7] Qingshan Jiang, A S M Touhidul Hasan "A General Framework for Privacy Preserving Sequential Data Publishing" 2017 31st International Conference on Advanced Information Networking and Applications Workshops.

[8] Machanavajjhala A., Kifer D., Gehrke J., Venkitasubramaniam M.,"ldiversity: Privacy beyond k-anonymity" 2007, ACM Transaction on Knowledge Discovery in Data, 1, 18-27.

[9] Li, N., Li, T., and Venkatasubramanian, S.,"t-Closeness: Privacy beyond k-Anonymity and l-Diversity" 2007, Proceedings, 23rd International Conference on Data Engineering, USA, 106-115.

[10] L. Sweeney, "Datafly: a system for providing anonymity in medical data. In Database Security", XI: Status and Prospects, IFIP TC11 WG11.3 11th Int'l Conf. on Database Security, 356-381, 1998

[11] K. Bache and M. Lichman. UCI Machine Learning Repository, 2013.

[12] J. Soria-Comas, J. Domingo-Ferrer, D. Sanchez, and S. Martınez, "Improving the Utility of Differentially Private Data Releases via k- Anonymity", In Proceedings of the 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, TRUSTCOM-13, pages 372–379, 2013.

[13] http://www.cs.waikato.ac.nz/ml/weka/

[14] http://www.nltk.org/

[15] https://opennlp.apache.org/

[16] UCI Machine Learning Repository http://archive.ics.uci.edu/ml/datasets