Privacy Preserving for Health Records

K Tejaswini[1], G Sai manasa[2], SK Dawood afroz[3], B Varun reddy[4], Mr. G.Vijay suresh[5]
[1,2,3,4]*Students,* [5] *Assoc. Professor*
Deptt. of CSE, LBRCE, Mylavaram, LBRCE, Mylavaram, Andhra Pradesh, India

Abstract - Nowadays, information publishing as an indispensable part appears in our vision, bringing about a mass of discussions about methods and techniques of privacy preserving data publishing which are regarded as strong guarantee to avoid information disclosure and protect individuals' privacy. Recent work focuses on proposing different anonymity algorithms for varying data publishing scenarios to satisfy privacy requirements, and keep data utility at the same time. K-anonymity has been proposed for privacy preserving data publishing, which can prevent linkage attacks by the means of anonymity operation, such as generalization and suppression. Numerous anonymity algorithms have been utilized for achieving k-anonymity. This paper provides an overview of the development of privacy preserving data publishing, which is restricted to the scope of anonymity algorithms using generalization and suppression. The privacy preserving models for attack is introduced at first. An overview of several anonymity operations follow behind. The most important part is the coverage of anonymity algorithms and information metric which is essential ingredient of algorithms.

Keywords - Privacy Preserving, Anonymity Operations, Generalization, Suppression.

I. INTRODUCTION

Due to the rapid growth of information, the demands for data collection and publishing increase sharply. A great quantity of data is used for analysis, statistics and computation to find out general pattern or principle which is beneficial to social development and human progress.The requirement for data publisher is that data to be published must fit for the predefined conditions. Identifying attribute needs to be omitted from published dataset to guarantee that individuals privacy cannot be inferred from dataset directly. Removing identifier attribute is just the preparation work of data processing, several sanitization operations need to be done further. However, after data processing, it may decrease data utility dramatically, while, data privacy did not get fully
preserved. In face of the challenging risk, some researches have been proposed as a remedy of this awkward situation, which target at accomplishing the balance of data utility and information privacy when publishing dataset. The ongoing research is called Privacy Preserving Data Publishing (PPDP). In the past few years, experts have taken up the challenge and undertaken a lot of researches. Many feasible approaches are proposed for different privacy preserving scenario, which solve the issues in PPDP effectively. New methods and theory come out continuously in experts' effort to complete privacy preserving.

A. Privacy Preserving - Personal records of individuals are logically being gathered by various government and organization foundations for the requirements of information examination. The information examination is encouraged by these associations to distribute "adequately private" thoughts over this data that are collected. Privacy could be a twofold edged brand -there should be sufficient protection to ensure that touchy information concerning the general population isn't revealed by the perspectives and at a comparative time there should be sufficient data to play out the investigation. Besides, an enemy who needs to gather delicate information from the uncovered perspectives in some cases has some data concerning the general population inside the data. The principle goal is to change over the first data into some mysterious sort to prevent from inducing its record owner's sensitive information as examined.

B. K-Anonymity - When referring to data anonymization, the most common data is two-dimensional table in relational database. For privacy preserving, the attributes of table are divided into four categories which are identifier, quasi-identifiers, non-quasi attributes and sensitive attribute. Identifier can uniquely represent an individual. Obviously, it should be removed before data processing. Quasi-identifiers are a specific sequence of attributes in the table that malicious attackers can take advantage of these attributes linking released dataset with other dataset that has been already acquired, then breaking privacy, eventually gaining sensitive information. Data sanitization operated by data publisher mainly targets on quasi-identifiers. Due to uncertainty of the number of quasi-identifiers, each approach of PPDP assumes the quasi-identifiers sequence in advance. Only in this way can the following processing carry out. Non-quasi attributes have less effect on data processing. For this reason, sometimes, these attributes does not turn up in the progress of data processing which tremendously decrease memory usage and improve the performance of the proposed algorithm. Sensitive attribute contains sensitive information, such as disease, salary.

Table 1: Census Data

Name	Birthday	Sex	Zip code
Myron	1990/10/01	Male	210044
Yoga	1980/05/22	Female	210022
James	1992/07/12	Male	210001
Sophie	1997/03/03	Male	210012

Table 2: Patient Data

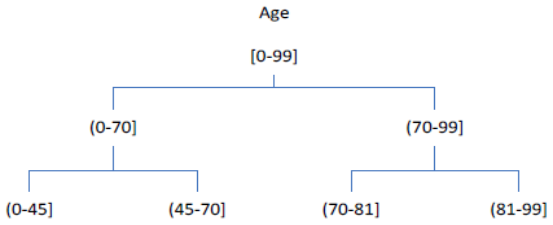
Id	Work	Birthday	Sex	Zip code	Disease
2310012	Learner	1992/12/12	Male	210044	Cardiopathy
2310022	Clerk	1998/12/04	Female	210033	Diabetes
2310032	Officials	1997/07/11	Male	210022	Flu
2310042	HR	1992/11/11	Male	210055	cancer

In this table:
Sensitive attribute : Diesese
Quasi-Attributes : Birthday, Sex, Zipcode
Non Quasi-Attributes : Work
Identifier : ID
By linking birthday, sex, Zipcode attributes we can retrieve the whole data from database without using Id.

C. K-Anonymity techniques -

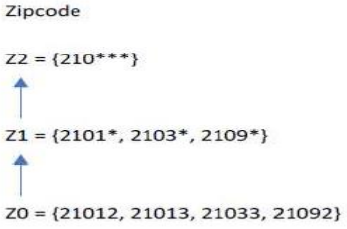
Generalization - Generalization is the way toward changing over an incentive into a less particular general term. For ex, "Male" and "Female" can be generalized to "Any". At the accompanying levels generalization procedures can be connected.

- Attribute (AG): Generalization is performed at the segment level; all the qualities in the section are generalized at a speculation step.
- Cell (CG): Generalization can likewise be performed on a solitary cell; at long last a summed up table may contain, for a particular section and values at various levels of generalization.



Suppression - Suppression comprises in averting delicate information by evacuating it. Suppression can be connected at the level of single cell, whole tuple, or
whole segment, permits diminishing the measure of speculation to be forced to accomplish k anonymity.

- Tuple(TS): Suppression is performed at column level; suppression operation evacuates entire tuple
- Attribute (AS): Suppression is performed at segment level; suppression operation shrouds every one of the estimations of a segment.
- Cell (CS): Suppression is performed at single cell level; at long last k-anonymized table may wipe out just certain cells of a given tuple/quality.



D. Paper Overview - This paper mainly refers to four topics that are privacy model, anonymity operation, information metric and anonymization algorithm. Due to different kinds of attacks to steal privacy, it forms different privacy preserving models for these attacks accordingly. Every privacy preserving model has its feature, so that researchers propose some theory and method for each type of attack. Algorithm implementation is based on specific theory and methodology. So each anonymity algorithm belongs to the specific privacy preserving model. As to anonymity operation and information metric, they are the details of algorithms. Anonymity operation is the core of algorithm, an algorithm often keep one or two operations in mind, and finally make the processed dataset to meet privacy requirement. The information metric is incorporated into the algorithm to guide its anonymity process or execution, and finally get better result rather than just get a rare result. Therefore, these four topics are essential parts of privacy preserving data publishing.There are several essential operations to implement data anonymization that are generalization, suppression, anatomization, permutation and perturbation.Generalization and suppression usually replace the specific value of quasi-identifiers with general value.Generally, there exists a taxonomy tree structure for each quasi-identifier that is used for replacement.Anatomization and permutation decouple the correlation of quasi-identifier and sensitive attribute by separating then in two datasets. Perturbation distorts dataset by the means of adding noise, exchanging value or generating synthetic data that must keep some statistical properties of original dataset.

Table 3: Patient Data

Work	Birthday	Sex	Zip code
Learner	1992/12/12	Male	210044
Clerk	1998/12/04	Female	210033
Officials	1997/07/11	Male	210022
HR	1992/11/11	Male	210055

Table 4: Background knowledge

Name	Birthday	Sex	Zip code
Myron	1990/10/01	Male	210044
Yoga	1980/05/22	Female	210022
James	1992/07/12	Male	210001
Sophie	1997/03/03	Male	210012

Table 4: 2-anonymous Patient Data

Work	Birthday	Sex	Zip code
Learner	1992/12/12	Male	210044
Clerk	1998/12/04	Female	210033
Officials	1997/07/11	Male	210022
HR	1992/11/11	Male	210055

II. ENCRYPTION

While your medical organization switching from paper records to electronic health records, Using an electronic health record system offers you much better control over

information security. Here are some reasons why electronic health records are more secure than paper records.

Encryption Keeps Information Secure - A paper record is open, giving anybody a chance to see it, translate subtleties, make a duplicate or even sweep or fax the data to an outsider. Electronic records can be ensured with hearty encryption techniques to shield vital patient data security from prying eyes.

Grant Access Only to Authorized Users - When you use a paper-based system for your patients' medical records, it's possible to access them without your knowledge. With an EHR system, you can control precisely who has access to patient information.

III. CONCLUSION

In this paper, we talked about the Privacy preserving information distributing and information anonymization. We likewise talked about different anonymization strategies and for the most part focused on k-anonymity which involves both generalization and suppression. The last part is about the generalization algorithm and its execution for securing the protection of information utilized for the most part for data analysis.

IV. REFERENCES

[1]. Kabir ME, Wang H, Bertino E. Efficient systematic clustering method for k-anonymization. Acta Informatica. 2011 Feb 1;48(1):51-66.

[2]. Byun JW, Kamra A, Bertino E, Li N. Efficient kanonymization using clustering techniques. In International Conference on Database Systems for Advanced Applications 2007 Apr 9 (pp. 188- 200). Springer, Berlin, Heidelberg.

[3]. Xiao X, Tao Y. Anatomy: Simple and effective privacy preservation. In Proceedings of the 32nd international conference on Very large data bases 2006 Sep 1 (pp. 139-150). VLDB Endowment.

[4]. Zhang X, Liu C, Nepal S, Yang C, Dou W, Chen J. Combining top-down and bottom-up: scalable subtree anonymization over big data using MapReduce on cloud. In Trust, Security and Privacy in Computing and Communications (TrustCom), 2013 12th IEEE International Conference on 2013 Jul 16 (pp. 501-508). IEEE.

[5]. Goldberger J, Tassa T. Efficient anonymizations with enhanced utility. In Data Mining Workshops, 2009. ICDMW'09. IEEE International Conference on 2009 Dec 6 (pp. 106-113). IEEE.

[6]. Terrovitis M, Mamoulis N, Kalnis P. Privacy preserving anonymization of set-valued data. Proceedings of the VLDB Endowment. 2008 Aug 1;1(1):115-25.

[7]. Huda MN, Yamada S, Sonehara N. On Enhancing Utility in k-anonymization. International Journal of Computer Theory and Engineering. 2012 Aug 1;4(4):527.

[8]. Bhaladhare PR, Jinwala DC. Novel Approaches for Privacy Preserving Data Mining in k-Anonymity Model. J. Inf. Sci. Eng.. 2016 Jan 1;32(1):63-78.

[9]. Mohammed N, Fung B, Hung PC, Lee CK.Centralized and distributed anonymization for high-dimensional healthcare data. ACM Transactions on Knowledge Discovery from Data(TKDD). 2010 Oct 1;4(4):18.

[10]. Fienberg SE, Slavkovic A, Uhler C. Privacy preserving GWAS data sharing. In Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on 2011 Dec 11 (pp. 628-635). IEEE.

[11]. Gkoulalas-Divanis A, Loukides G. PCTA: privacy constrained clustering-based transaction data anonymization. In Proceedings of the 4th International Workshop on Privacy and Anonymity in the Information Society 2011 Mar 25 (p. 5). ACM.

[12]. Data Anonymization", ACM 2011

[13]. Kisilevich S, Rokach L, Elovici Y, Shapira B. Efficient multidimensional suppression for kanonymity. IEEE Transactions on Knowledge and Data Engineering. 2010 Mar 1;22(3):334-47

[14]. Loukides G, Gkoulalas-Divanis A, Malin B. Anonymization of electronic medical records for validating genome-wide association studies. Proceedings of the National Academy of Sciences.2010 Apr 27;107(17):7898-903.

[15]. Cao J, Karras P, Raïssi C, Tan KL. ρ-uncertainty: inference-proof transaction anonymization. Proceedings of the VLDB Endowment. 2010 Sep 1;3(1-2):1033-44.