

Efficient Personalized Privacy Preservation Using Anonymization

#¹Ashwini N. Patil, #²Prof. R. N. Phursule

#¹anpatil.26@gmail.com

#²rphursule@gmail.com



#¹²JSPMs Imperial College Of Engineering and Research, Wagholi, Pune

ABSTRACT

The k-anonymity privacy for publishing micro data requires that each equivalence class contains at least k records. Many authors have studied that k-anonymity cannot prevent attribute disclosure. The technique of l-diversity has been introduced to address this; l-diversity requires that each equivalence class must have at least well-represented values for every sensitive attribute. In this paper, we show that l-diversity has many limitations. In particular, it is not necessary or sufficient to prevent attribute disclosure. Motivated by these limitations, we propose a new method to detect privacy which is called as closeness. We first present the base model t-closeness, which includes the distribution of sensitive attributes in any of the equivalence classes is near to the distribution of the attribute in the overall table (i.e., the difference between the two given distributions should be no more than threshold value t). tcloseness that gives higher utility. We present our methode for designing a distance measure between given two probability distributions and give two distance measures. Here we discuss the methode for implementing closeness as a privacy concern and illustrate its advantages through examples and experiments.

Keywords: Data Publishing, Privacy Preserving, Anonymity Algorithms, Information Metric, Generalization, Suppression

ARTICLE INFO

Article History

Received :15th April 2016

Received in revised form :

17th April 2016

Accepted : 19th April 2016

Published online :

23rd April 2016

I. INTRODUCTION

PRIVACY is very important issue when one wants to make use of data that includes sensitive information. Studies on protecting the privacy of individuals and the confidentiality of data is contributed from many fields, including computer science, statistics, economics. This is an field that attempts to answer the problem of how an organization, such as a hospital, government agency or any organisation, can release data to the people without harming the confidentiality of personal information. We focus on privacy measures that provide legal safety, present algorithms that protect data to make it safe for accessing while preserving useful information, and discuss methodes for analyzing the sensitive data. Many challenges still remain. It provides a summary of the current state, based on which we expect to see advances in years to come. As personal information is collected in increasingly detailed level by various organizations, privacy related concerns are

introducing significant challenges to the data management organisations. Data anonymization methodes have been proposed in order to allow processing of personal data without compromising users privacy. Nevertheless, practical problems like dependencies between values in personal records do not obtain a satisfying solution. Here, we focus on the anonymization of tree-structured personal records links. Personal information do not comprise just a single tuple in modern information systems. The information concerning a single person usually spans over several tables or it is kept in a more flexible representation as an XML record. Such tree structured data could not be anonymized effectively with table based anonymization techniques since the structural relation between different fields substantially differentiates the problem. The difficulty in anonymizing tree structured data has been considered in existing research literature, in the technique of multirelational k-anonymity. In our methode we consider general case for tree structured data and we propose an anonymization method that is not

dependent solely on the generalization of values, but also on the simplification of the data tree.

II. LITERATURE SURVEY

To introduce the concept of Efficient Personalized Privacy Preservation Using Anonymization. This paper analyzes many concepts of different authors as mentioned below:

In the paper Anonymizing Collections of Tree-Structured Data, Olga Gkountouna and Manolis Terrovitis [1] introduces real-world data which have implicit or explicit structural relations. Privacy preservation has focused on data with a very simple structure, e.g. data with very complex structure such as network graphs, but has ignored intermediate cases. Here we focus on tree structured data. Such data is required from various applications, e.g. XML documents. An example is a database where information about a person is scattered amongst tables that are associated through foreign keys. $k(m;n)$ anonymity, which provides protection and proposes a greedy anonymization technique that sanitizes large datasets.

Q Wang, C. Wang [21] introduces Enabling Public Verifiability and Data Dynamics for Storage Security in Computing, Computing has been thought as the next generation architecture of IT Enterprise. It moves the application software and databases to large data repositories, where managing data and services may not be fully trustworthy. This brings about many new challenges, which are not understood. This work studies the problem of ensuring the integrity of data storage in Computing. We consider the job of allowing a third party auditor (TPA), as a client. TPA removes the involvement of the client through the checking if the data stored in the is indeed intact. The support for data by the most general forms of operations performed on data, such as insertion and deletion, is also an important step toward practicality, since services in Computing are not limited backup data only.

Ateniese [3] developed a dynamic provable data possession protocol based on cryptographic hash function and symmetric key encryption. The main thing is to precompute a certain number of metadata during the setup period, so that the number of challenges is prevented and fixed beforehand. The author constructs a highly efficient and secure PDP technique based largely on symmetric key cryptography. This technique allows outsourcing of dynamic data, that is, it efficiently supports operations, such as block modification, deletion and append.

A. Juels and B. S. Kaliski [4], introduces HLA Based Solution. It supports public auditing without retrieving data block. It requires constant bandwidth. It is possible to compute an HLA which authenticates a linear combination of the individual data blocks.

N. Cao, S. Yu, S. Yang [5], tells us about Using Virtual Machine. They proposed Virtual machines that use RSA algorithm, for client data encryption and decryptions. Also SHA 512 algorithm is used which makes message digest and check the data integrity. Digital signature is used as an identity measure for client. It solves the problem of unauthorized access, integrity, privacy and consistency.

C. Erway, A. Kupcu [6], introduces Non Linear Authentication in which they suggested Homomorphic non linear authenticator with randomized masking techniques to obtain security. K. Govinda proposed digital signature method to protect the privacy and integrity of data. RSA algorithm is used for encryption and

decryption which uses the process of digital signatures for message authentication.

S. Marium [7] introduced Extensible authentication protocol through hand shake with RSA. They proposed identity based signature for class conscious architecture. They provide an authentication protocol for computing (APCC). APCC is more easy and efficient as compared to SSL authentication protocol. Here, Challenge handshake authentication protocol (CHAP) is used. When make request for any data or any service on the . The Service provider authenticator (SPA) orders the first request for client identity. Following are the steps:

- 1) When Client request for any service to service provider, SPA sends CHAP request challenge to the client.
- 2) The Client sends CHAP response or challenges which is calculated by using a hash function to SPA.
- 3) SPA compares the challenge value and its own calculated value. If they are similar then SPA sends CHAP success message to the client.

III. PROPOSED SYSTEM

We have proposed a novel method of privacy called closeness. We introduce two instantiations: a base model called t -closeness and a more flexible privacy method called (n, t) closeness. We explain the rationale of the (n, t) -closeness model and show that it gives a better balance between privacy and utility. The (n, t) -closeness model better protects the data while improving the utility of the released data. The t -closeness model was introduced to overcome attacks which were possible on diversity (like similarity attack). l -diversity model uses all values of a given attribute in a similar way (as distinct) even if they are semantically related. All values of an attribute are not equally sensitive. The algorithm to check (n, t) closeness could be given as follows.

input: P is partitioned into r partitions $\{P_1, P_2, \dots, P_r\}$
output: true if (n, t) -closeness is satisfied, false otherwise

1. **for** every P_i
2. **if** P_i contains less than n records
3. find=false
4. **for** every $Q \in \text{Parent}(P)$ and $|Q| \geq n$
5. **if** $D[P_i, Q] \leq t$, find=true
6. **if** find==false, **return** false
7. **return** true

The algorithm consists of following three subsections:

- 1) Choosing a dimension on which we have to partition : Find Number of rows in patient-enq
- 2) selecting a value to split and start Suppression : Here we suppress using a zipcode. This zipcode is having 5 digits like 46982. The variable inc is the value to split if we set inc= 4. The zipcode is displayed as first 4 digit numbers like 4698**. And we set threshold value $t=0.5F$ and n is the second highest value of table age-

count according to patients age in the table. For example we contain this data in our patient table,

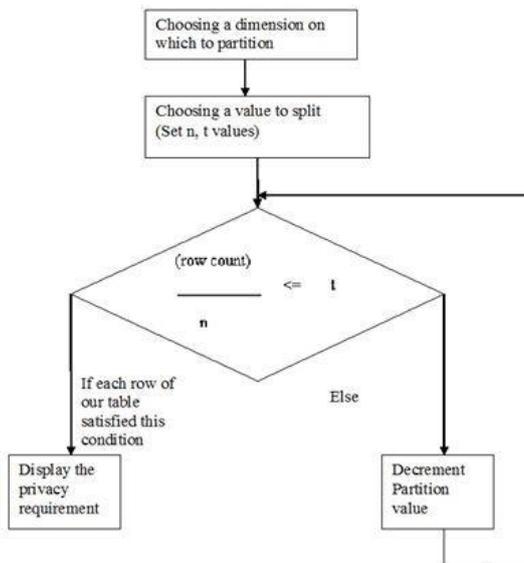
Age	Count
2*	6
3*	3
4*	10

- 3) Checking weather partitioning violates the privacy requirement : After that we check this following calculation.

$$t = \frac{(\text{rowcount})}{n} \quad (1)$$

If Each row of our table satisfies the condition, our privacy requirement is satisfied . Else we decrement our inc value and again we test this condition satisfied by each row or not till this condition will satisfied.

Flow of algorithm :



EARTH MOVER'S DISTANCE

Earth Movers distance is the difference between the probability distributions over a region d. The EMD is proposed on the minimal amount EMDmeasures the least amount of work needed to fill the of work needed to transform one distribution to another holes with earth. A unit of work corresponds to moving a unit of earth by a unit of ground distance.

IV. LIMITATIONS OF L-DIVERSITY

While the 'L-diversity principle represents an important step beyond k-anonymity in protecting against attribute disclosure, it has several shortcomings that we now discuss. 'L-diversity may be difficult to achieve and may not provide sufficient privacy protection.

L-diversity assumes an adversary who has knowledge of the form "Carl does not have heart disease," while closeness measures consider an adversary who knows the distributional information of the sensitive attributes. The goal is to propose an alternative technique for data publishing that remedies the limitations of L-diversity in

some application. An interesting question is how to effectively combine the existing techniques with generalization and suppression to achieve better data quality and privacy.

V. CONCLUSION

As seen above as k-anonymity protects against identity disclosure, it does not provide sufficient protection against attribute disclosure. The technique of l-diversity attempts to solve this problem. We have shown that -diversity has a number of limitations and especially discussed two attacks on l-diversity. Motivated by these limitations, we have proposed a novel privacy method called closeness. We propose two techniques: a base model called t-closeness and a more flexible privacy technique called (n, t) closeness. We explain the logic of the (n, t)-closeness model and show that it achieves a better balance between privacy and utility. Finally, through experiments on real data, we show that similarity attacks are a real problem and the (n, t)- closeness model better protects the data while improving the utility of the released data. (n, t)-closeness allows us to take advantage of anonymization techniques other than generalization of quasiidentifier and suppression of records.

ACKNOWLEDGMENT

I feel great pleasure to submitting this project paper on EFFICIENT PERSONALIZED PRIVACY PRESERVATION USING ANONYMIZATION. I wish to express true sense of gratitude towards my project guide, Prof. R. N. Phursule who at very discrete step in study of this project, contributed his valuable guidance and helped to solve every problem that arose. My great obligation would remain due towards Prof. S.R.Todmal (Head of Department), who was a constant inspiration during my project. He provided with an opportunity to undertake the project at JSPMs Imperial College Of Engineering and Research, Wagholi, Pune. I feel highly indebted to them who provided me with all my project requirements, and done much beyond my expectations to bring out the best in me.

I sincere thanks to our respected Principal Dr.S.V.Admane proved to be a constant motivation for the knowledge acquisition and moral support during our course curriculum.

REFERENCES

- [1] Anonymizing Collections of Tree-Structured Data Olga Gkountouna, Student Member, IEEE, and ManolisTerrovitis,IEEE Transactions on Data and Knowledge Engineering Vol No 27 Year 2015
- [2] G. Ateniese, R.D. Pietro, L.V. Mancini, and G. Tsudik, Scalable and Efficient Provable Data Possession, Proc. Fourth Intl Conf. Security and Privacy in Comm. Networks (SecureComm08), 2008.
- [3] G. Ateniese, R. D. Pietro, dynamic provable data possession protocol based on cryptographic hash function and symmetric key encryption 2007

- [4] A. Juels and B. S. Kaliski HLA Based Solution International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 1526-1532
- [5] N. Cao, S. Yu, Z. Yang, W. Lou, and Y. Hou, LT CodesBased Secure and Reliable Cloud Storage Service, Proc. IEEE INFOCOM, pp. 693-701, 2012
- [6] C. Erway, A. Kupcu, C. Papamanthou, and R. Tamassia, Dynamic Provable Data Possession, Proc. 16th ACM Conf. Computer and Comm. Security (CCS09), pp. 213-222, 2009
- [7] S. Marium, Q. Nazir, A. Ahmed, S. Ahthasham and Aamir M. Mirza, Implementation of EAP with RSA for Enhancing The Security of Cloud Computig, International Journal of Basic and Applied Science, vol 1, no. 3, pp. 177- 183, 2012
- [8] G. Aggarwal, T. Feder, K. Kenthapadi, S. Khuller, R. Panigrahy, D.Thomas, and A. Zhu, Achieving Anonymity via Clustering, Proc. ACM Symp. Principles of Database Systems (PODS), pp. 153- 162, 2006.
- [9] R.K. Ahuja, T.L. Magnanti, and J.B. Orlin, Network Flows: Theory, Algorithms, and Applications. Prentice-Hall, Inc., 1993
- [10] R.J. Bayardo and R. Agrawal, Data Privacy through Optimal k- Anonymization, Proc. Intl Conf. Data Eng. (ICDE), pp. 217-228, 2005.
- [11] F. Bacchus, A. Grove, J.Y. Halpern, and D. Koller, From Statistics to Beliefs, Proc. Natl Conf. Artificial Intelligence (AAAI), pp. 602- 608, 1992.
- [12] J.-W. Byun, Y. Sohn, E. Bertino, and N. Li, Secure Anonymization for Incremental Datasets, Proc. VLDB Workshop Secure Data Management (SDM), pp. 48-63 , 2006.
- [13] B.-C. Chen, K. LeFevre, and R. Ramakrishnan, Privacy Skyline: Privacy with Multidimensional Adversarial Knowledge, Proc. Intl Conf. Very Large Data Bases (VLDB), pp. 770- 781, 2007.
- [14] B.C.M. Fung, K. Wang, and P.S. Yu, Top-Down Specialization for Information and Privacy Preservation, Proc. Intl Conf. Data Eng. (ICDE), pp. 205-216, 2005
- [15] C.R. Givens and R.M. Shortt, A Class of Wasserstein Metrics for Probability Distributions, Michigan Math J., vol. 31, pp. 231-240, 1984.
- [16] V.S. Iyengar, Transforming Data to Satisfy Privacy Constraints, Proc. ACM SIGKDD, pp. 279-288, 2002.
- [17] D. Kifer and J. Gehrke, Injecting Utility into Anonymized Datasets, Proc. ACM SIGMOD, pp. 217-228, 2006.
- [18] N. Koudas, D. Srivastava, T. Yu, and Q. Zhang, Aggregate Query Answering on Anonymized Tables, Proc. Intl Conf. Data Eng.(ICDE), pp. 116- 125, 2007.
- [19] S.L. Kullback and R.A. Leibler, On Information and Sufficiency, Annals of Math. Statistics, vol. 22, pp. 79-86, 1951
- [20] D. Lambert, Measures of Disclosure Risk and Harm, J. Official Statistics, vol. 9, pp. 313-331, 1993
- [21] Enabling Public Verifiability and Data Dynamics for Storage Security in Cloud Computing Qian Wang ,

- Cong Wang , Jin Li , KuiRen , and Wenjing Lou
Illinois Institute of Technology, Chicago IL 60616
- [22] C. Wang, Q. Wang, K. Ren, and W. Lou, Privacy-Preserving Public Auditing for Data Storage Security in Cloud Computing, Proc. IEEE INFOCOM, pp. 525-533 , 2010



Ashwini N. Patil Department of Computer Engineering JSPMS ICOER Pune, India anpatil.26@gmail.com

Prof. R. N. Phursule Department of Computer Engineering JSPMS ICOER Pune, India rphursule@gmail.com