Comparative Analysis of Privacy Preserving Techniques in Distributed Database

Sapana Anant Patil [1], Dr. Abhijit Banubakode [2]

[1] Rajarshri Shahu College of Engineering,Tathawade, Pune ,India

[2] Rajarshri Shahu College of Engineering,Tathawade, Pune ,India

Abstract: ***In recent years, isolation takes an imperative role to secure the data from various probable attackers. For public advantage data need to be shared as required for Health care and researches, individual privacy is major concern with respect to sensitive information. For that anonymization of data with K-anonymity and L-diversity are studied. Existing system is depends on providers and is used generalization technique for anonymization. This will increase in data loss to avoid this slicing techniques are used. Data publishing is done in such a way that privacy of data should be preserved .While publishing collaborative data to multiple data provider's two types of problem occurs, first is outsider attack and second is insider attack. Outsider attack is by the people who are not data providers and insider attack is by colluding data provider who may use their own data records to understand the data records shared by other data providers. The paper focuses on insider attack, and makes some contributions. This problem can be overcome by combining slicing techniques with m-privacy techniques and addition of protocols as secure multiparty computation and trusted third party will increase the privacy of system effectively.***

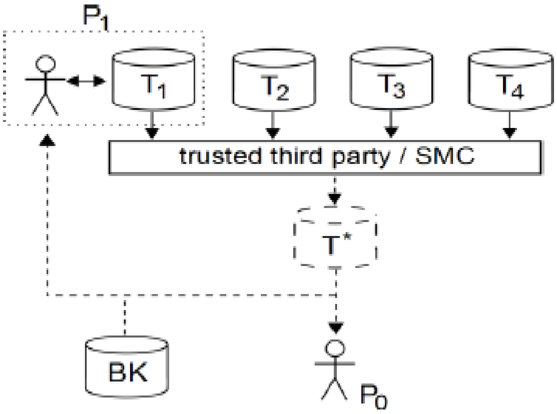
Keywords: Anonymization, Bucketization, Distributed database, Privacy, Security

1. Introduction

In distributed databases there is increasing need of sharing data that contain personal information. In healthcare domain, focus is to develop Information Network for distributing data among providers with privacy protection. Privacy preserving data publishing have established consideration in recent years as promising approaches are used for sharing data while preserving individual privacy.

Major goal is to distribute an anonymized view of combined data, T, which will be immune to attacks (figure 1). Attacker can be single or group of internal and external entities that want to break privacy of data using background knowledge. Collaborative data publishing is carried out successfully with the help of trusted third party (TTP) , which guarantees that information or data about particular individual is not disclosed anywhere, that means it maintains privacy. A more desirable approach for collaborative data publishing is, first aggregate then anonymize data into T* (figure1).

In figure 1 T1, T2 , T3 and T4 are databases for which data is provided by provider like provider P1 provides data for database T1. These distributed data coming from different providers get aggregated by TTP (trusted third party) or using SMC protocol. Then these aggregated data anonymized further by any anonymization technique. P0 is the authenticate user and P1 trying to breach privacy of data which is provided by other users with the help of BK(Background knowledge). This type of attack it can be called as a "insider attack", so protect system from such a type of attacks.



In health care all information related to patient is present in central network which includes disease details, corresponding treatment and test details. By using anonymization technique the data is anonymized and then released to the public. This process is called as the privacy preservation data publishing. The attributes are classified into three types as Key attribute, quasi identifier and sensitive attribute. Key attribute represents unique identification such as names, SSN (Social Security Number). Quasi-identifiers are segments of information that are not unique identifiers but well correlated with an entity; they can be combined with other quasi-identifier to create a unique identifier. Example birth date, gender, which can be used link unionized dataset with other data. Sensitive attributes contain sensitive value such diseases, policy detail, and salary etc. A data recipient may have access to some background knowledge which represents any publicly available information about released data, e.g., Census datasets. By m-privacy techniques, the information of the employee can be protected such as a sensitive attribute (SA) e.g. disease of patient, identifier (ID) e.g. name and quasi identifier (QI) i.e. age or zip code etc. But these methods have some limitation such as membership disclosure and data loss.

2. Generalization

It is the process of generalizing attribute separately. Using generalization the correlation between attribute is lost..

3. Suppression

Suppression is used to prevent the membership disclosure in the k-anonymity thus it be an assignment technique of placing * for the attribute values instead of their original values. This suppression technique is used in the quasi identifier fields to preserve the individual data

4. Bucketization

In Bucketization SAs are separated from the QIs by doing the random permutation on the SA values in each bucket. The anonymized data is collection of buckets, those bucket undergo the permutation on sensitive attribute values. Bucketization does not prevent membership disclosure. Bucketization requires the clear separation of SA and QI attributes and it breaks the attribute correlation between them.

5. Slicing

Slicing is a technique in which data is divided into vertical partition and horizontal partition. Vertical partition is a group of attributes in column based on correlationship among attribute. Horizontal partition is a group of tuples into buckets. In each bucket, each column consists of randomly permitted value.

6. Literature Survey

This section highlights the different methods which are previously used for anonymization. Also discuss some advantages and limitation of these systems. Privacy preserving data analysis and collaborative data publishing has received considerable attention in current years as promising approaches for sharing data while preserving individual privacy.

Period	Total Reference Collected	Paper Related To Generalization	Paper Related To Anonymization	Paper Related To Bucketization	Paper Related To Slicing
Privacy Technique Initiation on 2004					
2004-2008	10	4	5	-	-
2009-2010	15	5	8	1	-
2011-2014	14	2	4	2	2

Keyword	Content Type	Range of Publication Year	Author
Comparative Analyses of Privacy Preserving Technique	1) Conference Publications		
	2) Journals and Magazines		All
	3) Books and eBooks	From 2004 to 2014	Author
	4) Early Access Articles		

Sr. No	Name of Resource	Total No of Journal Preceeding	No.of Journal Proceeding on generalization	No. of Journal Proceeding on Anonymization	No.Of Journal Preceeding On Bucketization	No.Of Journal Preceeding on Slicing
1	**IEEE Xplore Digital Library**	08	02	04	02	01
2	**IEEE Xplore Conference Paper**	31	09	12	02	02
3	**Springer Link**	07	-	-	-	-
4	**ACM Digital Library**	08	01	03	02	01
5	**Advanced Search**	02	02	-	-	-
6	**Other**	05	02	02	02	01

Paper 1: ANGEL Technique published in IEEE 2009 [1]

In this paper [1] author has developed new anonymization technique that is that is effective in generalization in privacy protection but it able to retain significantly more as micro data. ANGEL(Anatomy and Generalization on Multiple Sensitive) is relevant to any monotonic principles . Author shows that ANGEL provides itself sophisticatedly to the hard problem of bordering publication. In generalization can issue only restricted marginal, ANGELM method can be used to publish any marginal with strong privacy guarantees.

To develop this approach they have use k-anonymity, data distribution, E-M generalization, anonymization principle and monotonicity. They also establish the privacy guaranty with generalization and anonymization algorithms.

Definition 1 (k-anonymity): E satisfies k-anonymity if every Equivalence Class in E comprises at least k tuples.

Definition 2 (SA-distribution): Given a multiset S of sensitive values, the SA-distribution in S is considered by a pdf

Definition 3 (E-M Generalization modeling): A generalization of a microdata table T can be effectively represented as a pair of E and M, denoted by (E,M).

Definition 4 (Anonymization principle): An anonymization principle is a constraint on an SA-distribution. A generalization (E,M) satisfies the principle if the SAdistribution of every EC in E satisfies the constraint.

Definition 5 (Monotonicity): An anonymization principle is monotonic if the following is true: given any two multisets of sensitive values S1 and S2 whose SA-distributions obey the principle, the SA-distribution of the union S1 ^{U} S2 also obeys the principle.

Paper 2: Slicing published in IEEE in year 2012 [2]

This paper [2] presents a new technique slicing which undergoes horizontal and vertical partitions of the data. Paper shows that slicing provide better data utility than generalization. Slicing can work efficiently on high-dimensional data and it can also be used for attribute disclosure protection. Slicing partitions attributes into columns which undergoes generalization, and divide tuples into buckets.

Paper 3: Privacy-Preserving for Anonymous and Confidential Databases in IEEE in year 2011 [3]

This system [3] is develop without knowing John and Harish content of tuple and database, inserted tuple is checked for K-anonymity. Author has proposed two protocols based on suppression and generalization. These protocol based on cryptographic assumption. This paper provides theoretical analyses to proof and experimental results to show their efficiency. This paper has data anonymization techniques to address the problem of privacy.

Paper 4: Privacy Preserving Research for Sensitive Attributes in Data in IEEE in year 2011 [4]

In this system they have develop a new generalization principle that effectively limits the risk of Multiple Sensitive Attributes privacy disclosure in re-publication. The results show that algorithm has higher degree of privacy protection and lower hiding rate. This approach the below definitions for execution

Definition 1(Identifier) Identifier can uniquely identify a single individual attribute such as name, id etc.

Definition 2(Quasi-identifier) Quasi-identifier cans connection with external data sources which can identify individual attribute such as age, sex etc.

Definition 3(Sensitive Attribute) Sensitive Attribute contains the properties of private dataset, such as disease doctor’s salaries.

Definition 4(Generalization) Generalization is a popular methodology of privacy preservation. It divides the tuple into QI-group, and then transforms the QI values in every group to a uniform format.

Definition 5(QI group) For a micro data table T(j), a QI-group is a subset of the tuples in T(j), which have the same generalized value for each non-sensitive attribute.

Definition 6(Signature) Let QI* be a QI group in T*(j) for any j. The signature of QI* is the set of distinct sensitive values in QI*.

Definition 7(Candidate Update Set) Suppose a is an element in the domain of attribute A, its candidate update set is the union of same elements in dom(A), such that a has non-zero update probability to it.

Paper 5: Slicing Models in IEEE in year 2013 With ICCTET [5]

This paper [5] has given suppression slicing is done by suppressing any one of the attribute value in the tuples and then perform the slicing. Thus utility is maintained with minimum loss by suppressing only very few values and privacy is maintained by random permutation. The next model is Mondrian slicing in this the random permutation is done with all the buckets not within the single bucket. Thus same utility of the original dataset is maintained. This approach use slicing, data publication, bucketization and generalization in the database.

7. Comparative Studies

Paper	*Observation*	*Remarks*
ANGEL:[1]	The last experiment on this approach gives comparison results when ρ(Anonymization principle) is 10-diversity (0.2-closeness). In all cases, releasing marginal always reduces reconstruction error. The improvement becomes more obvious when a marginal has a lower dimensionality.	This paper proposes angelization as a new anonymization technique for privacy preserving publication, which is applicable to any monotonic anonymization principle.
Slicing [2]	Workload experiments shows that slicing preserves data more accurately than Generalization. Slicing is better than Bucketization in workload consist of sensitive attribute. Experiment shows better performance with Slicing technique. Drawback of Bucketization is overcome by slicing.	A Slicing is a privacy-preserving technique for data publishing. Drawbacks of Bucketization and Generalization are overcome by Slicing. Slicing protects against privacy threat. Data characteristics is analyzed before anonymization of data.

Privacy-Preserving for Anonymous and Confidential Databases [3]	Some observation are done. 1: If none of the tuples in the chunk matches the User tuple, then the loader reads another chunk of tuples from the k-anonymous DB. Note the communication between the prototype and User is mediated by an anonymizer (like Crowds) and that all the tuples are encrypted. 2: The experiments confirm the fact that the time spent by both protocols in testing whether the tuple can be safely inserted in the anonymized database decreases as the value of k increases. Intuitively, this is due to the fact that the larger the k is, the smaller the witness set. Fewer are the partitions in which table T is divided Consequently, fewer protocol runs are needed to check whether the update can be made. Further, we report that the experiments confirm the fact that the execution times of Protocols	This paper proposed two secure protocols to check K-anonymous database for anonymity when a new tuple is inserted. With the use of proposed protocol new database become K-anonymous, query result returned by user is also K-anonymous. Privacy of provider never be affected by any query. As long as the database is updated properly using the proposed protocols, the user queries under our application domain are always privacy-preserving.
Privacy Preserving Research for Sensitive Attributes in Data [4]	Experiments generate original dataset T with 50k records, comprising Name, Gender, Age, Zip code, Disease and Doctor attributes. Disease attribute is self-defined which contains seven categories of diseases and every one is a candidate update set, a total of 60. Name, Gender, is categorical attributes and Age, Zip code are numerical attributes, Disease and Doctor are multiple attributes. In this experiments, name as the identifier of individuals which is suppressed in publishing table, Gender, Age, Zip code as the quasi-identifier.	This paper presents an analytical study that various inference channels of publishing of dynamic multiple sensitive attribute dataset and discuss how to avoid such inferences. As a second step, It provides an efficient algorithm that improving the limitations of previous studies, which adequately protects privacy and has low Number of Counterfeits.
Slicing Models [5]	In the tuple partitioning algorithm takes two phases. In the first phase tuples are partitioned into buckets. The tuple partition algorithm is defined by modifying the Mondrian algorithm for better performance and security. All results got on satisfactory level.	This paper has two enhanced techniques to preserve the privacy in data publishing. Thus the both techniques will preserve the membership disclosure and provide more utility than the existing system. The diversity checks in the Mondrian and suppression slicing will ensure that these techniques will satisfy privacy requirement of l-diversity.

8. Conclusion

This paper having lot of enhanced techniques to preserve the privacy in data publishing. Drawback of Generalization and Bucketization is overcome by Slicing The diversity checks in the Mondrian and suppression slicing will ensure that these techniques will satisfy privacy requirement of l-diversity. Basically slicing is the important technique with all available methodologies like data publication, bucketization and generalization in the database.

9. Acknowledgement

I would like to thank my guide Dr. A. N. Banubakode and Dr. P. K. Deshmukh for their help and guidance throughout this project and the semester, without him this would not have been possible.

References

[1] Yufei Tao, Hekang Chen, Xiaokui Xiao, “ANGEL: Enhancing the Utility of Generalization for Privacy Preserving Publication” IEEE transaction on knowledge and data engineering VOL 21, No. 7 jully 2009R. Caves, Multinational Enterprise and Economic Analysis, Cambridge University Press, Cambridge, 1982. (book style).

[2] Tiancheng Li, Ninghui Li, Jian Zhang, Ian molloy “Slicing: A New Approach for Privacy Preserving Data Publishing” in IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24,NO. 3, MARCH 2012.

[3] Alberto Trombetta, Wei Jiang, Elisa Bertino, Lorenzo Bossi “Privacy-Preserving Updates to Anonymous and Confidential Databases” in IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, VOL. 8, NO. 4, JULY/AUGUST 2011.

[4] Xiaolin Zhang, Lifeng Zhang “Privacy Preserving Research for Re-publication Multiple Sensitive Attributes in Data” in 978-1-4244-8728-8/11/$26.00 ©2011 IEEE.

[5] S.Kiruthika, Dr.M.Mohamed Raseen “Enhanced Slicing Models For Preserving Privacy In Data Publication” in International Conference on Current Trends in Engineering and Technology, ICCTET’13.

[6] Xiao X,Tao Y.m-invariance:Towards privacy preserving republication of dynamic datasets[C].In:Proceedings of the 2007 ACM SIGMOD international conference on Management of data(SIGMOD),Beijing:ACM Press,2007.689-700.
[7] Ying-yi Bu, Ada Wai-Chee Fu, Raymond Chi-Wing Wong, et al. Privacy Preserving Serial Data Publishing By Role Composition[C].In:The 34TH Int1 Conf on Very Large Data Bases[VLDB].Auckland, New Zealand:2008.
[8] Li F, Zhou S. Challenging More Updates: Towards Anonymous Republication of Fully.
[9] Dynamic Datasets[C].Computing Research Repository.New York:2008.
[10] Jha S, Kruger L, Mcdaniel P. Privacy Preserving Clustering[C].In:Proceedings of the 10TH European Symposium on Research in Computer Security.Milan,Italy:2005.
[11] Younho Lee proposed “secure ordered data bucketization ” Dependable and Secure Computing, IEEE Transactions on (Volume:11 , Issue: 3) in June 2014.

Author Profile

Sapana Anant Patil had completed Bachelor of Engineering in Computer Engineering from Shri Sant Gajanan Maharaj College Of Engineering from Amaravati University in 2008 and currently pursuing Master Of Engineering in Computers, from Rajarshri College Of Engineering Pune under University of Pune, MH, India.

Dr. Abhijit Banubakode received Ph.D. degree in Computer Studies from Symbiosis Institute of Research and Innovation (SIRI), a constituent of Symbiosis International University (SIU), Pune, India in April 2014 and ME degree in Computer Engineering from Pune Institute of Computer Technology (PICT), University of Pune ,India in 2005 and BE degree in Computer Science and Engineering from Amravati University, India, in 1997. His current research area is Query Optimization in Compressed Object-Oriented Database Management Systems (OODBMS). Currently he is working as Professor and Head of Department (HOD) in Department of Information Technology, Rajarshi Shahu College of Engineering, Pune, India. He is having 16 years of teaching experience. He is a member of International Association of Computer Science and Information Technology (IACSIT), ISTE, CSI and presented 12 papers in International journal and conference.