IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 6, August 2014.
www.ijiset.com
ISSN 2348 – 7968

Privacy in "Anonymizing Horizontally Partitioned Data"

E. Uday Reddy[1], Ms V Uma Rani[2] and Dr M Srinivasa Rao[3]

[1] PG Scholar, School of IT, JNTU-Hyderabad, Telangana, India.

[2] Assistant Professor, School of IT, JNTU-Hyderabad, Telangana, India.

[3] Professor, School of IT, JNTU-Hyderabad, Telangana, India.

Abstract

Privacy preserving data analysis and publishing has received considerable attention in recent years. Most work has been focused on a single data provider setting and considered the data recipient as an attacker. A limited background data is assumed from the literature of the attacker, and by considering specific types of attacks we define privacy using relaxed adversarial notion. In this, malicious users are colluding the data and Anonymization techniques cannot control the all different attackers. In this, we consider the collaborative data publishing setting with horizontally distributed data across multiple data providers each grant a subset of records Ti considering all the privacy constraints. Each record has an owner, whose identity should be protected. Data recipient may have access to some background knowledge, which represents any information which is available publicly about released data. As a result it removes all duplicate records' information and gives the better utility and efficiency results.

Keywords: Collaborative data publishing, data recipient, privacy preserving, data provider and privacy constraints.

1. Introduction

Data mining is discovering knowledge or interesting patterns from huge amount of data. In recent years, with the explosive development in Internet, privacy preservation, data storage and processing technologies, has been one of the greater concerns in data mining. A number of methods have been developed for privacy preserving data mining. Privacy preserving data mining is an important issue in the areas of data mining and security on private data in the following scenario: Multiple data providers, each having a private data set, want a group of people organized for a joint purpose rule mining without revealing their private data to other parties. Because of the influencing nature among parties, to develop a framework and to achieve such a computation is both desirable and challenging. There is an increasing need for sharing data repositories containing personal information across multiple distributed, possibly entrusted, and private databases. Such data sharing is subject to constraints imposed by privacy of data subjects as well as data confidentiality of institutions or data providers. We developed a set of decentralized protocols that enable data sharing for horizontally partitioned databases given these constraints.

2. Related Work

2.1 Privacy-Preserving Data Publishing:

In the most basic form of privacy-preserving data publishing (PPDP), the data holder has a table of the form
D (Explicit Identifier, Quasi Identifier, Sensitive Attributes, on-Sensitive Attributes)
where Explicit Identifier is a set of attributes, such as Aadhar number and name, containing information that explicitly identifies record owners; Quasi Identifier (QID) is a set of attributes that could potentially identify record owners; Sensitive Attributes consist of sensitive person-specific information such as abnormality, earnings, and disability status; and Non-Sensitive Attributes contains all attributes that do not fall into the previous three categories.
Anonymization refers to the PPDP approach that seeks to hide the identity and/or the sensitive data of record owners, assuming that sensitive or personal data must be retained for analysis. Clearly, explicit identifiers of record owners must be removed and Adversary is a coalition of colluding data providers or data owners and attempts to infer data records contributed by other data providers. To prevent record linkage through Quasi Identifier we consider K-anonymity i.e., If one record in the table has some value QID, at least k − 1 other records also have the value QID. In other words, the minimum equivalence group size on QID is at least k. A table which satisfies this need is called k-anonymous.
In a k-anonymous table, with respect to QID each record is indistinguishable from at least k −1 other records. Therefore, the probability of connecting a victim to a specific record through QID is at most 1/k.
Our aim is to provide an anonymized view of the integrated data T* which will be immune to attacks. Attackers can be from internal or external entities which want to break

ISSN 2348 – 7968

privacy of data using user background knowledge and anonymized data.

3. Existing system

Most work has been focused on single data provider setting and considered the data recipient as an attacker. A huge body of composition assumes limited background knowledge of the attacker, and defines privacy using relaxed adversarial conviction by considering specific types of attacks.

A trusted third party protocols are used to warranty that there is no admission of intermediate information during the anonymization. Finally neither trusted third party nor secure multiparty computation protects against inferring information using the anonymized data.

3.1 Drawbacks in Existing system

Till now problem of deducing information from anonymized data has been studied extensively in a single data provider settings. A data recipient that is an attacker e.g., Po, attempts to deduce additional information about data records using the published data, T* and background knowledge.

- So k-anonymity protects against admission attacks by taking each QID which inturn contain at least k-records.
- Malicious users are colluding the data E.g., shilling attackers.
- Anonymization techniques can't control the all different attackers.

Finally we conclude that with differential privacy we cannot deduce the statistical data with includes little background data.

4. Proposed System

- Each contributing a subset of records (Ti).
- Each element attribute is an identifier, which identifies the owner, or a quasi identifier (QID) directly.
- As a special case, a data provider could be the data owner itself who is contributing its own records.
- A data recipient may have access to some background knowledge e.g., Census datasets.

When we compare previous external recipient, each provider has additional data knowledge if its own records which helps with the attack. It is further deteriorated when multiple data providers collude with each other.
We explain the adversary threats with an example shown in Table 1.

Table 1: *Adversary and privacy example which Assume that Software industries want to anonymize their respective employee databases B1, B2, B3, and B4.*

B1

Name	Age	Id	Designation
Ramchand	27	1001	Clerk
Sania	36	1011	Analyst
Sachinjoshi	21	1022	Programmer

B2

Name	Age	Id	Designation
Arun	33	1002	Clerk
Mallesh	38	1012	Technocrat
Santosh	32	1013	Technocrat

B3

Name	Age	Id	Designation
Bhavani	33	1002	Analyst
Bhupesh	38	1012	Technocrat

B4

Name	Age	Id	Designation
Arun	33	1002	Clerk
R.mahesh	38	1012	Programmer

In each of the databases B1, B2, B3, B4, Name is an identifier, {age,id} is a quasi- identifier(QI), and Designation is a sensitive attribute. We observe one record, owned by Arun is contributed by two providers P2 and P4, and is represented as a single record in anonymized dataset. So, we anonymize the employee records to choose a small class of individual employee information for data analysis. Since it contains the subset of the complete data, this inherent data has to be modeled externally when the data are anonymized. In general the attacker use more attributes as a quasi identifier and back ground knowledge to scale the linking attack. Access to the unified data may have entry to the multiple databases.
Eg., a employee switching to another company and using information about her previous employees.
In the same way in social network sites a user may infer private info about others using the anonymized data.

ISSN 2348 – 7968

Table 2: Ta*

Provider	Names	Age	Id	Designation
P1	Ramchand	[20-30]	****	Clerk
P1	Scahin joshi	[20-30]	****	Programmer
P3	**Bhavani**	**[20-30]**	****	**Analyst**

Provider	Names	Age	Id	Designation
P2	Santosh	[31-34]	****	Technocrat
P2,P4	Arun	[31-34]	****	Clerk
P4	R.Mahesh	[31-34]	****	Programmer

Provider	Names	Age	Id	Designation
P1	Sania	[35-40]	****	Analyst
P2	Mallesh	[35-40]	****	Technocrat
P3	Bhupesh	[35-40]	****	Technocrat

*T*a is one possible anonymization that guarantees k-anonymity and l-diversity i.e., each Quasi-identifier group contains records with at least l different sensitive values.*

However, an attacker from the Software industry P1 may remove all records from P1. In the first QI group there will be only one remaining record, which belongs to an employee between 20 and 30 years old.

*Table 3: Tb**

Provider	Names	Age	Id	Designation
P1	Ramchand	[20-40]	****	Clerk
P2	Mallesh	[20-40]	****	Technocrat
P3	Bhavani	[20-40]	****	Clerk

Provider	Names	Age	Id	Designation
P1	SachinJoshi	[20-40]	****	Programmer
P2,P4	Arun	[20-40]	****	Clerk
P3	Bhupesh	[20-40]	****	Technocrat

Provider	Names	Age	Id	Designation
P1	Sania	[20-40]	****	Analyst
P4	R.Mahesh	[20-40]	****	Technocrat
P2	Santosh	[20-40]	****	Technocrat

In the above table Tb is an anonymized table which satisfies k-anonymity and l-diversity.*

By joining this record with the background knowledge BK using Quasi identifier attributes, P1 can identify Bhavani as the owner of the record(highlighted in the table) and his designation Analyst.

5. Conclusion

In this paper we discussed various privacy issues in getting horizontally partitioned data considering insider attacks. In distributed privacy preserving data mining, we try to develop more efficient algorithms and look for a balance between admission cost, computation cost and communication cost. So to deploy privacy-preserving techniques into practical applications also needs to be further studied.

6. References

1. *Benjamin C. M. Fung, Ada Wai-Chee Fu ,Ke Wang, and Philip S. Yu. Introduction to Privacy-Preserving Data Publishing Concepts and Techniques*
2. *B. C. M. Fung, K.Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent development".*
3. *Adi Armoni, Tel-Aviv University, Israel, Data Security Management In Distributed Computer Systems.*
4. *Z. Zhan and S. Matwin, "A crypto-approach to privacy preserving data mining," In IEEE International Workshop on Privacy Aspect of Data Mining, Hong Kong.*
5. *J. Vaidya and C. Clifton, "Privacy preserving association rule mining in vertically partitioned data," In Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, Canada.*

Author Profiles

E. Uday Reddy is currently pursuing his M.Tech (Computer Science) in School of IT, JNTU-Hyderabad. He did his B.Tech in Computer Science &Engineering from Nishitha College of Engineering and Technology, Hyderabad. His research area interest includes data mining and information security

Ms Uma Rani is presently working as assistant professor in School of IT, JNTU Hyderabad .She has more than nine years of experience. Her area of interest is data mining.

ISSN 2348 – 7968

Dr. M Srinivasa Rao is former director of School of IT and he is currently working as professor in it. His articles and publications are published all over the world.
His area of interest includes Web Technologies, Artificial Neural Networks, Data mining, Software Testing Tools and IT workshop.