International Journal for Rapid Research in Engineering Technology and Applied Science
ISSN (Online): 2455-4723
Vol 1 Issue 1 December 2015

Data Attribute Security for High Dimensitional Data Set

Harsha A. Chaudhari*, Prof. Chhaya Nayak**
II[nd] Year Mtech, Department of Computer Science and Engineering, Indore, MP, India*
HOD, Department of Computer Science and Engineering, Indore, MP, India**
*Harsha.chaudhari22@gmail.com**, *hod.computers@bmcollege.ac.in***

Abstract: In the recent year, the privacy takes major role to secure the data from various potential attackers. While publishing collaborative data to multiple data provider's two types of problem arises, 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. In the proposed approach problem can be resolved by using different approaches as m-privacy, which is a technique which guarantees that the anonymized data satisfies a given privacy constraint against any group of up to m-colluding data-providers. Second, Heuristic algorithms is also exploiting the equivalence group monotonicity of privacy constraints and adaptive ordering techniques for efficiently checking m-privacy given a set of records. Data provider aware anonymization algorithm with adaptive m- privacy checking strategy is to ensure high utility and m-privacy of anonymized data with efficiency. Privacy for collaborative data publishing can further enhanced by combining techniques of m-privacy with Slicing techniques. And by using secure protocols as trusted-third party (TTP), secure multiparty computation (SMC) or enhancement in the protocol security can be done effectively.

Keywords: M – privacy, L-diversity, Data Anonymization, Slicing, Bucketization.

I. Introduction

Privacy-preserving publishing of micro data has been studied extensively in recent years. Micro data contain records each of which contains information about an individual entity, such as a person, a household, or an organization. Several micro data anonymization techniques have been proposed. Several anonymization techniques, such as generalization and bucketization, have been designed for privacy preserving micro data publishing.

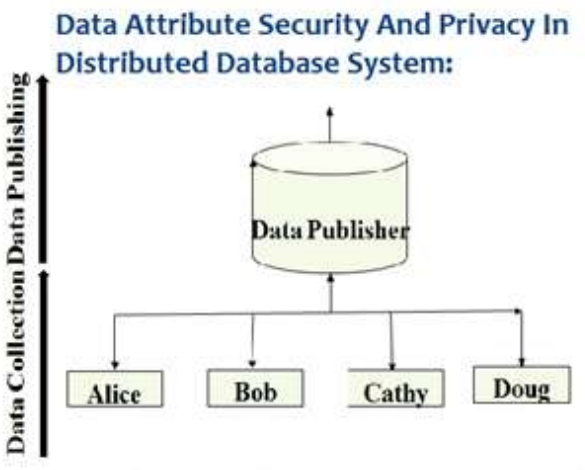

Figure 1: Data collection and data publishing

The most popular ones are generalization for k-anonymity and bucketization for ℓ- diversity. Recent work has shown that generalization loses considerable amount of information, especially for high dimensional data. Bucketization, on the other hand, does not prevent membership disclosure and does not apply for data that do not have a clear separation between quasi-identifying attributes and sensitive attributes. In this paper, we present a novel technique called slicing, which partitions the data both horizontally and vertically.

We present a novel technique called slicing, which partitions the data both horizontally and vertically. We show that slicing preserves better data utility than generalization and can be used for membership disclosure protection. Another important advantage of slicing is that it can handle high-dimensional data. We show how slicing can be used for attribute disclosure protection and develop an efficient algorithm for computing the sliced data that obey the ℓ-diversity requirement. Our workload experiments confirm that slicing preserves better utility than generalization and is more effective than

bucketization in workloads involving the sensitive attribute.

Privacy preservation techniques are mainly used to reduce the leakage of formation about the particular individual while the data are shared and released to public. For this, the sensitive information should not disclose. Data is getting modified first and then published for further process. For this various anonymization techniques are followed and they are generalization, suppression, permutation and perturbation.

By various anonymization techniques data is modified which retains sufficient utility and that can be released to other parties safely. Single organization does not hold the complete data. Organizations need to share data for mutual benefits or for publishing to a third party. For banking sector want to integrate their customer data for developing a system to provide better services for its customers. However, the banks do not want to indiscriminately disclose their data to each other for reasons such as privacy protection and business competitiveness.

Main goal is to publish an anonymized view of integrated data, T, which will be immune to attacks(fig1). Attacker runs the attack, i.e. a single or a group of external or internal entities that wants to breach privacy of data using background knowledge. Collaborative data publishing is carried out successfully with the help of trusted third party (TTP) or Secure Multi Party Computation (SMC) protocols, which guarantee that information or data about particular individual is not disclosed anywhere, that means it maintains privacy. Here it is assumed that the data providers are semi honest. A more desirable approach for collaborative data publishing is first aggregate then anonymize.

II. Literature Survey

In this paper [1] they have 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 is applicable to any monotonic principles (e.g., l-diversity, t-closeness, etc.), with its superiority (in correlation preservation) especially obvious when tight privacy control must be enforced. We show that ANGEL lends itself elegantly to the hard problem of marginal publication. In particular, unlike generalization that can release only restricted marginal, our technique can be easily used to publish any marginal with strong privacy guarantees.
For develop e 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.

This system [2] develop for to check whether the database inserted with the tuple is still k-anonymous, without letting Alice and Bob know the contents of the tuple and the database, respectively. In this paper, we propose two protocols solving this problem on suppression-based and generalization-based k-anonymous and confidential databases. The protocols rely on well-known cryptographic assumptions, and we provide theoretical analyses to proof their soundness and experimental results to illustrate their efficiency.

This paper having the techniques addressing the problem of privacy via data anonymization has been developed, thus making it more difficult to link sensitive information to specific individuals. One well-known technique is k-anonymization.

This paper [3] focuses on the organization of the collection and anonymization phases at the data source (i.e., at each SPT) while compromising neither privacy nor data utility compared to a trusted central server approach. The problem is difficult due to three assumptions: (1) the data publisher and the data recipients are untrusted, (2) the SPTs are trusted but there is no direct communication between them and (3) there is no certainty about the connection frequency and duration of each SPT connection.

Given system focused precisely addresses this issue and proposes to adapt the traditional Generalization privacy mechanism to an environment composed of a large set of tamper-resistant smart portable tokens seldom connected to a highly available but untrusted infrastructure. This conjunction of hypothesis makes the problem fundamentally different from any previously studied privacy-preserving data publishing problem we are aware of.
This system [4] present a novel technique called slicing, which partitions the data bothhorizontally and vertically. System shows that slicing preserves better data utility thangeneralization and can be used for membership disclosure protection.

Another importantadvantage of slicing is that it can handle high-dimensional data. It show how slicing can beused for attribute disclosure protection and develop an efficient algorithm for computing the sliced data that obey the ℓ-diversity requirement.

It shows that slicing can be effectively used for preventing attribute disclosure, based on the privacy requirement of ℓ-diversity. Efficient algorithm for computing the sliced table that satisfies ℓ-diversity. Our algorithm partitions attributes into columns, applies column generalization, and partitions tuples into buckets. Attributes that are highly-correlated are in the same column.

Privacy issues arise in various area such as health care, intellectual property, biological data etc. It is one of the challenging issues when sharing or publishing the data between one to many sources for research purpose and data analysis. Sensitive information of data owners must be protected. There are two kinds of major attacks against privacy namely record linkage and attribute linkage attacks Earlier, researchers have proposed new methods namely *k*-anonymity, ℓ-diversity, t-closeness for data privacy. K-anonymity method preserves the privacy against record linkage attack alone. It fails to address attribute linkage attack. ℓ-diversity method overcomes the drawback of *k*-anonymity method. But it fails to address identity disclosure attack and attribute disclosure attack in some exceptional cases. *t*-closeness method preserves the privacy against attribute linkage attack but not identity disclosure attack. But it computational complexity is large.

In this paper [5], the authors propose a new method to preserve the privacy of individuals' sensitive data from record and attribute linkage attacks. In the proposed method, privacy preservation is achieved through generalization of quasi identifier by setting range values and record limitation. The proposed method is implemented and tested with various data sets.

The suppression slicing [6] 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 proposed database.

III. Protocols

A. Mathematical Model Of Proposed Work

Let, S={s,e,X,Y,F}
Where S is a system of collaborative data publishing consist of database with certain attributes related to patient data for hospital management system. S consist of
s = distinct start of system
e = distinct end of system
X= Input of system from users
Y = output of system
F = algorithms or functions having certain computation time
Let,
s = {Ru} // Request from users
= {Rud, Rua } // Rud=request from doctors, Rua= request from admin
X= {DBp1,DBp2…….DBpn }
// database i.e data provided by providers
// Apply F on s and P.
F = {slicing algorithm(SA), L diversity (LD), provider aware algorithm(PA)}
Y = {T1*,T2*,T}
T1*= {Rud^'DBpn}
// collaborative data according to user request and database which we have. Slicing and L diversity provides privacy and security to input data.
T2*={Rud^DBpn}
// After applying PA on database after user request
T1={Rua^DBpn}
// Original data view to authenticated user admin.
e = output in table format according to user authentication.
Success condition,
Ru ≠ NULL, DBpn ≠ NULL

Failure condition,

Ru = = NULL, DBpn = = NULL

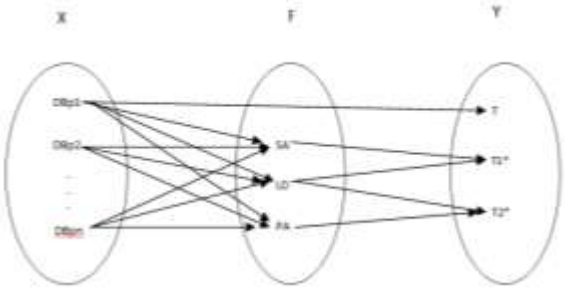

Figure 2: NP Hard and NP Complete

NP Hard: NP hard is a system where there is only one distinct answer is possible.

NP Complete: NP complete is a system where there is at least one way exist for a solution i.e deterministic solution. For anonymization there are many algorithm available or other user can develop another algorithm for it. Slicing technology with permutation which I used for anonymization. Therefore my system is NP complete system.

Flow diagram

A diagram showing the flow of information through the function and the transformation it undergoes is presented.

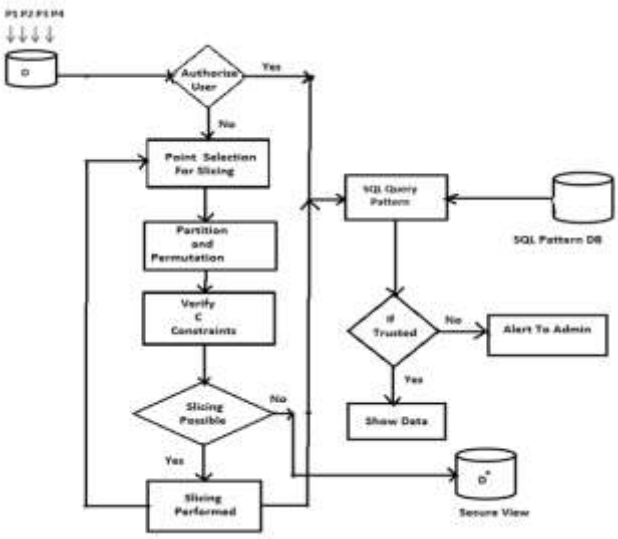

Figure 3: Flow diagram

IV. Proposed System

Main goal is to publish an anonymized view of integrated data, T, which will be immune to attacks. Attacker runs the attack, i.e. a single or a group of external or internal entities that wants to breach privacy of data using background knowledge. Collaborative data publishing is carried out successfully with the help of trusted third party (TTP) or Secure Multi Party Computation (SMC) protocols, which guarantees that information or data about particular individual is not disclosed anywhere, that means it maintains privacy. Here it is assumed that the data providers are semi honest. A more desirable approach for collaborative data publishing is, first aggregate then anonymize .

In above diagram, 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 aggregate by TTP(trusted third party) or using SMC protocol. Then these aggregated data anonymized further by any anonimization 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 we can call as a "insider attack". We have to protect our system from such a type of attacks

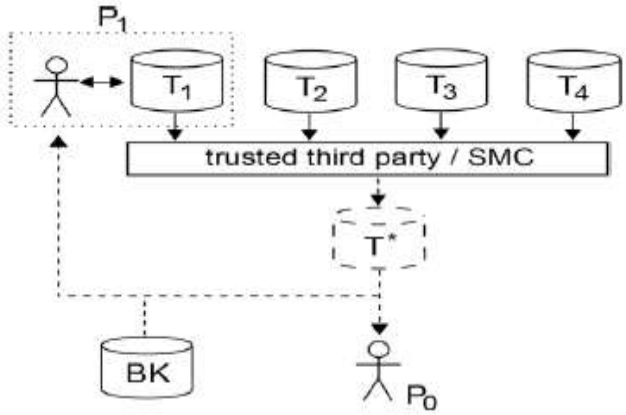

Figure 4: System Module

Proposed System Architecture

A system architecture or system's architecture is the conceptual model that defines the structure, behaviour, and more views of a system. An architecture description is a formal description and representation of a system, organized in a way that supports reasoning about the structures of the system.

Below methodology we use when we develop the proposed approach

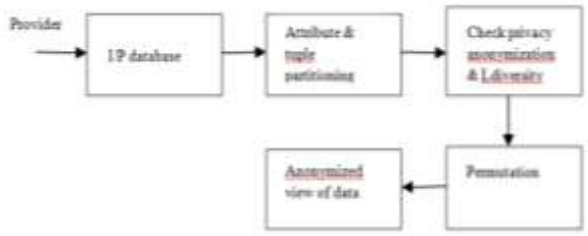

Figure 5: Proposed block architecture

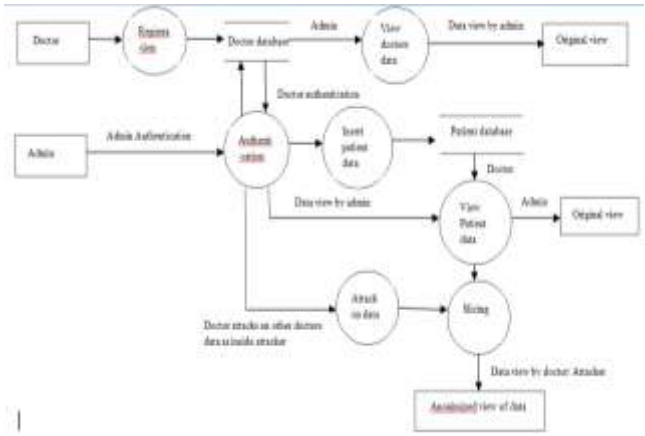

Figure 6: System Flow

Proposed Algorithms

Input: Data set with D, providers n, with C
Output: Slice view (T*) with provider
Step 1: read data from (D up to null)
Step 2: for each (attributes in table)
For each (tupels in tables)
Step 3: set quasi identifier (QIfr) and sensitive attributes (SA)
Step 4: Apply generalization technique it will classify the tupples in QIfr groups
Step 5: Apply anonymization on relative information attributes
Step 6:While(verify data-privacy(D, n, C) = 0) do
if (Di → D) verified with QIfr then
add Di up to when K-anonimty
elseealy stop
Bucket(i1) → D;
Step 7: permute the data with (I=(I(null-1)))
Step 8: Apply Pruning on(D)
Step 9: Apply step 1,2,3 on Becket(i1)
Step 10: if (C fails with (D)&& (p#1))
Bucket(i2) → Bucket(i1(j))
Step 11: Display all (Bucket (i2)!=null)
Step 12: end while
Other Algorithm Use In Proposed System As SQL Injection And Prevention.

INPUT: Query=User Generated Query
SPL[ ]=Static Pattern List with m AnomalyPattern
2: For j = 1 to m do
3: If (AC (Query, String. Length(Query), SPL[j][0]) = =)then
4: Calc anomaly score
5: If ( ) Score Value Anomaly = Threshold
6: then
7: Return Alarm. Administrator
8: Else
9: Return Query. Accepted
10: End If
11: Else
12: Return Query. Rejected
13: End If
14: End For
End Procedure

V. Conclusion

We consider a potential attack on collaborative data publishing. We used slicing algorithm for anonymization and L diversity and verify it for security and privacy by using binary algorithm of data privacy.

This proposed system help to improve the data privacy and security when data is athered from different sources and output should be in collaborative fashion.
Slicing algorithm is very useful when we are using high dimensional data. It divides data in both vertical and horizontal fashion. Due to encryption we can increase security. But the limitation is there could be loss of data utility.

Above system can used in many applications like hospital management system, many industrial areas where we like to protect a sensitive data like salary of employee. Pharmaceutical company where sensitive data may be a combination of ingredients of medicines, in banking sector where sensitive data is account number of customer, our system can use. It can be used in military area where data is gathered from different sources and need to secured that data from each other to maintain privacy.

This proposed system help to improve the data privacy and security when data is gathered from different sources and output should be in collaborative fashion.

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 2009

[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.MohamedRaseen "Enhanced Slicing Models For Preserving Privacy In Data Publication" in International Conference on Current Trends in Engineering and Technology, ICCTET'13

[6] Younho Lee proposed "secure ordered databucketization " Dependable and Secure Computing, IEEE Transactions on (Volume:11 , Issue: 3) in June 2014.

[7] Luong The Dung proposed Privacy Preserving Classification in Two-Dimension Distributed Data in 2010 Second International Conference on Knowledge and Systems Engineering

[8] Tristan Allard, Benjamin Nguyen, Philippe Pucheral proposed Safe Realization of the Generalization Privacy Mechanism 2011 Ninth Annual International Conference on Privacy, Securityand Trust

[9] Jing Yang and Ziyun Liu , yangyue ,Jianpei Zhang A Data Anonymous Method based on Overlapping Slicing in Proceedings of the 2014 IEEE 18th International Conference on Computer Supported Cooperative Work in Design

[10] R. Mahesh and T. Meyyappan proposed Anonymization Technique through Record Elimination to Preserve Privacy of Published Data Proceedings of the 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering.