International Journal of Engineering Inventions
e-ISSN: 2278-7461, p-ISSN: 2319-6491
Volume 3, Issue 12 (July 2014) PP: 60-65

Data attribute security and privacy in Collaborative distributed database Publishing

Ms. Pragati J. Mokadam[1], Dr. S.T.Singh[2]
(PKTC, Department of computer science and engineering, Pune university)

Abstract: *In this era, there are need to secure data in distributed database system. For collaborative data publishing some anonymization techniques are available such as generalization and bucketization. We consider the attack can call as "insider attack" by colluding data providers who may use their own records to infer others records. To protect our database from these types of attacks we used slicing technique for anonymization, as above techniques are not suitable for high dimensional data. It cause loss of data and also they need clear separation of quasi identifier and sensitive database. We consider this threat and make several contributions. First, we introduce a notion of data privacy and used slicing technique which shows that anonymized data satisfies privacy and security of data which classifies data vertically and horizontally. Second, we present verification algorithms which prove the security against number of providers of data and insure high utility and data privacy of anonymized data with efficiency. For experimental result we use the hospital patient datasets and suggest that our slicing approach achieves better or comparable utility and efficiency than baseline algorithms while satisfying data security. Our experiment successfully demonstrates the difference between computation time of encryption algorithm which is used to secure data and our system.*

Keywords: *Distributed database, privacy, protection, security, SMC, TTP*

I. INTRODUCTION

There is an increasing need for sharing data that contain personal information from distributed databases. For example, in the healthcare domain, a national agenda to develop the Nationwide Health Information Network (NHIN) to share information among hospitals and other providers, and support appropriate use of health information beyond direct patient care with privacy protection. Privacy preserving data analysis, and data publishing have received considerable attention in recent years as promising approaches for sharing data while preserving individual privacy.

1.1 Idea and Motivation

In this era large amount of data are getting shared throughout world. So privacy preserving data publishing has been studied extensively in recent years. This data may contain private data like personal information of any person, household or an organization that should not be disclosed to other have to protect. To maintain the security some privacy preservation techniques are available. Privacy preservation techniques are mainly used to reduce the leakage of formation. Now a days for anonymization of data from multiple data providers by using techniques like generalization for k anonymity and bucketization for maintaining L diversity are getting studied. In these techniques, the information have to protect can call as a sensitive attribute (SA) i.e. disease of patient, salary of employee etc. The second part of any database is identifier (ID) i.e. name and third is quasi identifier (QI) i.e. age, zip code etc. But these techniques have some disadvantages like data lose , membership disclosure etc. To avoid this we studied slicing technique.

1.2 Existing work

In this section we are presenting the different methods which are previously used for anonymization. We discuss some advantages and limitation of these systems.

Traditional k anonymization model used generalization technique. But the limitation is, this model causes a loss of data in high dimensional database system. Beyond this LKC[3] privacy model for high dimensional relational data gives better result than traditional k anonymization model. But LKC model consider only relational data and healthcare data is complex, may be a combination of relational data, transaction data and textual data. But only k anonymity is not sufficient to preserve privacy. Hence this two party protocol DPP2GA helps in it but major disadvantages of DPP2GA[4] is it may not produce a precise data when data are not partitioned. It is only privacy preserving protocol not SMC because it introduces certain inference problem. Then there are DkA[5] model which is proven to generate k anonymous dataset and satisfying security definition of SMC. But DkA is not a multiparty framework. Beyond this when attacker attacks on anonymized system with the help of BK(background knowledge) all above systems will fail. L diversity[6] helps to overcome this problem.

In further paper we introduce a new system which is depends on slicing technology [10] which improves result than generalization and bucketization technique. This is very useful technique for high dimensional data. But there could be a loss of data utility.

II. PROBLEM DEFINITION AND SCOPE

Due to different attacks attackers can attack on our system. For our system we consider certain insider attack like background knowledge attack. Privacy protection is impossible due to the presence of the adversary's background knowledge [6]. Second is linkage attack in which when an adversary is able to link a record owner to a record in a published data table called record linkage, to a sensitive attribute in a published data table called attribute linkage, or to the published data table itself called table linkage. In this attack adversary may know some victims data like QID etc. In some cases provider himself can be an attacker. His own record which might be a subset of database. Maintaining security and privacy of document without using encryption have been a challenging problem in distributed network. Various methods and strategies are developed to make maximum probability to make it possible. To overcome these problems we proposed a system.

Problem definition: Our main goal is to publish an anonymized view of integrated data, P* which will be immune to attacks. We improve the security and privacy with the help of slicing technique, data privacy verification algorithm and secure data analysis with the help of classifier.

2.1 Goals and objective

- We use slicing algorithm which gives better result than generalization and bucketisation
- Binary algorithm use to verify data privacy for every section of data by using pruning strategy.
- Check EG monotonic for checking the privacy of equivalent group.
- We have to decrease computation time of system.

2.2 Scope

- Proposed system is run on LAN network.
- Distributed system like hospital patient data management, companies employers salary system, banking system like personal information of account holders etc where we need to secure collaborative data.

III. PRIVACY PRESERVATION OF DATA

To protect data from external recipients with certain background knowledge BK, we assume a given privacy requirement C is defined as a conjunction of privacy constraints: C1∧C2∧...∧Cw. If a group of anonymized records T* satisfies C, we say C(T*)=true. By definition C(Ø) is true and Ø is private. Any of the existing privacy principles can be used as a component constraint Ci. We now formally define a notion of data-privacy with respect to a privacy constraint C, to protect the anonymized data against data-adversaries. The notion explicitly models the inherent data knowledge of an data-adversary, the data records they jointly contribute, and requires that each QI group, excluding any of those records owned by an data-adversary, still satisfies C.

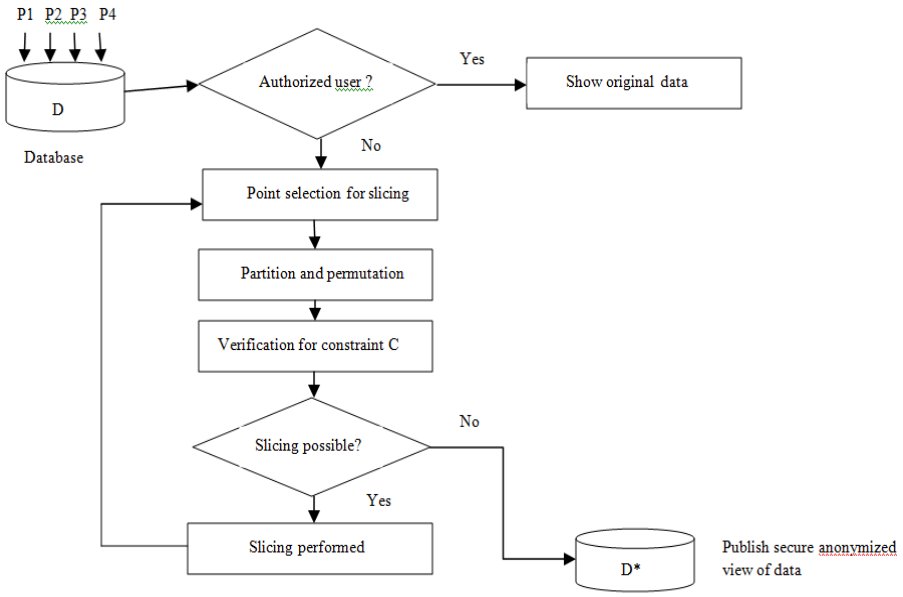


Fig 1: Proposed system

Fig.1 shows our proposed system in which input data is given from different provider. Select point for slicing. Check that input data against privacy constraint C for data privacy Check further is slicing is possible or not. If it possible for input data perform slicing. Our final output are also anonymized data(D*). Any adversary(coalition of data users with data providers cooperating to breach privacy of anonymized records) cannot breach privacy of data. In this system we are using horizontal as well as vertical partitioning over database.

IV. FLOW OF SYSTEM

4.1 Mathematical flow of proposed system:

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

4.2 Proposed system algorithms:

1. Anonymization by slicing:

Slicing is basically depends on Attribute and tuple partitioning. In Attribute partitioning I partitioned data as {name},{age-zip} and {Disease} and tuple partitioning as {t1,t2,t3,t4,t5,t6}. In attribute partitioning age and zip are partitioned together because they both are highly correlated because they are quasi identifiers (QI). These QI can be known to attacker. While tuple partitioning system should check L diversity for the sensitive attribute (SA) column. Algorithm runs are as follows.

1. Initialize bucket k=m, int i= rowcount, column count=C, Q={D}, // D= data into database, Arraylist= a[i];
2. While Q is not empty
   If i<=m
   Check l diversity;
   Else
   i++;
3. Q=Q-{D*+a[i]};
4. Repeat step 2 and 3 with next tuple in Q
5. D*=D*U A[D] // next anonymized view of data D

2. L diversity:

L diversity is the concept of maintaining uniqueness within data. In this system I used this concept on SA i.e on disease. Our anonymized bucket size is 6 and I maintain L=4 i.e from 6 disease record 4 must be unique. (Algo)

1. Initialize L=n, int i;
2. If i= n-m+1;
   Then a[0].....a[1], insert these values as they are in Q;
   i++;
3. Else
   Check privacy constraint for every incremented value in Q
   If
   L=n then
   Fscore=1
   Insert value in the row
   i++;
   else
   Add element to arraylist a[i];
4. Exit

3. Permutation:

Permutation means rearrangement of records of data. In my project. I used permutation process for rearrangement of quasi identifier i.e {Zip-Age}

4. Fscore:

Fscore is privacy fitness score i.e the level of fulfillment of privacy constraint C. If fscore=1 then C(D*)= true.

5. Constraint C:

C is a privacy constraint in which D* should fulfill slicing condition with L diversity as explain above. Consider value of L diversity is 4. Fscore should be 1 when system fulfills L diversity condition.

Some verification processes are carried out.

1) Verification for L diversity: For verification of L diversity I used Fitness score function.
   1. Generate fake values of SA
   2. Check for privacy constaint and fscore=1;
   3. If
      Privacy breach;
      Then early stop;
      Else
      Return (Fscore);
   4. Exit

2) Verification for strength of system against number of provider: For verification against number of provider, add one more attribute in anonymized data as a provider.
   1. Generate fake values of SA by providers P= 1.....n
   2. Check for privacy constraint and Fscore=1 with respect to number of provider If
      Privacy breach;
      Then early stop;
      Else
      Return(Fscore);
   3. Exit

By using above algorithm we can obtain anonymization and l diversity both. This two technique maintains the privacy of data.

4.3 Result:

Table I shows the input to the system.

TABLE I

Name	Age	Zip code	Disease
Alice	22	12311	Cancer
Mark	24	12334	Epilepsy
Sara	25	12365	Flu
Bob	43	23411	High BP
Marry	31	23433	High Bp
Frank	33	23433	Attack

John	15	12311	Flu
Siyara	19	23411	Epilepsy
Rean	32	23412	BP
Margaret	70	12365	Cancer
Elizabeth	68	23433	Attack
Soman	53	12365	Attack

By using our proposed slicing algorithm we get result as in Table II. This maintain 6-anonymization with quasi identifier QI {Age, Zipcode} and 4-diversity with disease(SA). Slicing performs permutations within buckets on. This system protect our database from attacks like background knowledge, linkage attack as this result provide no linking between tuples and attribute. Attacker can not breach this security and privacy of sensitive data.

TABLE II

Name	{Age, Zip code}	Disease
********	{33,23433}	Cancer
********	{31,23433}	Epilepsy
*********	{43,23411}	Flu
********	{25,12365}	High BP
********	{24,12334}	High Bp
********	{22,12311}	Attack
*********	{53,12365}	Flu
*********	{68,23433}	Epilepsy
*********	{70,12365}	BP
**********	{32,23412}	Cancer
*********	{19,23411}	Attack
*********	{15,12311}	Attack

4.3.1 Experiment:

For study the performance of our system with another existing systems we consider above input data and following s/w and h/w specification.

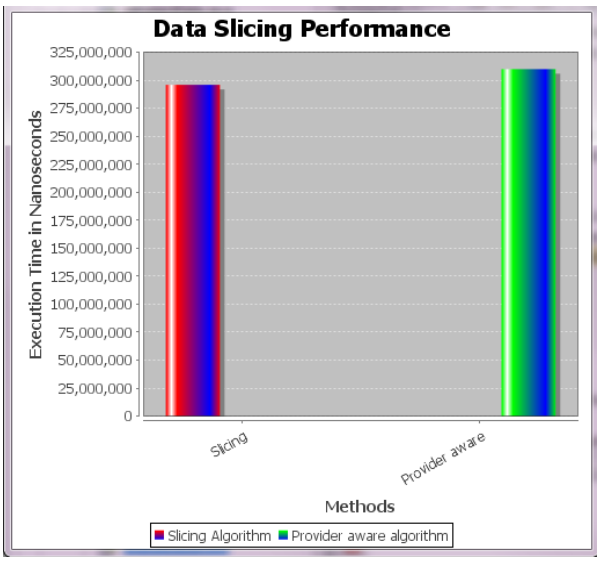
Hardware specification :-

i) Processor - Pentium –IV ii) speed 1.1 Ghz iii) RAM- 256 MB(min) iv)hard disk-20 GB(min)

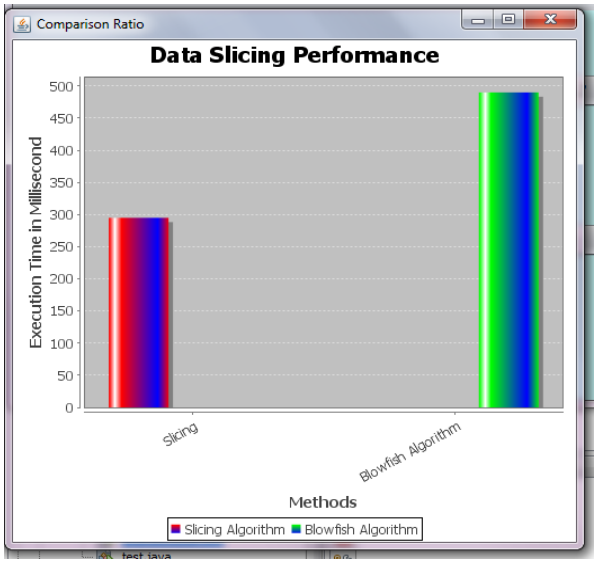
Software specification :-

i) Operating system- windows 7/8 ii) Programming language- Java iii) Tool- Net beans

Graph 1 shows the performance of anonimization with slicing shown by red bar and provider aware algorithm[1] shown by green bar. Graph 2 shows the performance difference between anonimization with slicing shown by red bar and blowfish encryption algorithm shown by green bar.



Graph 1: Slicing with provider aware algorithm Performance time in nanosecond



Graph 2: Slicing with Blowfish encryption algorithm performance time in millisecond

Performance of the system are measure in computation time of system. This performance is depends on code, platform, software and hardware specification used to sun the system. For above result we run slicing algorithm, blowfish encryption algorithm and provider aware algorithm on above specified experimental setup.

V. CONCLUSION AND FUTURE SCOPE

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 verification algorithm of data privacy. Slicing algorithm is very useful when we are using high dimensional data. It divides data in both vertical and horizontal fashion. Time computation of slicing is very less as compare to blowfish encryption algorithm. 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 balance of customer, our system can use..

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. In future this system can consider for data which are distributed in ad hoc grid computing. Also the system can be considered for set valued data. Slicing can be performed at insertion time.

REFERENCES

[1] S. Goryczka, L. Xiong, and B. C. M. Fung, "m-Privacy for collaborative data publishing," IEEE transaction on knowledge and data engineering 2013

[2] S. Goryczka, L. Xiong, and B. C. M. Fung, "m-Privacy for collaborative data publishing," in Proc. of the 7th Intl. Conf. on Collaborative Computing: Networking, Applications and Worksharing, 2011.

[3] C.Dwork,"Differential privacy: A survey of results", in Proc. of the 5th Intl. Conf. on Theory and Applications of Models of Computation, 2008, pp. 1–19.

[4] N. Mohammed, B. C. M. Fung, P. C. K. Hung, and C. Lee, "Centralized and distributed anonymization for high-dimensional healthcare data," ACM Trans. on Knowledge Discovery from Data, vol. 4, no. 4, pp. 18:1–18:33, October 2010.

[5] W. Jiang and C. Clifton, "Privacy-preserving distributed k-anonymity,"in DBSec, vol. 3654, 2005, pp. 924–924.

[6] W. Jiang and C. Clifton, "A secure distributed framework for achieving k-anonymity," The VLDB Journal Special Issue on Privacy Preserving Data Management, vol. 15, no. 4, pp. 316–333, 2006

[7] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam,"l-Diversity: Privacy beyond k-anonymity," in ICDE, 2006,p. 24

[8] R. Sheikh, B. Kumar, and D. K. Mishra, "A distributed k-secure sum protocol for secure multi-party computations," J. of Computing, vol. 2, pp. 68–72, March 2010 (2002)

[9] C. Dwork, "A firm foundation for private data analysis," Commun. ACM, vol. 54, pp. 86–95, January 2011

[10] P. Jurczyk and L. Xiong, " Distributed anonymization: Achieving privacy for both data subjects and data providers," in DBSec, 2009, pp. 191–207

[11] C. M. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," ACM Comput. Surv., vol. 42, pp. 14:1–14:53, June 2010.

[12] Tiancheng Li, Ninghui Li, Jian Zhang,Ian Molloy,"Slicing: A New Approach for Privacy Preserving Data Publishing" IEEE transactions on knowledge and data engineering, vol. 24, no. 3, March 2012

[13] O. Goldreich, Foundations of Cryptography: Volume 2, 2004