IJCTA, 9(28) 2016, pp. 49-55

CRYPTOHASHING: AN EFFICIENT PRIVACY PRESERVING DATA PUBLISHING

Nithya. M*, and Sheela.T*****

Abstract : Preserving Privacy and retaining accuracy are important requirements in the field of data mining and publishing. Privacy preserving focuses on breaking down associations between data attributes, meanwhile accuracy expects relationship intact between attributes. Achieving balance between privacy and accuracy without compromise of either of them is always a challenge

Methods / Analysis: Breaking down of association between attributes involves high data utility loss, expensive and will be difficult to reconstruct the original data. PPDM techniques like data partitioning, data modification, data restriction are common techniques aiming to achieve balance between these 2 factors. Unfortunately they have their own limitation in handling privacy and accuracy. This paper explains a new technique called Hashing which handles privacy and accuracy factors effectively.

Findings: This new Hashing technique looks promising as it offers better privacy and accuracy of data. Hashing algorithm is derived and realized using ORANGE tool. **Novelty /Improvement:** Hashing algorithm brings relative significance on privacy on data when compared to accuracy as reconstruction of original data post modification is quite a challenge. This algorithm leaves room for improving the data reconstruction part. Experimental results are analyzed to justify the performance of Hashing technique.

Keywords: Privacy;Accuracy; Hashing; Modification.

1. INTRODUCTION

Data mining and publishing involves collection of large scale raw data and converting them into beneficial information. Advanced techniques like artificial intelligence, neural networks and statistics are used in data mining to arrive at data extraction models which could be in the form of rules, patterns and decision trees. These models are directly related to individual privacy of humans. PPDM ensures individual privacy is not breached while still able to extract useful information from the data extraction models. Predictive rules and techniques[1] can help in predicting privacy information easily. Thus data anonymization becomes a requirement to avoid sensitive data leakage. Anonymization techniques like Generalization[2,3], Bucketization[4,5,6] and Slicing are well known which handle data anonymization in their own way. In general, these techniques manage in manipulating the original data to avoid sensitive data made available for data analysts. In this course of data manipulation, there are always possibilities of data utilization going down. Utilization loss becoming predominant shall directly affect the accuracy of data analysis. In few occasions the analysis results go completely wrong finally unable to solve the very purpose of data mining and publishing. In general there is a strong assumption that privacy and accuracy

* Research scholar, Faculty of computer science and engineering, Sathyabama University, Chennai-600119,Tamil Nadu, India - 600119, Tamil Nadu, India; Email: nithya.cse@sairam.edu.in

** Assistant Professor, Computer Science and Engineering Sri Sairam Engineering College, Chennai - 600044, Tamil Nadu, India

*** Professor & HOD, Dept of IT, Sri Sairam Engineering College, Chennai, INDIA

are trade off features[7], practically impossible to achieve both. Other techniques like Cryptography methods are relatively better as they tweak the source data to an extent to maintain data privacy and accuracy. Methods like noise additions and space transformation are well known among them. This paper explains a new Hashing method which offers better privacy and accuracy of data. Open source ORANGE data mining tool is used to design this algorithm. Initial part of this paper will detail on merits and demerits of conventional cryptographically techniques.

2. Data Analysis

Any source data shall have identifiers which can uniquely identify individual (Name, SSO), Quasi Identifiers (Age, Sex) which are available for the analyst and finally sensitive data (Disease, Salary) whose privacy need to be secured. Medical records from Hospital, salary records from Company are considered as sensitive data which are prone to security issues and attacks[8]. These data when leaked out could be a threat to individual privacy. These data could be of any data type, volume and size. The data source can be manipulated with certain level of privacy maintained and released to certain group of people. In parallel another group of people might receive manipulated data with different degree of privacy. If both the data variants are somehow accessible by an intruder then there is always a possibility to compare both the data variants and exploit the privacy factor. Further data analysis results should always respect analysis requirement. In few occasions maintaining data privacy is expected than accuracy of data. In other cases accuracy of data is mandate. Thus data publishing technique should be flexible for generating reports as per need.

3. INSPIRATION FOR HASHING

Cryptographic techniques on secured multiparty computation is explained[9] with the help of two billionaire's who which to understand the richest person among the two without revealing each other's wealth data. This infact gave an insight to compare two different data sources to get useful information out of it. Further it was demonstrated that any problem which can be described by a polynomial size boolean circuit of logarithmic depth[10] can be solved securely. The level of privacy to be maintained is proportional to the intensity of encryption applied on the source data. The intensity of encryption can be referred as protocol which determines the security level of the multiplayers involved in data sharing. Here 2 or more players share their data to a third party protocol which will finally publish the desired output which are defined and agreed by all the players. This formulation is being followed by most of the secured computations[11]. Goldwasser-Micali[12] (GM) cryptosystem, which is the first homomorphic cryptosystem, falls under public key encryption methods. This method involved in exhaustive message expansion during encryption resulting in becoming unusable for data mining. Benaloh[13] cryptosystem was the successor of the Goldwasser-Micali (GM) cryptosystem. Although this method was better than the earlier one, it was not an efficient method. Paillier[14] cryptosystem was proposed to avoid the drawbacks in the earlier homomorphic cryptosystem. The Paillier cryptosystem houses speedy encryption and decryption algorithms, encrypting 1024-bit messages in ciphertexts of at least 2048-bits.

Collaborative similarity measure approach protocol is used in[15] with lightweight overhead. An efficient aggregation operator fused into advanced encryption algorithm is discussed in[16]. Zero data and query privacy leakage is achieved in Efficient Conjunctive Query (ECQ) scheme in[17]. By linking the benefits of RSA public key cryptosystem and homomorphism encryption scheme, a model of hierarchical management on the cryptogram is derived in[18]. Similarly a secure k-means data mining approach is proposed in[19] which offers better efficiency. Normalization techniques[20] are also used to achieve privacy preserving in data mining.

3.1 Hashing Technique

Methodology

Considering the drawbacks of earlier cryptographic methods there is a need to create an improvement in publishing technique. Hashing technique is designed considering the above drawbacks. A source table with Place, Gender, Age, Disease information is used for demonstrating Hashing method. Since disease attribute is a sensitive factor it needs to be handled carefully to avoid leakage of privacy. To improve data utility correlated attributes (age and gender) are grouped together. Similarly (Place and disease) attributes are grouped together. The data owner generates dependent hash keys and retains them. The data owner performs hashing encryption exclusively for the two groups and sends the encrypted data to the data analyst. The data which is sent may be vulnerable to hacking before it reaches the data analyst. Since the data is hashed, the hackers may not be in a position to merge the 2 groups and find the sensitive data. Once the data reaches the data analyst, dependent hash keys are sent separately to him from the data owner. With the help of dependent hash keys the data analyst can decode the encryption and reconstruct the original table without losing accuracy. In this process privacy of data is also retained as the hackers could not hack the source data during the data transfer.

3.2 Algorithm

Encryptor 1:

```
import Orange
from random import randint
import hashlib
data = Orange.data.Table (in_data)
age, gender, hash_one = [Orange.feature.String(x) for x in [“Age”,”Gender”,”Hash_One”]]
Domain = Orange.data.Domain ([age,gender,hash_one])
out_data = Orange.data.Table (Domain)
print “%-15s %-15s %s” % (“Age”, “Gender”,”Hash_One”)
for i in range(len(data)):
hash_one = hashlib.sha224 (str(randint(1000,10000))).hexdigest()
print “%-15s %-15s %s” % (data[i][“age”],data[i][“gender”], hash_one)
out_data.append([str(data[i][‘age’]) ,str(data[i][‘gender’]),hash_one])
```

Encryptor 2:

```
import Orange
from random import randint
import hashlib
data = Orange.data.Table (in_data)
place, disease, hash_two = [Orange.feature.String(x) for x in [“Place”,”Disease”,”Hash_Two”]]
Domain = Orange.data.Domain ([place,disease,hash_two])
```

```
out_data = Orange.data.Table (Domain)
print (“Place”, “Disease”,”Hash_Two”)
for i in range (len (data)):
hash_two = hashlib.sha224 (str (randint (1000, 10000))).hexdigest ()
print (data[i][“place”],data[i][“disease”],hash_two)
out_data.append([str(data[i][‘place’]) ,str(data[i][‘disease’]),hash_two])
```

Shuffler:

```
import Orange
out_data = Orange.data.Table (in_data)
out_data.shuffle ()
```

3.3 Working Procedure

Step 1: Source data table to be published is first classified into identifiers, quasi identifiers & sensitive attributes. Since attribute “disease” is a confidential data for an individual which can reveal personal information when linked with other attributes it is considered as sensitive data.

Step 2: Hash_One is created using random number generation and reiterated to prevent cracking.

Step 3: Hash_Two is created using Hash_One and reiterated to prevent cracking.

Step 4: The source data table is next divided into columns. This division brings certain quasi identifiers together on one side (vertical 1) and the other with a combination of quasi identifier and sensitive attribute (vertical 2).

Step 5: Vertical 1 is hashed with Hash_One key.

Step 6: Vertical 2 is hashed with Hash_Two key.

Step 7: Vertical 1 with Hash_One key is random shuffled.

Step 8: Vertical 2 with Hash_Two key is random shuffled.

3.4 Experimental Analysis

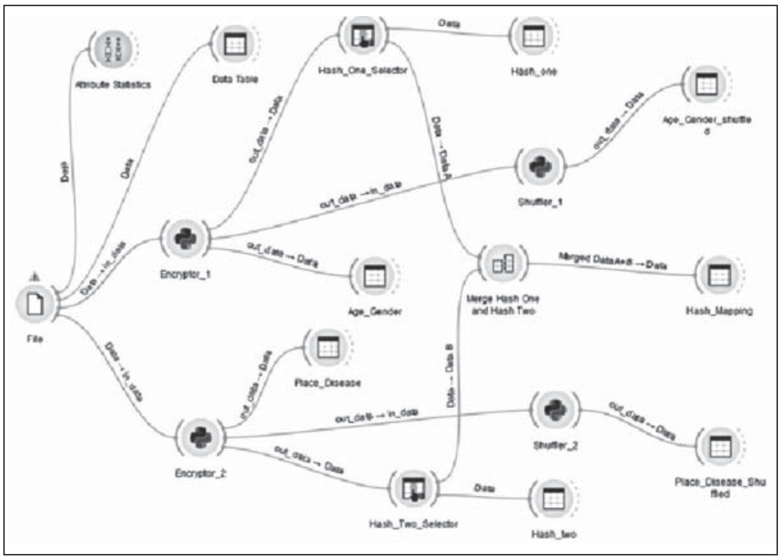

Figure 1. Implementation Schema

Figure 1 shows the implementation of hashing algorithm using ORANGE data mining tool. Encryptor_1 and Encryptor_2 are used to encrypt (age, gender) and (Place, disease) attributes respectively. Shuffler_1 and Shuffler_2 are used to shuffle the hashed and encrypted data tables.

hash_two	hash_one
2fe10492280608c5573e9de4397fa178eed89532ef6a90bb532528557d359e51bc...	884d247c6f65a96a7da4d1105d584ddd63d4f5773e932c635aa50947814af0952e8aaa3e1.
5658eb3bf03461888d599d8a2799e7a03c81554cadb7549da172b706b769a5216...	c622c085c04eadc473f08541b255320ebd250711e7559a484b90d187a34810ba92c3c41de
20877befbd58c865e224346e6b92d7727f3077498ca8dca2c2f991c94531594497...	f1b6fac213a8baf8b79472063996 5fde95c754ce6d4c1d43ff6ea509fa5694b1d47ba73e18.
71029625d9350f1fe3ce2c6f2211b69240883b2b2bed501ba260864d6ad8821cc...	0d87a1c0ff1500e3c4febc7582848dc2ab038cc1174e981bebb6cc8af1759b9cbb68b9a26.
ea0f827fe45c91c956bacfd78d91d47ba483c7786110b8982cffe5f2ebb00145899...	4851703a0471110c71ec7f7c8fdc073faddbb03e0a580e1f4796e890a5e3be8643aa01f610.
a813f6133031aa997c1223aca7ef7cf8fe6f02dc5f44961e9a83b24c6c62418e6f29...	36203d7da31576b98485dc648ee525e2434b708211268b29d893b21729a79a42ac81a1c2.
ead97089aae476d362a942d978947c32e1fd414956aec1b0a86454eb4c1c12b73...	fb6c4e0b4b90ebfb5a35ca7a9cbf1d16d580a6162e7a8e3d08bd764584121a8259794fe3d
8cad7770bea867c44a6cd63bae19903d1c2adbdb610a4ec06244c468a7eb59d7...	9dc69f4a9d78d28dc1ba5697a159c546cc96accc463558a72fc997caeedc90d9f2eadb2c2.
3b922393a3cb462762ff5dd2bec1b20fb396dd630343d2b747263d2767e935847...	0c5534f554a26f7aeb7c780e12bb152568c6e2ec7a330d601bf619acf67d8fc01305f63e11e
6e3adb1ae0e02c934766182313b6775d67ed8f6b41c1c98850ce95c9d24811f77...	4e2545f819e67f0615003dd7e04a608769a9279e4885a267a39075c47cce507cf2aa03e6bd
793d8e745d2b346c4ddc27a53408324383c9cce8f35cf1a6cf7399cad36c91c99e...	47c917b09f2bc64b2916c0824c71592363bbcaf6d23d8037b7065634ade3856104f5ee26d
c254e7753095807e1cca159e48eceb213b7ef29d4c21d369f4ddbf6cc818ecec59...	e13748298cfb23c19fdfd134a2221e7ba8d469d5a42ff66c78a7e54d9eb6b0c269683fdb17
843f61d9c31569e3ecdc8ffaf5a4accadb6633da62601955f63411ffa247b446a07...	67179258750202 8b6cd4be7c4d662d089c3e0e94403ac07ea1316cfaa4ea80feaa7ced167.
9d53b7a44f7aea7ef05b4bc3c1e37d09f1a25d97850646a0aa61b2a5aedebba77f...	808e53023ea4a8a9d6ecbc1290580f72388611985651c58a3566baf82aa8b35834d5360bd.
2d36b5821f8affc6868b59dfc9af6c9f458b261598c075888c37459a5470a0f723b...	01cbec07301846508 6c9752e6508e0ec27ad20032f5bf485ed3bd4011caca1ea998ea2c54.
faa453efde4ac6a36849ba381feb9e87c3b4930783eaaced9509aa44037c5eff246...	f6a8dd1c954c8506aadc764cc32b895ef2d102148f5897d09b3d73f7f05fd752732f0d2f285
9d7f6fe926fac1ccc6c2cc32c94d82385b159c824f3419618d22011437ed2dc018...	ebf12cb74e96e67e63783d93c534ef27e52bed6fe00a98191abc4d0631cfc8ff246ac62953f
0e19a8bac63f97a513063dcb9a64442ba0c64ee63586476e10fe3be818d242a32f...	ac52c626afc10d4075708ac4c778ddfc8956c3e6f251f3d14b7e4acf3cd14ac5d6df62a942.
4469eee129fe5da0d3edce5404418f598a9ede938eaeb195f36b72304d4c21334c...	a6b964c0bb675116a15ef1325b01ff45bfbfe23a4ac617437c066147085deea62916cbc537
a4bc254def844da9f771cd03eb0cb6d4aa05aa1373fa4de143db513044b65f57a...	d7da8f0d99e5ed3ac7ab00568ae369151771fa59c930ca0a37335569539bce0e096a772b0.

Figure 2. Hash Table

Figure 2 shows the dependent hash table generated. Hash_One is random generated and Hash_Two is generated based on Hash_One. Further the two hash keys are used to encrypt vertical 1 and 2 tables respectively. These two tables are sent to the data analyst. Since they are encrypted it will be impossible for the hacker to hack the original data. Once the tables reach the data analyst, he can combine the tables using the dependent hash table which is separately sent to him. To validate the accuracy of the data received, an accuracy finder algorithm is used. This algorithm compares the original data table and the reconstructed data table received in the data analyst end.

Accuracy finder Algorithm

1 Compare the hash key in both tables and rearranges them according to key-value pair in hash_values table
2. Hash_one and hash_two are the hash entries from hash_table
3. Count is the number of entries in table and Count1 is the number of entries that match with hash keys
4. Count - Count1 is the number of invalid entries
5. Iteration to move through all entries in table
6. Increment Count variable count + = 1
7. Iteration to move through all entries
8. Increment Count1 if hash_keys match count1 += 1
9. End of Iteration
10. valid_entries gives the total number of valid entries evaluated
11. valid_entries = 0 implies that there is no missing of entries
12. invalid_entries = count - count 1
13. if invalid_entries > 0
14. if there is missing entries, find the error percentage error_percent = (count1 / count) * 100

4. EXPERIMENTAL RESULTS

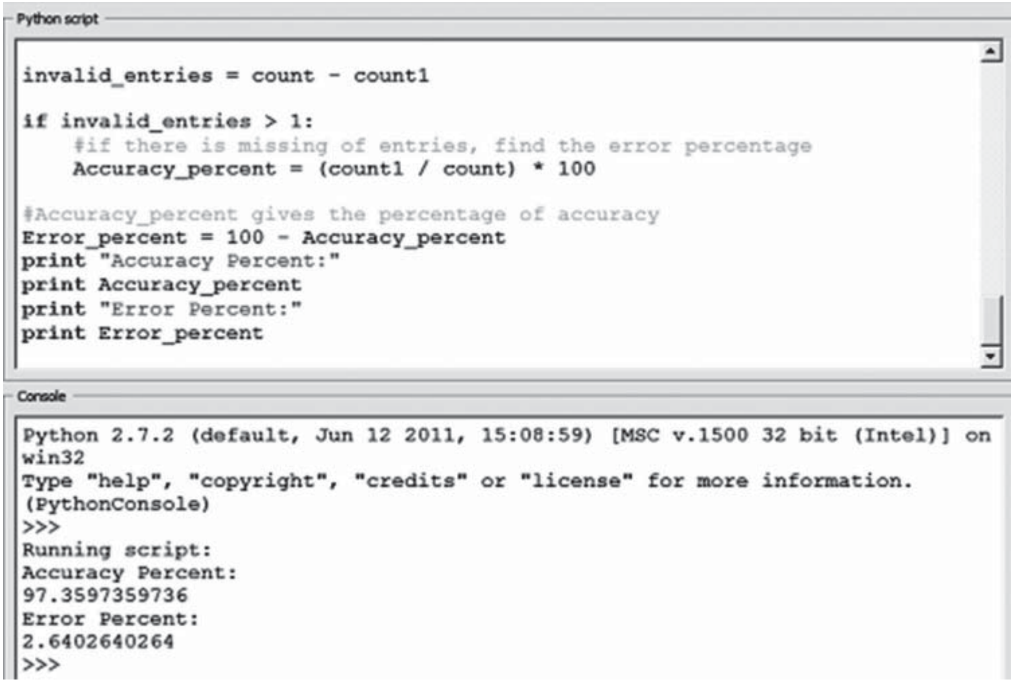

Figure 3. Accuracy Results

Figure 3 shows the results of accuracy finder comparing the source table and the reconstructed table. It is evident that accuracy is 97.3 % which is promising when compared to other cryptographic methods. From the experimental results it is clear that the hash encrypting is performing well in terms of efficiency and cost. Time taken for reconstructing the source data is negligible as the program takes 7.4 seconds for reconstructing 1920 records.

5. CONCLUSIONS

This new Crypto Hashing technique looks promising as it offers better privacy and accuracy of data in the field of data mining and publishing. Considering the efficiency, cost and run time factors this technique can be used for managing large amount of data. Memory required for encrypting is also limited as the hash keys occupy less data space. This technique offers privacy and accuracy during the phase of data transmission from the data owner to data analyst. It seals the source data from the intermediate hackers. Care should be taken once the data reaches the data analyst.

References

1. Vimala, R. Shakouri & Hooman, “Integrating association rules to predict retinopathy”, Maejo Int. J. Sci. Technol. 2012, 6(03), 334-343.
2. P. Samarati, “Protecting Respondent’s Privacy in Microdata Release,” IEEE Trans. Knowledge & Data Eng., 2001, vol. 13, pp. 1010-1027.
3. L. Sweeney, “k-Anonymity: A Model for Protecting Privacy,” Int’l J. Knowledge-Based Systems, 2002, vol. 10, pp. 557-570.
4. X. Xiao and Y. Tao, “Anatomy: Simple and Effective Privacy Preservation,” Proc. Int’l Conf. Very Large Data Bases, 2006, pp. 139-150.
5. D.J. Martin & J.Y. Halpern, “Worst-Case Background Knowledge for Data Publishing,” Int’l Conf. Data Eng., 2007, pp. 126-135.

6. N. Koudas, Yu & Q. Zhang, "Aggregate Query Answering on Anonymized Tables," Int'l Conf. Data Eng., 2007, pp. 116-125.

7. J. Brickell and V. Shmatikov. The cost of privacy: destruction of data-mining utility in anonymized data publishing. KDD, 2008, pages 70–78.

8. Phuwanai & Sangsuree, "Privacy-preserving emergency access control for health records", Maejo Int. J. Sci. Technol. 2015, 108-120.

9. A. C. Yao, "Protocols for secure computations" (extended abstract). In 23rd Annual Symposium on Foundations of Computer Science. IEEE, 1982.

10. A. C. Yao, "How to generate and exchange secrets". In Proceedings of the twenty-seventh annual IEEE Symposium on Foundations of Computer Science, 1986, pages 162–167. IEEE Computer Society.

11. O. Goldreich. The Foundations of Cryptography —Basic Applications. Cambridge University Press, May 2004, Volume 2.

12. S. Goldwasser and S. Micali. Probabilistic encryption. J. COMP. SYST. SCI., March 1984, 28(2):270–299.

13. J. Vijayan. House committee chair wants info on cancelled dhs data-mining programs. Computer World, September 18, 2007.

14. P. Paillier. Public-key cryptosystems based on composite degree residuosity classes. In Advances in Cryptology EUROCRYPT'99, LNCS 1592, 1999, pages 223–238. Springer.

15. Kikuchi, H., Aoki, Y; Terada, M. ; Ishii, K., 2012, Accuracy of Privacy-Preserving Collaborative Filtering Based on Quasihomomorphic Similarity, 9th International Conference on Ubiquitous Intelligence & Computing and 9th International Conference on Autonomic & Trusted Computing (UIC/ATC), 2012, pp: 555- 562

16. Shu Qin Ren, Khin Mi Mi Aung; Jong Sou Park, 2010, A Privacy Enhanced Data Aggregation Model, Computer and Information Technology (CIT), 2010 IEEE 10th International Conference on, pp: 985 – 990

17. Mi Wen, Rongxing Lu; Jingshen Lei; Xiaohui Liang, 2013, ECQ: An Efficient Conjunctive Query scheme over encrypted multidimensional data in smart grid, Global Communications Conference (GLOBECOM), 2013 IEEE, 796 – 801

18. Gui Qiong, Cheng Xiao-hui, 2009, A Privacy Preserving Distributed Method for Mining Association Rules Artificial Intelligence and Computational Intelligence, 2009. AICI '09. International Conference on Volume: 4 DOI: 10.1109/AICI.2009. 486 pp: 294 – 297

19. Mittal, D.; Kaur, D.; Aggarwal, A., 2014, Secure Data Mining in Cloud Using Homomorphic Encryption IEEE International Conference on Cloud Computing in Emerging Markets (CCEM), 2014, pp: 1 – 7

20. G. Manikandan, N. Sairam, S. Sharmili, S. Venkatakrishnan, "Achieving Privacy in Data Mining using Normalization", Indian Journal of Science and Technology,2013 Apr, 6(4), Doi no: 10.17485/ijst/2013/v6i4/31852