International Journal of Computer Science and Network (IJCSN)
Volume 1, Issue 6, October 2012 www.ijcsn.org ISSN 2277-5420

Review Paper on Preserving Confidentiality of Data in Cloud Using Dynamic Anonymization

[1] Bhushan Mahajan, [2] Swati Ganar

[1] Department of Computer Science and Engineering
G.H.Raisoni College of Engineering, Nagpur

[2] Department of Computer Science and Engineering
G.H.Raisoni College of Engineering, Nagpur

Abstract

Cloud computing is a model that enables Convenient and On-demand network access to a shared pool of configurable computing resources where millions of users share an infrastructure. Security and Privacy concerns are significant obstacle that is preventing the extensive adoption of the public cloud in the Industry. Multi-tenancy where multiple tenants share cloud infrastructure poses an additional concern about the deliberate or accidental exposure of data. Data Anonymization makes data worthless to anyone except the owner of the data. It is one of the methods for transforming the data in such a way that it prevents identification of key information from an unauthorized person. Data can also be anonymized by using techniques such as, Hashing, Hiding, and Shifting etc. The proposed system uses novel model of security i.e k-anonymity to improve data anonymization. It uses dynamic anonymization technique, key distribution mechanism to preserve confidentiality of cloud data.This paper gives literature survey of few methods that have been applied to preserve privacy for static and dynamic anonymization.

Keywords— *Anonymization, k-anonymity, l-diversity, t-closeness*

1. Introduction

Private organizations publish their data on to the cloud for some research or other purpose. The confidentiality of this data must be preserved. i.e. any sensitive information should not be disclosed. Data anonymization is one of the privacy preserving techniques that translate the information, making the data worthless to anybody except the owners [1]. It is different from that of data encryption. The data also called microdata is stored in a table which has multiple records. These records may be categorized as explicit identifiers, quasi identifiers and sensitive identifiers. Explicit identifiers are the attributes which identifies an individual. For eg: Name, social security number etc. Quasi identifiers are the attributes which can be linked with other information to identify an individual. For eg: gender, birth-date etc. And sensitive identifier is the attribute with sensitive value.

When releasing a dataset, it is necessary to prevent the sensitive information of individual from being disclosed. There are 2 types of disclosure: Identity disclosure and Attribute disclosure. Identity disclosure takes place when an individual is linked to a particular record in the released microdata table. Whereas, Attribute disclosure occurs when new information about some individuals is revealed [2].

In order to protect the sensitive values, Generalization [3] techniques can be used. This technique replaces quasi identifier attributes with less specific values. The data in microdata table is generalized using anonymization K-Anoymization technique. To effectively limit information disclosure, it is necessary to measure the disclosure risk of anonymized table.

Most of the anonymization work has been done on static datasets. But, the real datasets are dynamic. So, dynamic anonymization is required. The dynamic datasets are made complex by using data updates. Following hierarchy shows the different techniques that are used for anonymization.

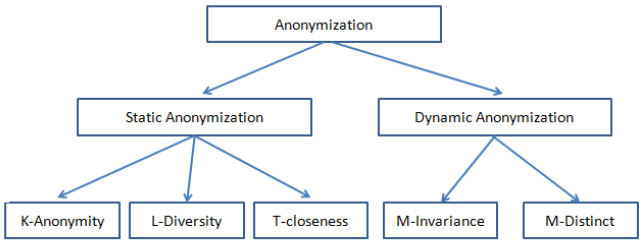

Fig. 1: Anonymization techniques

Data updates can be either external or internal [10]. External update leads to update of records in dataset.

Whereas internal update leads to update of records attribute value. There is always a correlation between old value and new value of a record. For example, a person's current salary in one particular organization is 4.5 lakhs per annum. After several years, even if we cannot determine her/his highest salary without complementary knowledge, we can conclude that it will not be lower than 4.5 lakhs per annum and will be one of {6 lakhs, 8 lakhs, more than 8 lakhs} with different nonzero probabilities.

The rest of this paper is organized as follows: We give an overview of static anonymization in section II. Then dynamic anonymization techniques are discussed in Section 3. And the key distribution techniques are discussed in Section 4.

2. Static Anonymization

In the literature, many static anonymization techniques were developed to generalize the data. These methods are discussed in following sections.

2.1 k-Anonymity

Samarati [4] and Sweeney [5] introduced k-anonymity as the property that each record is indistinguishable from a defined number (k) if attempts are made to identify the data. For any data record with a set of attribute values, if there are atleast k-1 other records that match those attribute values then, the dataset is said to be k-anonymzed.Assume that the dataset contains 3 quasi identifier attribute: Gender, Age, Zip code. Now the dataset is said to be k-anonymized if, for any one particular record, there exists k-1 other records that have same Age, Gender and Zip code. Accordingly, value of k may be 2, 3, 4 and so on. More privacy is achieved for large value of k.

Fig.2 gives the original microdata. Fig.3 gives the data to be published in which uniquely identifying attribute is removed to avoid identification of records from microdata. Fig.4 gives 3-anonymized dataset. K-anonymity can prevent only identity disclosure, it cannot prevent disclosure of attribute information and there are 2 attacks that can take place. Consider following examples. Given the published tables, if I know that sujata is 44 years old, lives in zipcode 400182 and there is one record of sujata in these tables then, I can conclude that she belongs to 3rd section of an anonymized table and is suffering from Flu. This leads to Homogeneity attack.

Given the published tables, if I have some background knowledge about Pankaj such as, Pankaj is 37 years old, and he lives in zip code 440182. Then there are two possibilities of Pankaj having viral infection or Heart problem. But as per my background knowledge, he does not suffer from Heart problem. That means here, I can conclude that Pankaj has some viral infection.

ID	QID			SA
Patient Name	Gender	Age	Zip code	Health Problem
Amit	Male	35	400071	Viral Infection
Pankaj	Male	37	400182	Viral Infection
Vishal	Male	39	400095	Heart problem
Sheetal	Female	54	440672	Flu
Pallavi	Female	58	440123	Heart problem
Nilesh	Male	54	440893	Viral Infection
Sagar	Male	41	400022	Flu
Mahesh	Male	46	400135	Flu
Sujata	Female	44	400182	Flu

Fig.2: Original Microdata

QID			SA
Gender	Age	Zip code	Health Problem
Male	35	400071	Viral Infection
Male	37	400182	Viral Infection
Male	39	400095	Heart problem
Female	54	440672	Flu
Female	58	440123	Heart problem
Male	54	440893	Viral Infection
Male	41	400022	Flu
Male	46	400135	Flu
Female	44	400182	Flu

Fig.3: Published dataset

If 3-anonymization is applied, the dataset becomes as follows:

QID			SA
Gender	Age	Zip code	Health Problem
*	<40	400*	Viral infection
*	<40	400*	Viral Infection
*	<40	400*	Heart problem
*	5*	440*	Flu
*	5*	440*	Heart problem
*	5*	440*	Viral Infection
*	>40	400*	Flu
*	>40	400*	Flu
*	>40	400*	Flu

Fig.4: 3-Anonymized dataset

2.2 l-Diversity

Machanavajjhala et al. [3] introduced a new model, called l-diversity, which requires that there are 'l' different sensitive values for each combination of quasi identifiers. The definition of l-diversity is given as follows:

An equivalence class is said to have l-diversity if there are at least l "well-represented" values for the sensitive attribute. A table is said to have l-diversity if every equivalence class of the table has l-diversity.

3-diverse version of above original table can be given as follows:

QID			SA
Gender	Age	Zip code	Health Problem
*	<50	4000*	Viral Infection
*	<50	4000*	Heart Problem
*	<50	4000*	Flu
*	>50	440*	Flu
*	>50	440*	Heart problem
*	>50	440*	Viral Infection
*	<50	4001*	Flu
*	<50	4001*	Flu
*	<50	4001*	Viral Infection

Fig.5: 3-diverse dataset

Similar to k-anonymity, l-diversity does not prevent attribute disclosure. And there are some attack that may occur on l-diversity such as, Skewness attack and Similarity attack. The information leakage occurs in l-diversity because it does not consider semantically closeness of sensitive values [6].

Following microdata has 2 different sensitive values: one is numeric attribute and another is categorical attribute.

QID			SA	
Gender	Age	Zip code	Health Problem	Bank Balance
Male	35	400071	Viral Infection	10K
Male	37	400182	Viral Infection	11K
Male	39	400095	Heart problem	12K
Female	54	440672	Flu	13K
Female	58	440123	Heart problem	15K
Male	54	440893	Viral Infection	16K
Male	41	400022	Flu	16K
Male	46	400135	Flu	17K
Female	44	400182	Flu	18K

Fig.6: Original Microdata

It can be anonymized using 3-diverse generalization as follows:

QID			SA	
Gender	Age	Zip code	Health Problem	Bank Balance
*	>30	400*	Viral Infection	10K
*	>30	400*	Heart Problem	11K
*	>30	400*	Flu	12K
*	5*	440*	Flu	13K
*	5*	440*	Heart problem	15K
*	5*	440*	Viral Infection	16K
*	>40	400*	Flu	16K
*	>40	400*	Flu	17K
*	>40	400*	Viral Infection	18K

Fig.7: 3-Diverse version of dataset

If an adversary somehow knows that Amit has record in first equivalence class that means Amit's bank balance ranges from 10K-12K. This gives knowledge that Amit has less Bank balance. Thus, the information is disclosed here also.

2.3 t-closeness

Ninghui Li, Tiancheng Li and Suresh Venkatasubramanian proposed a new privacy model known as *t*-closeness which

requires that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table (i.e. the distance between the two distributions should be no more than a threshold *t*). The definition for t-closeness is given as follows:

An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t. A table is said to have t-closeness if all equivalence classes have t-closeness.

t-closeness uses Earth Mover Distance (EMD) to calculate the distance between two distributions [2]. And it also considers semantic closeness of attribute values. Consider the original datasets given in fig.6 and its 3-diverse version given in fig.7.

Now, EMD can be calculated by using the solution of transportation problem.

Consider following example to calculate EMD of D [P1, Q] where, P1 is the distribution of sensitive attribute in 1st equivalence class and Q is the distribution of sensitive attribute in overall dataset. The classification of equivalence classes is done by using 3-diversity as shown in following figure.

Here, Q={10K,12K,14K,18K,20K,22K,26K,28K,30K}
And P1={10K,12K,14K}. Now the distance between P1 and Q is calculated as follows:
D[P1,Q]=1/9* (1+2+2+3+4+5+4+5+6)/8=29/72=0.402

Thus, the t-closeness calculates the EMD and finds closeness between distributions.

QID			SA	
Gender	Age	Zip code	Health Problem	Bank Balance
*	<50	4000*	Viral Infection	10K
*	<50	4000*	Heart problem	12K
*	<50	4000*	Flu	16K
*	>50	440*	Flu	13K
*	>50	440*	Heart problem	15K
*	>50	440*	Viral Infection	16K
*	<50	4001*	Viral Infection	11K
*	<50	4001*	Flu	17K
*	<50	4001*	Flu	18K

Fig.8: 0.402 t-closeness for Bank Balance

One cannot determine particular individual with low bank balance or health problem.t-closeness prevents attribute disclosure but it cannot protect the dataset against identity disclosure.

The static anonymization can be summarized as follows in table 1.

Table 1: Summary of Static Anonymization

Sr.No.	Type of Anonymization	Information Disclosure
1.	k-Anonymity	Prevents Identity disclosure
2.	l-Diversity	Prevents Identity disclosure
3.	t-closeness	Prevents Attribute disclosure but does not prevent Identity disclosure

3. Dynamic Anonymization

Above models were considering only static datasets. They did not considered updates in the published datasets. Serial publishing for dynamic databases is necessary whenever there are insertions, deletions and updates in datasets [7]. Here, two different techniques used for dynamic anonymization are discussed.

3.1 M-Invariance

Byun [5] first identified the attacks that occur while republishing the data and proposed a solution to effectively prevent those attacks. But, it only supports the insertions of data, and does not consider deletions and updates. To understand this concept, consider one example where the Hospital republishes the dataset after few months. Meanwhile some records are inserted and some are deleted.

ID	QID			SA
Patient Name	Gender	Age	Zip code	Health Problem
Amit	Male	35	400071	Viral Infection
Pankaj	Male	37	400182	Heart Problem
Vishal	Male	39	400095	Flu
Sheetal	Female	54	440672	Flu
Pallavi	Female	58	440123	Heart problem
Nilesh	Male	54	440893	Viral Infection
Sagar	Male	41	400022	Flu
Mahesh	Male	46	400135	Flu
Sujata	Female	44	400182	Flu

Fig.9: 1st release of Dataset

QID			SA
Gender	Age	Zip code	Health Problem
*	<40	400*	Viral infection
*	<40	400*	Heart problem
*	<40	400*	Flu
*	5*	440*	Flu
*	5*	440*	Heart problem
*	5*	440*	Viral Infection
*	>40	400*	Flu
*	>40	400*	Flu
*	>40	400*	Flu

Fig.10: k-anonymized data

ID	QID			SA
Patient Name	Gender	Age	Zip code	Health Problem
Amit	Male	35	400071	Viral Infection
Sneha	Female	36	440574	Gastritis
Vishal	Male	39	400095	Flu
Sheetal	Female	54	440672	Flu
Pallavi	Female	58	440123	Heart problem
Neha	Female	53	400342	Gastritis
Sagar	Male	41	400022	Flu
Mahesh	Male	46	400135	Flu
Atul	Male	42	440234	Gastritis

Fig.11: 2nd release of Dataset

Now, consider that an adversary knows that Sagar has the record in both the tables in fig.9 and fig.11. Based on fig.10, adversary can conclude that Sagar has Flu. And based on fig.12, adversary can conclude that Sagar has either Flu or Gastritis. So, by combining both these tables, he/she can conclude that Sagar has Flu and not Gastritis [11].

QID			SA
Gender	Age	Zip code	Health Problem
*	3*	400071	Viral Infection
*	3*	440574	Gastritis
*	3*	400095	Flu
*	>50	440672	Flu
*	>50	440123	Heart problem
*	>50	400342	Gastritis
*	>40	400022	Flu
*	>40	400135	Flu
*	>40	440234	Gastritis

Fig.12: k-anonymization of republished data

To avoid this determination, M-Invariance was proposed by X. Xiao and Y. Tao which makes use of counterfeited generalization as follows [8]:

ID	QID			SA
Patient Name	Gender	Age	Zip code	Health Problem
Amit	Male	35	400071	Viral Infection
f1	Male	37	400182	Heart Problem
Sneha	Female	36	440574	Gastritis
Vishal	Male	39	400095	Flu
Sheetal	Female	54	440672	Flu
Pallavi	Female	58	440123	Heart problem
f2	Male	54	440893	Viral Infection
Neha	Female	53	400342	Gastritis
Sagar	Male	41	400022	Flu
Mahesh	Male	46	400135	Flu
f3	Female	44	400182	Flu
Atul	Male	42	440234	Gastritis

Fig.13: Republished data with counterfeit records

Consider another example. If the adversary refers first published dataset in fig.9 and republished dataset in fig.13, he/she cannot conclude the possible disease of Amit. Because these groups encompass same set of sensitive attribute values.But M-invariance supports only external updates.

3.2 M-Distinct

To overcome the problem in m-invariance, Feng Li and Shuigeng Zhou proposed a new generalization principle m-Distinct to effectively anonymize datasets with internal as well as external updates. M-Distinct uses m-unique which is used to maintain the sensitive values which are not different in separate publication [9]. In order to maintain this indistinguishability of sensitive values, records must be partitioned carefully while releasing new publication.

Along with providing anonymization, M-Invariance minimizes number of counterfeit records to be added and generalization of QI attribute. More generalization of attributes may exhibit more loss of information.

Major difference between both these techniques is given as follows in table 2.

Table2:Summary of Dynamic Anonymization

Sr.No.	Techniques Available	Description

1.	M-invariance	Anonymize the dataset with external update
2.	M-Distinct	Anonymize the dataset with external as well as internal updates

4. Key Distribution

A key distribution technique must be adapted to authenticate the receivers in the cloud. Many algorithms exist for key distribution and key management. These algorithms are classified as Contributory schemes and Distributive scheme [12]. Distributive scheme is again classified as Symmetric scheme and Asymmetric scheme. In Contributory scheme, communicating parties generate and distribute the key to be used. Whereas in Distributive scheme, it is the responsibility of trusted third party to generate and distribute the keys. These schemes include Diffie-Hellman key exchange algorithm, Ingemarsson, Tang and Wong (ING), Hypercube and Octopus (H&O), Autonomous Key Management (AKM) etc. These techniques may be used to distribute the keys to participating entities.

5. Conclusion

In this paper, an analytical study of how static and dynamic anonymization techniques have been used in the literature has been presented. While k-anonymity provides protection against identity disclosure, it does prevent attribute disclosure. Similarly, l-diversity also does not prevent attribute disclosure. It has a number of limitations as we have discussed. Then a new privacy model t-closeness is discussed which incorporates EMD measure to calculate the distance between two distributions. Along with publishing a dataset, an organization must also republish an updated dataset to preserve the privacy. Two different approaches exist which provides these dynamic updations: M-invariance and M-distinct.

These techniques can be used as a better approach to secure a data in a cloud. K-anonymity can be used along with M-Distinct to provide dynamic anonymization. Then, use a key distribution approach which will authenticate the users and provide an anonymization table to valid users. Along with these methods, anonymization using other methods such as Hashing, Hiding, and Permutation can also be used to secure a cloud data.

References

[1] Jeff sedayao, "Enhancing cloud security using Data Anonymization", Intel white paper, June 2012

[2] Ninghui Li Tiancheng Li, Suresh Venkatasubramanian, "*t*-Closeness: Privacy Beyond *k*-Anonymity and *l*-Diversity", 2007.

[3]Ninghui Li, Member, IEEE, Tiancheng Li, and Suresh Venkatasubramanian, "Closeness: A New Privacy Measure for Data Publishing" ,IEEE transactions on knowledge and data engineering, vol. 22, no. 7, July 2010

[4] P. Samarati, "Protecting Respondent's Privacy in Microdata Release," IEEE Trans. Knowledge and Data Eng., vol. 13, no. 6, pp. 1010-1027, Nov./Dec. 2001.

[5] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy," Int'l J. Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 5, pp. 557-570, 2002.

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

[7] Xiaolin Zhang, Hongjing Bi, "Secure and Effective Anonymization against Re-publication of Dynamic Datasets", IEEE 2010

[8] X. Xiao and Y. Tao, "m-invariance: Towards privacy preserving republication of dynamic datasets," in SIGMOD. 2007.

[9] Priyanka Gupta, Ashok Verma , "Establishing a Service Model of Private Elastic VPN for cloud computing", ijcsn, vol 1, issue 4, 2012

[10] J. Byun, Y. Sohn, E.Bertino, and N.li, "Secure anonymization for incremental datasets," in Secure Data Management. 2006, pp.48–63.

[11] J. Pei, J. Xu, Z. Wang, W. wang, and K. Wang, "Maintaining k-anonymity against incremental updates," in SSDBM. 2007.

[12] Erdal Çayırcı and Chunming Rong," Security in Wireless Ad Hoc and Sensor Networks" @2009 John Wiley & Sons, Ltd. ISBN: 978-0-470-02748-6

[13] Feng Li and Shuigeng Zhou, "Challenging More Updates: Towards Anonymous Re-publication of Fully Dynamic Datasets" arXiv: 0806.4703v2 [cs.DB] 24 Jul 2008.