Enhancing Security in Public Clouds using Data Anonymization Techniques

N. Nishara
School of Computing Science and Engineering
VIT University, India

Reeta Pandey
School of Computing Science and Engineering
VIT University, India

ABSTRACT
Security issues have given rise to immerging an active area of research due to the many security threats that most of the organizations have faced at present. Despite the advancements in cloud computing, the organizations are slow in accepting it, due to security threats that make a cloud environment to be source of data breaching. Maintaining privacy for the high dimensional database has become an important aspect of security. This paper, emphasizes on protecting the data in public cloud using data anonymization techniques. Anonymization is the process of making the sensitive data to be de-identified and preventing this data to be linked with identities of an individual or an organization. The data has to be anonymised, thereby preventing it from malicious attack & at the same time data must be also made available for the owner of the data. To preserve the data from the attacker, two methods of privacy preserving models are used - k-anonymity and l-diversity. Finally, in this paper an algorithm for graph anonymisation is presented, called the Evolutionary Algorithm for Graph Anonymization (EAGA) that is based on k-anonymity model.

Keywords
Data anonymization, Cloud Computing, k-anonymity, l-diversity

1. INTRODUCTION
The amount of complexity involved in protecting the personal and sensitive organizational data remains to be an ever growing concern. Security remains to be directly proportional with that of the cost factor. Large organizations can afford to pay more for infrastructure as well as for security purposes, whereas medium & low level organizations heavily invest on product development and delivery. Hence there was a need for a computing environment that will boost the existing infrastructure as well as improve the utilization of resources, all at an affordable cost. Cloud computing emerged into the market to allow organizations to add up resources as per their need & allow them to use the existing resources.

Merrill Lynch (2008***),*** stated that, Cloud computing is the concept of delivering business productivity and personal applications from centralized servers, over the internet. The major advantage to be considered when using a cloud is reduction over the investment on maintaining a huge data center. Another significant quality would be the integration among various programming models provided by the cloud. But at the same time, the organizations are hesitant in accepting cloud, to store their data as it becomes the source for information breaching. The major security threat is the concept of running an individual's data on a remote hard disk of some service provider and using a remote CPU which may be daunting experience for many of them. Such kind of storing data always have a possibility of suffering from data loss, phishing attacks etc. The multi-tenancy architectural model of the cloud required a novel approach in providing security against the misuse of data stored in it. We need to find a more practical approach to store data in the cloud that easy to implement as well as protects privacy.In this paper, the proposed idea is the use of Data anonymization techniques , that can be employed for better privacy protection of sensitive data. The anonymised data can still be processed to obtain a useful information. This can be done by adding fictitious data to the released data content which is generally stored in a cloud environment. To improvise the data hiding techniques, privacy preserving models k-anonymity & l-diversity are used. Organization of the paper is as follows: section II Related work, section III Proposed system

2. RELATED WORK
The existing systems for preserving privacy of data were highly dependent on the data mining techniques. They were:

2.1 Data Mining for Privacy Preserving
Generally when discussed about privacy, it is said that hiding the sensitive information about an organization or an individual is necessary. In some of the western countries like UK and US relieving the medical details or say making a patient's sensitive detail available to an individual or a group of people is considered as an offence. It is this intrusion, which might cause negative impacts in the life of the patient. In order to ensure this technical and social solutions are to be considered in such a way that the private details of a person are not revealed as well as the administrators also have enough space & independence to work with the available data.

For eg, an individual may not care if an outsider comes to know about his birth date, social security number, mother's maiden name. But when an intruder combines all these, there is a possibility of identity theft. The disclosure of knowledge about a single entity or attribute has chance that leads to an individual's privacy violation. Hence this method is prone to intrusion attack.

2.2 Data for Privacy Publishing
To prevent the breaching of data there were several transformation methods used. The techniques includes methods such as, k-anonymity, l-diversity, randomization techniques. But when combined with the perturbed data these methods do not confirm the provision of best utility of underlying data set. This data set includes keys, quasi-identifiers, explicit identifiers.

For eg, in case of hospital management system, the Sensitive attribute of a person may include details of disease, disability status & salary. Non sensitive attributes include all those that do not come under the category of Sensitive attributes. These techniques do not go well with the rule association mining. Also the memory needed to store all these data was another major concern. As the world is now moving towards the

concept of Distributed Systems, there came a gateway opened for dealing with the storage of data as well as the best utilization of available resources at a maximum level. Hence cloud computing came into light. This technology proved worthy in terms of cost and scalability .But the major threat remained the security. Hence this paper proposes a method to combine the anonymization techniques with the public cloud deployment model thereby improving the security provided in the cloud environment.

2.2 Anonymization–Concepts& Techniques

Generally, there is a need to find a flexible approach to store data in the cloud that is not too difficult to implement as well as protects privacy, which is the ultimate goal. Data anonymization technique is employed to improve the data security in the public clouds but still allows the data to be effectively used & analyzed. This methodology is used in the areas where preserving the personal data about an individual (for eg, treatment related data of a patient) and sensitive data of an organization (Revenue of a company) is essential. It is a technique where the content published will be in the form that prevents key information from being identified. Anonymised data can be stored in cloud without any concern that a hacker might access the data. The details of anonymised Anonymization is different from Encryption though both techniques are imbibed in enhancing the quality of data stored in cloud.

The former technique is the process of transforming data so that it can be processed in an effective way, but at the same time prevents data from being linked to individual identities of people or a company. The latter involves transforming data in an unreadable format with a provision of a key to decrypt the data on the receiver's side. Two privacy preserving models have been established: k-anonymity and l-diversity which will be discussed later in this paper.

2.3 Cloud – A Comprehensive approach

Cloud (Gartner 2008) is a style of computing where IT – oriented functionalities are provided "as a service" that uses Web technologies to service a large number of service requestors. The major objective of cloud computing was best sharing of resources thereby reducing the cost overhead.

2.4 4.1 Architecture

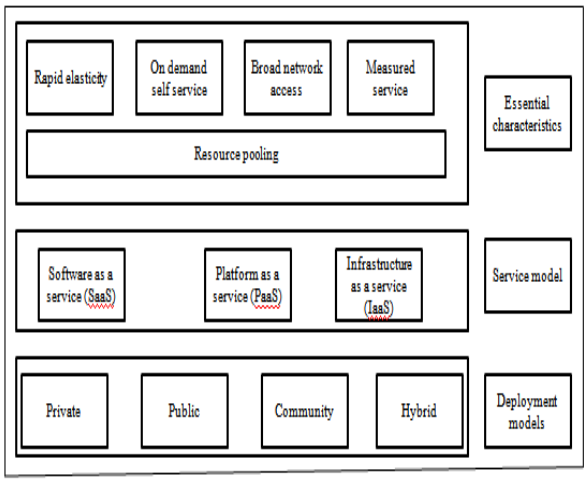

Figure 1: Architecture of Cloud Computing

2.5 Terminologies associated with Cloud

Deployment Models- Cloud commonly supports four deployment models [5][6]. They are broadly classified as:

(i) Public cloud: The cloud infrastructure is made available to general public or a large organization which is managed by cloud service provider.

(ii) Private cloud: The infrastructure is completely dedicated to a single company, and is managed either by the company or is handed over to any third party organization.

(iii) Community cloud: The community cloud infrastructure is shared among the organizations that work on a common criteria.

(iv)Hybrid cloud: It involves the combination of public and private clouds that helps in application and data portability.

Service Models: Three major service models are studied commonly. They are:

Software as a Service (SAAS): The consumers are allowed to use the service provider's application but are denied access from managing the servers, operating system and networks.

Platform as a Service (PAAS): The consumers are provided with the tools and technical support and are allowed to deploy consumer created or acquired applications but are denied the permission to manage the networks, operating system.

Infrastructure as a Service (IAAS): The consumers cannot control the cloud infrastructure but can manage the operating systems, storage, networks (includes host firewalls).

3. PROPOSED SYSTEM

3.1 Anonymization in achieving privacy

As we propose that anonymization is an efficient tool in protecting privacy of the sensitive data, in order to accomplish it in the example that is used in this paper, the names of the companies(X, Y, Z) are converted to Aron, Mike, Stephen and fictitious names or data (Sara and Vista) are added and are stored in the cloud based data storage.A secure enclave which is generally a well secured area in the cloud is selected and the converted data is stored in a transition table. This secure enclave finally maps the translation of the names and also identifies the fictitious data. Using this anonymised data total revenue of the companies can be populated without revealing the sensitive information. The final result can be calculated by subtracting the fictitious company data, internally from the transition table. This method can be effective as it is difficult for the hacker to identify the actual number of companies or the particular company on which he is interested on hacking.

This method can have many negative effects if done incorrectly. For example, a popular online shopping dealer released a database of users and the products purchased and liked by them. Researchers warn that the identities and the details of the customers can be found out by correlating the publicly available database.Several privacy models were diffused in order to safe guard the data, which includes k-anonymity & l-diversity which will be discussed in detail in the next section.

3.1.1 k-ANONYMITY

The major goal is to make every available record on the cloud to be indistinguishable from k number of records[1]. It guarantees that every released data is accurate. A data set is said to be k anonymised if there exist k-1 records that match the set of attributes for every data record available. Attributes such as key, quasi-identifier, sensitive are present as given below,

Table 1: k-Anonymity Attributes

Attribute Type	Characteristics	Example	Action to be taken
Key	Directly identifies an individual	Name, SSN.	Remove or obscure
Quasi identifier(QI)	linked with other external information to identifies an individual	Birthday ZIP code, gender	Generalize or suppress
Sensitive	Sensitive data about an individual	Type of illness, income.	Needs to be de-linked from the individual

The most sensitive data can be delinked from the table, so that the attacker can never identify the individual. Table 2 contains some sample hospital records. Zip and Age are quasi-identifiers and Disease is a sensitive attribute. This method will prevent database linkages that are definite. One of the most promising aspects of k-anonymity is that information cannot be linked to group less than k-individuals.

Table 2: k-Anonymised patients record sample

Zip	Age	Disease
140·	4.	Diabetes
140·	4.	Diabetes
140·	4·	Diabetes
140·	4·	Jaundice
140·	5·	Lung cancer
140·	5·	Lung cancer

Represents a suppressed value

K-anonymity focusses mainly on Generalization and Suppression. ***Generalization*** is the commonly used approach, which replaces the quasi identifier values with less specific values [2]. Due to high dimensionality of QI's, generalization may suffer information loss. To make generalization to be more effective, records in the same bucket must be closer with each other. ***Suppression*** works on and suppresses the free availability of data.

Limitations

1. No protection against the background knowledge and homogeneity attacks.
2. It is not applicable on high dimensional data integrity without losing data [3].
3. When a data set is anonymised and gets published more than once, the efficiency of generalization is a concern.

3.1.2 l-Diversity

l-Diversity overcomes both background and homogeneity attacks that k-anonymity suffers. The difference between both the privacy models is that k-anonymity requires each combination of QI's in order to have k-entries whereas l-diversity requires that there are l-different sensitive values for every combination of QI's. Thereby l-diversity proves to be more effective providing data security.

Table 3: Patient records sample created with l-Diversity, where l=4

Zip	Age	Disease
140·	4·	Diabetes
140·	4·	Diabetes
140·	4·	Diabetes
140·	4·	Lung cancer
140·	4·	Lung cancer
140·	4·	Jaundice
140·	4·	Jaundice
140·	5·	Jaundice
140·	5·	Jaundice

The cloud architecture holds the anonymised data sets with a higher degree of security.

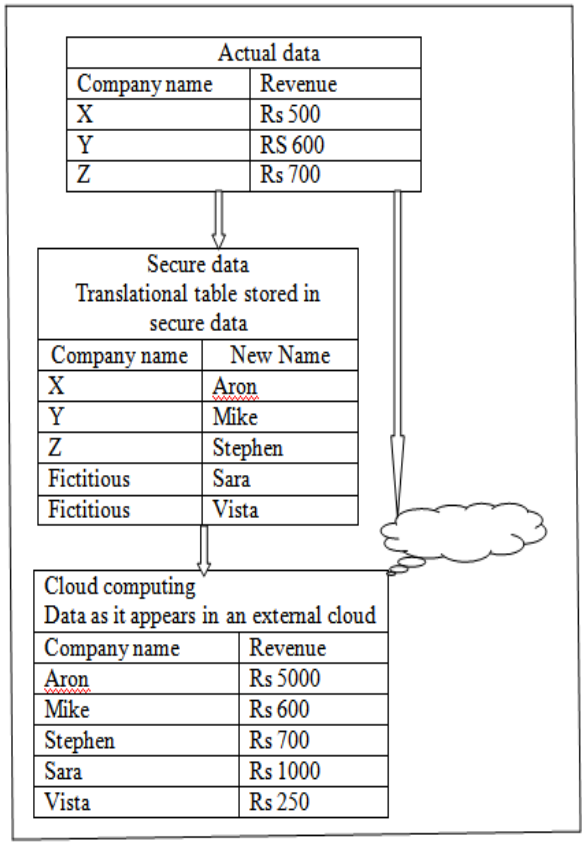

Fig 2: Data anonymization helps enable safer computing in the cloud

3.2 EAGA Algorithm

In recent years there has been a significant increase in the utilization of social networks. These Social Medias become the source for researchers and the third parties, whose major concern is business.Hence there arises a need to safe guard the customer's details who appears in the network using

anonymisation technique. In this approach we present the idea of graph anonymisation called ***EAGA*** (Evolutionary Algorithm for Graph Anonymisation)[4]. The description of the algorithm can be featured in the following 2 steps.

1). Generating the sequence ~d from the original sequence of G=(V,E) and d={d1,d2,…d_n} and thereby minimizes the distance(Δ) from original sequence which is given by, Δ=(~d,d) =∑ |~d_i -d_i| (for i=0 to n)

2). Modifying the original graph: ~G= (~V,~E) where ~V=V, ~E∩E≈E and the degree of the sequence is equal to ~d.

Algorithm-Step 1: Generating anonymous degree of k-sequences

Require: Original degree sequence (d) and the k-anonymity value (k).
Ensure: k-degree anonymous sequence ($\tilde{d}$).
INITIALIZE *population* ⇐ d
k_actual ⇐ GET_K *population*
while *k_actual* < k **do**
MUTATE *population*
EVALUATE *new candidates*
population ⇐ SELECT *individuals*
k_actual ⇐ GET_K *individuals*
end while
$\tilde{d}$ ⇐ SELECT *best candidate*
return $\tilde{d}$

In order to construct such a sequence there are few conditions to be met. They are:

1. The number of nodes is determined by using the number of elements in the degree sequence and hence it must be stable.
2. The value of degree sequences are the degrees of nodes,
3. The total number of edges will be half the sum of degree sequence since each edge is counted twice in degree sequence.

Algorithm- Step 2: Modifies Original Graph
The output of first algorithm will be k degree sequences form which we need to find out k degree in each node in anonymized graph. The difference between original and anonymized degree sequence identifies the nodes for which the degrees can be increased or decreased. The difference vector, $\mathbf{d_{dif}=d-\sim d}$ identifies the nodes for which the degrees has to be modified. The algorithm removes incident edges to nodes to remove to decrease the degrees and add new edges in order to increase them.(V_P,V_Q) Є E where VQ belongs to nodes whose degree must be decreased & adding new edge VR (VP,VR) which belongs to nodes that has to increase the degree.

Algorithm 2 Algorithm pseudo-code for modifying the original graph.

Require: Original graph $G(V, E)$, original degree sequence d and the k-degree anonymous sequence $\tilde{d}$.
Ensure: The graph $\tilde{G}(V, \tilde{E})$ where the degree sequence is $\tilde{d}$ and $\tilde{E} \cap E \approx E$.
$\tilde{G}(V, \tilde{E}) \Leftarrow \tilde{G}_0(V, \tilde{E})$
$d_{dif} = d - \tilde{d}$
$V_{del} = \{v_i \in V | d_{dif}(i) < 0\}$
$V_{add} = \{v_i \in V | d_{dif}(i) > 0\}$
while $V_{del} \neq \emptyset$ and $V_{add} \neq \emptyset$ **do**
$\tilde{E} = \tilde{E} \setminus \{(v_p, v_q)\}$ where $(v_p, v_q) \in E$ and $v_q \in V_{del}$
$V_{del} = V_{del} \setminus \{v_q\}$
$\tilde{E} = \tilde{E} \cup \{(v_p, v_r)\}$ where $v_r \in V_{add}$
$V_{add} = V_{add} \setminus \{v_r\}$
end while
return $\tilde{G}$

4. CONCLUSION

The research area is wide open in terms of security provision in the cloud environment. In this paper, different privacy preserving models based on data anonymization were discussed followed by the introduction of a new methodology of EAGA algorithm which stabilizes the anonymization technique for the graph networks further. This anonymised data when stored in the public cloud will persist with high levels of security. As the technology of social networks and larger organizational data sets is now focused on Public cloud, this method might prove to be a promising area of research in near future.

5. ACKNOWLEDGEMENT

We would like to express our sincere thanks to our faculty, G.Lydia Jane, Assistant Professor, School of Computing Science & Engineering, VIT University , for her guidance and support.

6. REFERENCES

[1] k-Anonymity" – P.Samarati, S.Foresti, S. De Capitani, Universit`a degli Studi Milano, Italia.

[2] Privacy Preserving for high dimensional data with Anonymisation Techniques"- Prof. Girish Agarwal, Prof.Pragati Patil, ABHA Gaikwad Patil College of Engineering, Nagpur. IJARCSSE, Volume 3, June 2013.

[3] "Anatomy- Simple & Effective privacy preservation"- X.Xiao and Y.Tao, Proc.Int'l Conf. Very Large Databases (VLDB), pp. 139-150, 2006.

[4] Anonymous Publication of Sensitive Transactional Data" – Gabriel Ghinita, Member IEEE, Panos Kalnis, Yufei Tao, in Proc. Of IEEE Transactions on Knowledge & Data Engineering Feb 2011(vol. 23) pp.161-174.

[5] White papers on Cloud Security. Cloud Computing & Information Security, June 2012.

[6] Cloud Computing – Principles and Paradigms, Andrzej Goscinski, R.K.Buyya, James Broberg, Wiley Publishers, 2013.

[7] NIST cloud computing standards roadmap, U.S. Department of Commerce, Special Publication 500-291, Version 2.