Universal Journal of Communications and Network 2(5): 87-92, 2014 http://www.hrpub.org
DOI: 10.13189/ujcn.2014.020501

User Profiles and Identifing User Behaviour in the Cloud Computing Environment

Rasim Alguliev, Fargana Abdullaeva*

Institute of Information Technology, Azerbaijan National Academy of Science, Baku, Azerbaijan
*Corresponding Author: farqana@iit.ab.az

Copyright © 2014 Horizon Research Publishing All rights reserved.

Abstract In this paper, for the detection of the masquerade attacks in the cloud infrastructure collaborative filtering algorithm based on the cloud model is proposed. One of the advantages of this model is the identification of the similarity between the users on the basis of the cloud model. While using the similarity measurement method based on the cloud model, it does not require a strict comparison between the score value of operations used by different users. Here we provide the calculation of the statistic features of the score values of all operations used by the user at the access point, then we provide a comparison of statistics features of the input data and based of these we determine the similarity between the input data.

Keywords Cloud Computing, Masquerade Attack, Cloud Model, User Similarity, Collaborative Filtering

1. Introduction

The emergence of the cloud computing technologies nowadays has significantly changed the way we use the computer means as well as the way we access and store our personal and business information. The new computing and communications paradigms based on these technologies cause the emergence of the new security problems. To be more precise, in comparison with the traditional technologies the nature of the threats brought by the cloud technologies to the organization infrastructure has changed. Thus, by the huge CSA (The Cloud Security Alliance) which is actively operating in the area of the standardization of the security issues of cloud technologies classifies the main threats brought by this technology as follows [1]:

- Threat 1. Abuse and Nefarious Use of Cloud Computing;
- Threat 2. Insecure Interfaces and APIs;
- Threat 3. Malicious Insiders;
- Threat 4. Shared Technology Issues;
- Threat 5. Data Loss or Leakage;
- Threat 6. Account or Service Hijacking;
- Threat 7. Unknown Risk Profile.

Among these threats the 1st and the 3d threats can be implemented in each layer of the cloud computing SPI (Service Platform Infrastructure) model. In both of these threats in order to use the resources illegally the attacker tries to present himself to the system as a legitimate user by capturing legitimate user's identity information. These threats are known as masquerade attacks in the literature sources [2]. Moreover, according to the cybercrime watch survey [3] conducted by the organization CERT in 2010, the first place among the top 5 electronic crimes belongs to viruses, worms and other malicious codes, the second part takes the unauthorized access.

The Twitter incident committed by the French hacker can be introduced as an example of this type of attacks. Several Twitter corporate and personal documents were leaked to the technological website called "The TechCrunch" [4, 5] and customers' accounts including the account of the U.S. President Barack Obama were illegally accessed [6, 7]. The attacker used a Twitter administrator's password in order to gain access to Twitter's corporate documents hosted on Google's infrastructure as Google Docs. The damage was significant both for Twitter and for its customers.

There are number of research works devoted to the discovery of masquerade attacks. In one of these studies [8] in order to discover the masqueraders a new approach has been presented. This approach implements the comparison between the two users' command sequence. This method is based on the similarity measured between the 10 most recent commands and a user's profile. But this approach is not suitable for cloud infrastructure because cloud systems are heterogeneous systems [9, 10]. Here the virtual machines which form cloud systems are executes under the different operating systems (Windows, Linux, UNIX, Network). It is also likely that different commands are used for the same operation. For this reason, it is impossible to evaluate similar users by their command sequence.

In the other approach a sequence alignment method, which is broadly used in bioinformatics, is applied to discover the illegal accesses [11, 12]. As it can be seen most of the studies are focused on the detection of unusual or variable command sequence applied by the users.

Generally, the legitimate users of the computer system

are familiar with the files on that system and are aware of their location. Any search for the specific files is likely to be targeted and limited here. A masquerader, however, who gets access to the victim's system illegitimately is unlikely to be familiar with the structure and the contents of the file system. For this, his search is likely to be widespread and untargeted as well as the type and sequence of commands applied by the users in the system are being different. Taking this into account, the user's profile reflecting his search behavior is created at [13, 15] and the one-class modeling technique is developed in order to detect the illegal intrusion.

However it should be noted that, when the attacker is being more deeply familiar with the genuine user's behavior he can more accurately imitate him in the system. In this case, the above mentioned methods lose their influence and the detection of the illegal accesses by that way does not allow to obtain good results.

Typically, a user who gain access to the cloud system by masquerading under the genuine user always has a unique interest on the selected target resources. In this case the detection of the illegal accesses by modeling the user's specific interest may allow to obtain significant results.

In this paper we created a profile which reflects the interest behavior of each user and this profile is considered as a cloud model. This cloud model reflects not a command sequence used by the user but the statistical characteristics of the operations conducted by the users. According to this method, in order to discover the illegal attack the normal behavior of a user is modeled as a cloud model first; then these models are compared and the deviations from this behavior are evaluated. If the deviation value becomes above the threshold the user who gained access to the system is evaluated as the illegal user.

2. Masquerade Attack Detection System

Masquerade attacks refer to attacks that use a fake identity in order to gain unauthorized access to the personal computer information through the legitimate access identification.

The most common information which can be used to detect the masquerade attacks is contained within the actions the masquerader performs. This set of actions is known as a behavioral profile. The masquerade detection techniques are based on the premise that when a masquerader attacks the system he will sufficiently deviate from the user's behavior. In other words, at first the normal behavior of the user is modeled as a profile there, then this profile is compared with his current behavior and the obtained deviation cases are considering as masquerader attacks.

In this paper we propose a new approach for detecting a masquerader in the cloud environment. This approach is schematically described in Figure 1. According to the architecture the masquerade detection is implemented in two phases:

2.1. Profile Creating Phase

The genuine user data collected and on the basis of the colleccted information a behavioral profile is created for each user. This phase consists of two components. The Event logs implements the collection of the information on all events during the session. The Feature Extraction Tool generates the feature vector based on cloud model for the user.

2.2. Detection Phase

The new user profile generating on the basis of the new user records and compared with the genuine user profile. There similarity value calculation block for each input vector there calculates the cosine similarity value. The similarity value here varies between 0 and 1.

If similarity value is high, then the input data become very close to the user profile and if it becomes very low, then it is estimated as very different. According to the input data, the detection module classifies all users as genuine or imposter. For this purpose, first of all by using filtering formula it calculates the deviation value and on the basis of the accepted threshold value in the system it makes a decision determining the user as genuine or imposter.

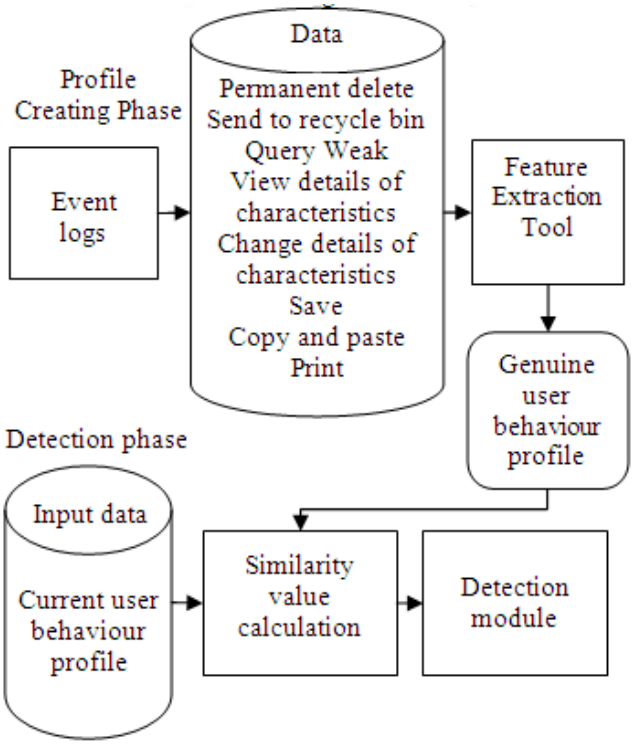

Figure 1. Masquerade attack detection system

3. User Profile Generation

In order to recognize the anomalous behavior for the detection system, first it must form a user profile to characterize a normal behavior. In this section we will describe the model underlying our approach to the user profiling and will discuss the implementation details of how

profiles are formed from the user's interest data.

3.1. Data Collection

The detection of the masquerade attacks is carried out on the basis of operations performed by the masqueraders in the system. This set of operations forms a user's behavior profile.

Based on these real data, it is possible to construct the interest profile in terms of the cloud model. This interest profile also can be called the feature vector of the user.

3.2. Feature Vector Construction

There every action of the user has a score which can be regarded as a drop of interest cloud, but the score value of all operations is regarded as an interest cloud. The cloud uses expectation E_x, entropy E_n and excess entropy H_e to represent one digital value. Together these three digital characteristics of i-th user's interest cloud form the c_i-th characteristic cloud vector. While computing the similarity between users i and j we can judge from similarity between vector c_i and c_j. The key innovation there is getting characteristic vector c of the user's interest cloud.

Assume that matrix $A(u_j, n_i)$ represents the score data set of operations used by the users. Here, u_j implies the users, n_i the operations used by the users, $Score_{i,j}$ - the score value of the i-th operation used by the j-th user. This assumption can be represented as follows in figure 2:

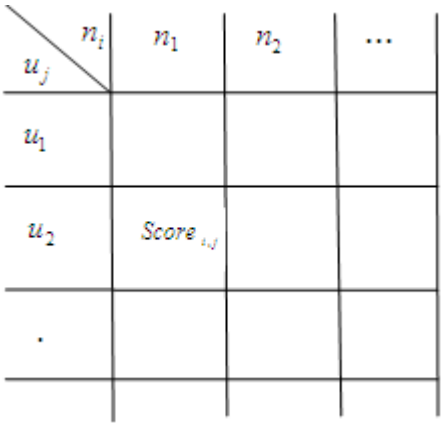

Figure 2. Score data matrix

In order to get the cloud feature vector $c=(E_x, E_n, H_e)$ of the user we need a backward cloud generator.

Assume that a set of score values used operations denoted as X_k, $k=(1,2,3,...,n)$. Then an algorithm to determine the digital value (E_x, E_n, H_e) can be constructed as follows:

Input: a set of score values of operations $X_k, k=\{1,2,...,n\}$

Output: $\{E_x, E_n, H_e\}$ numerical characteristics

1. The calculation of the average $\overline{X}$ score value, according to score value of the number of X_k operations applied by the user

$$\overline{X}=\frac{1}{n}\sum_{k=1}^{n}x_k \tag{1}$$

2. Absolute central moment calculation

$$M=\frac{1}{n}\sum_{k=1}^{n}\left|x_k-\overline{X}\right| \tag{2}$$

3. The calculation of the variance of the event

$$S^2=\frac{1}{n-1}\sum_{k=1}^{n}\left(x_k-\overline{X}\right)^2 \tag{3}$$

4. The calculation of the mean operation value

$$E_x=\overline{X} \tag{4}$$

5. The calculation of the dispersion of operations

$$E_n=\sqrt{\frac{\pi}{2}}\times\frac{1}{n}\sum_{k=1}^{n}\left|x_k-E_x\right| \tag{5}$$

6. The calculation of the entropy

$$H_e=\sqrt{S^2-E_n^2} \tag{6}$$

In this work the characteristic vector $c=(E_x, E_n, H_e)$ of the user cloud is determined by the above mentioned algorithm.

Let us consider the following example of the construction of the user's interest profile according to the cloud model.

Example 1. Assume that a list of operations used to create the user profile is described in Table 1. Each score value applied to every action there represents the degree of the user's interest. The higher the score is, the deeper the interests that the user has on the selected target object are.

Table 1. List of operations

User operations		Score value
Permanent delete		1
Send to recycle bin		2
Weak Query		3
View the details of the characteristics	Permissions	3
	Sharing	
	Owner	
	Type	
Change the details of the characteristics	Permissions	4
	Sharing	
	Owner	
	Type	
Save		4
Copy and paste		5
Print		5

For example, printing a file implies a deep interest to the file, while sending it to the recycle bin indicates no interest to it. Among the characteristics showing the user's interest other types of characteristics can also be appropriate.

In order to identify the legality of the user who gained access to the cloud recourses, we will consider a cloud vector construction process.

Table 2. The actions of the users U_1 and U_2 on the cloud recourses

User operations		U1	U2
Permanent delete		√	
Send to recycle bin			
Weak Query		√	
View the details of the characteristics	Permissions	√	
	Sharing	√	
	Owner		
	Type		
Change the details of the characteristics	Permissions		√
	Sharing		√
	Owner		√
	Type	√	√
Save			√
Copy and paste			√
Print			√

The actions of the users U_1 and U_2 on the cloud recourses respectively are shown in Table 2. Based on the relevant digital values of each action conducted by the users U_1 and U_2 we can construct two vectors as follows:

$$V_1 = (1, 3, 3, 3, 4); \quad V_2 = (4, 4, 4, 4, 4, 5, 5)$$

By using above mentioned equations E_x, E_n, H_e we can construct a relevant interest cloud for both users as follows:

$$E_x = \overline{X} = \frac{1}{5}\sum_{k=1}^{5} x_5 = \frac{1}{5}\times(1+3+3+3+4) = \frac{1}{5}\times 14 = 2.8$$

$$E_n \sqrt{\frac{\pi}{2}} \times \frac{1}{5}\sum_{k=1}^{5}|x_5 - E_x| = \sqrt{\frac{\pi}{2}} \times \frac{1}{5}\times(|1-2.8| + |3-2.8| +$$
$$+ |3-2.8| + |3-2.8| + |4-2.8|) =$$
$$= \sqrt{\frac{\pi}{2}} \times \frac{1}{5}\times 3.6 = 1.25 \times 0.72 = 0.9$$

$$S^2 = \frac{1}{5-1}\sum_{k=1}^{5}\left(x_5 - \overline{X}\right)^2 = \frac{1}{4}\times((1-2.8)^2 + (3-2.8)^2 + (3-2.8)^2 + (3-2.8)^2 +$$
$$+ (4-2.8)^2) = \frac{1}{4}\times(3.24 + 0.04 + 0.04 + 0.04 + 1.44) =$$
$$= \frac{1}{4}\times 4.8 = 1.2$$
$$He = \sqrt{S^2 - E_n^{\ 2}} = \sqrt{1.2^2 - 0.9^2} = \sqrt{1.44 - 0.81} = 0.79$$

Thus, according to the calculations conducted above, the cloud vector for the first user can be noted as $C_1 = E_x, E_n, H_e = (2.8, 0.9, 0.79)$. Also by conducting the calculations for second user by the same way we can construct the cloud vector as $C_2 = (4.29, 0.51, 0.39)$.

According to the elements of this cloud, we can tell that U_1 is not the same user as U_2 and their interests to be present in the system also differ from each other. In other words, the interest of the user U_2 in comparison with U_1 becomes very malicious.

Only three characteristics (E_x, E_n, H_e) of the above determined two interest clouds show that these clouds do not belong to the same user. But by conducting certain calculations on this characteristic values and by comparing the value obtained in the result of the calculation with certain threshold value maybe we can predict that these vectors belong to the same user. In other words, if the obtained value becomes less than the threshold value, we can tell that these vectors belong to the same user. The determination of this difference value has become one of the target issues of this research. Note that in [14] this problem is not being solved. In the presented research the following approach is proposed in order to solve this problem.

4. Detecting Anomalous Behavior

For the detection of the anomalous behavior in the cloud infrastructure first of all let us provide comparison of the cloud vectors. Let the vector $c = (E_x, E_n, H_e)$ present the cloud of i-th user, then the similarity between the i-th and the j-th users is labeled as $\mathrm{sim}(\mathrm{i}, \mathrm{j})$ and is calculated as follows:

$$\mathrm{sim}(\mathrm{i}, \mathrm{j}) = \cos(c_i, c_j) = \frac{c_i \bullet c_j}{\|c_i\| \bullet \|c_j\|} \qquad (7)$$

where

$$\mathrm{c_i} = (Ex_i, En_i, He_i), \quad c_j = (Ex_j, En_j, He_j).$$

After the calculation of the similarity measure the detection module in Figure 1 classifies the input vector and identifies the access as normal or anomalous. For the implementation of this process the following method is proposed.

5. Classifying User Behavior

It should be noted here that even between the input data of the user and his own profile deviation cases can be take place. In order to accept the decision about normal or anomalous accesses it is necessary to calculate the degree of this deviation value. This deviation value can be calculated via following filtration formula:

$$D_{i,j} = \frac{\sum_{j \in N} \left(S_{i,j} \times Score_{i,j} \right)}{\sum_{j \in N} \left(\left| S_{i,j} \right| \right)} \quad (8)$$

Here $S_{i,j}$ - is a weight ratio, which represents the similarity value between the i -th and the j -th users; $Score_{i,j}$ - is a score value of the i -th operation used by the j -th user.

The classification of the users here fulfilled on the basis of the threshold value accepted in the system. If $D_{i,j}$ becomes less than the threshold value, then the system evaluates the user as genuine, otherwise he is evaluated as anomalous.

6. Existing Masquerade Datasets

There are numbers of datasets which enable to evaluate the performance of the masquerader attack detection techniques.

SEA dataset. Most papers about the masquerader detection use this dataset [15] with its associated configuration. The SEA consists of the commands collected from the UNIX account audit data. Among all fields of audit data provided by account only the username and the command were taken. The data describe 50 different users each issuing 15000 commands. The first 5000 commands are considered as genuine. The remaining 10000 commands of each user are divided into 100 blocks consisting of 100 commands each.

Greenberg dataset. This dataset [16] contains the data from 168 UNIX users. Users are classified into four groups: novice programmers, experienced programmers, computer scientists and non-programmers. The data are stored in the plain text files that record the following information: the session start and end times, the command line as entered by the user, the current working directory, any alias expansion of the previous command, an indication whether the line entered has a history expansion or not, and any error detected in the command line.

Purdue University dataset. Purdue or just PU dataset [17] consists of the UNIX command histories of 4 users of the Purdue Millennium Lab, collected in four months. A few works use this dataset and this may be due to its low number of users.

RUU dataset. This dataset was collected by Columbia IDS group [18] and consists of Windows commands. The dataset was collected from 34 normal volunteer users and 14 masquerade users. They model how normal users typically search a file system and use these models to detect the unusual searches that may be considered as masquerades. The dataset includes the records of the search conducted by the masquerader who should perform a specific task to find any data useful for his financial gain on a previously unknown file system within 15 minutes.

This datasets suffer partially or fully from the several deficiencies which prevent their adoption for the cloud environments. In this perspective, their most significant weakness is the lack of real masquerade data. No command sequence was issued by the attackers, only the RUU dataset includes the real masquerades but they are predefined limited scenario. The main shortcomings of these datasets are:

- The cloud systems are heterogeneous and the user audits may be distributed among VMs running distinct (e.g., Windows, Linux, and network) operating systems.
- The absence of the command arguments and other useful details such as when the user commands were issued, the duration of each user's session and the type of operations implemented by the users in the system.
- They are not suitable for training or testing any cloud detection systems because of their small size.
- Any efficient Cloud dataset should coverage for attacks in all cloud service models (SaaS, PaaS, and IaaS).

At present, there is no standard dataset allowing to test the masquerade attacks detection techniques in the cloud infrastructure. This relates to a necessity of special tools for the access of the cloud infrastructure, a necessity of special permissions, with distributions of audit data across different environments (e.g., Windows, Linux, and network), with necessity of a huge size of the audit data for cloud systems (more than 20GB). At feature it would be significant to provide investigations in this area to create such datasets.

7. Conclusion

In this paper detection method for the illegal access to the cloud infrastructure is proposed. Detection process is based on the collaborative filtering algorithm constructed on the cloud model. Here, first of all, the normal behavior of the user is formed in the shape of a cloud model, then these models are compared with each other by using the cosine similarity method and by applying the collaborative filtering method the deviations from the normal behavior are evaluated. If the deviation value is above than the threshold, the user who gained access to the system is evaluated as illegal, otherwise he is evaluated as a real user.

REFERENCES

[1] “Top Threats to Cloud Computing”, Cloud Security Alliance, https://cloudsecurityalliance.org/topthreats/csathreats.v1.0.pdf

[2] M. B. Salem, S. J. Stolfo. Data Collection and Analysis for Masquerade Attack Detection: Challenges and Lessons Learned, Columbia University Computer Science Technical Reports, Columbia University, 2011, 8 p.

[3] 2010 cybersecurity watch survey: cybercrime increasing

faster than some company defenses, CERT, 2010, 17 p.

[4] M. Arrington. In our inbox: Hundreds of confidential twitter documents, 2009, http://techcrunch.com/2009/07/14/in-our-inbox-hundreds-ofconfidential-twitter-documents/

[5] D. Takahashi. French hacker who leaked Twitter documents to TechCrunch is busted, 2010, http://venturebeat.com/2010/03/24/french-hacker-wholeaked-twitter-documents-to-techcrunch-is-busted/

[6] D. Danchev. ZDNET: french hacker gains access to twitter's admin panel, 2009, http://www.zdnet.com/blog/security/french-hacker-gains-access-totwitters-admin-panel/3292

[7] P. Allen. Obama's Twitter password revealed after french hacker arrested for breaking into U.S. president's account, 2010, http://www.dailymail.co.uk/news/article-1260488/Barack-Obamas-Twitter-password-revealed-French-hacker-arrested.html

[8] T. Lane, C.E. Brodley. Sequence matching and learning in anomaly detection for computer security, Proceedings of the AAAI Workshop on AI Approaches to Fraud Detection and Risk Management, AAAI Press, 1997, 43-49.

[9] R. M. Alguliev, F. C. Abdullayeva, "Identity management based security architecture of cloud computing on multi-agent systems," Proceedings of the Third International Conference on Innovative Computing Technology (INTECH), London, 29-31 Aug 2013, pp. 123–126.

[10] R. M. Əliquliyev, F. C. Abdullayeva, "Bulud texnologiyalarının təhlükəsizlik problemlərinin tədqiqi və analizi," İnformasiya Texnologiyaları Problemləri, 2013, №1(7), s. 3–14.

[11] S.E. Coull, J. Branch, B. Szymanski, E. Breimer. Intrusion detection: A bioinformatics approach, Proceedings of the 19th Annual Computer Security Applications Conference, 2003, 24-33.

[12] S.E. Coull, B.K. Szymanski. Sequence alignment for masquerade detection, Computational Statistics and Data Analysis, Vol. 52, No. 8, 2008, 4116-4131.

[13] S. J. Stolfo, M. B. Salem, A. D. Keromytis. Fog Computing: Mitigating Insider Data Theft Attacks in the Cloud, Proceedings of the IEEE Symposium on Security and Privacy Workshops, 2012, 125-128.

[14] X. Cheng, J. Chen. Modeling User Interests Based on Cloud Model for Masquerade Detection, Proceedings of the IEEE International Conference on Computational Intelligence and Software Engineering (CiSE), 2009, 1–4.

[15] http://www.schonlau.net/intrusion.html

[16] S. Greenberg. Using unix: Collected traces of 168 users. Report, Univercity of Calgary, 1988.

[17] T. Lane, C.E. Brodley. An application of machine learning to anomaly detection, Proceedings of the 20th National Information Systems Security Conference, 1997, 366-380.

[18] RUU dataset http://sneakers.cs.columbia.edu/ids/RUU/data/