International Journal of Control and Automation
Vol. 11, No. 6 (2018), pp.1-12
http//dx.doi.org/10.14257/ijca.2018.11.6.01

Scenario-based Log Dataset for Combating the Insider Threat

Seungwoo Kim, Jangju Kim, Dongwook Ha and Yeonseung Ryu*

Dept. of Security Management and Engineering, Graduate School, Myongji University
Cheoin-Gu, Yongin, Gyonggi-Do, Korea
ohokimsw@gmail.com, doujj@naver.com, hdu0105@gmail.com
**ysryu@mju.ac.kr*

Abstract

Insider threat is the most important security issue to various organizations in recent years. Insiders are people who can gain access to confidential information within the organization by using their legitimate access rights, and can cause enormous damage to the organization by leaking it to the outside. Various studies have been conducted to handle with insider threat such as scenario-based threat detection schemes. Recently, machine learning-based abnormal behavior detection has become the main focus of research. Most of machine learning schemes need enormous data to learn but there exist few dataset to deal with insider threat except for CERT dataset of Carnegie Mellon University. In this paper, we proposed log data based on actual insider threat events and various scenarios to define the log. The logs presented in this paper will help organize learning data for big data analysis and machine learning, and it is expected that more accurate insider threat detection will be possible.

Keywords: *Insider threat, log analysis, information leakage prevention, threat detection*

1. Introduction

Insider threat has become a serious security issue to various organizations including corporations as well as countries. According to the Korea Agency of Industrial Technology Security, 78.2% of the company's core technology thieves for the past five years were former and current employees. For examples of country-wide insider threats, the Edward Snowden case and the Chelsea Manning case are well known. Since insiders can steal information by using their legitimate access rights, it is difficult to know their malicious acts in advance. Further it takes long time to recover from damage after becoming aware of leakage. Figure 1 shows the time taken to resolve an attack [1]. It takes more time to solve the insider threat than other attacks.

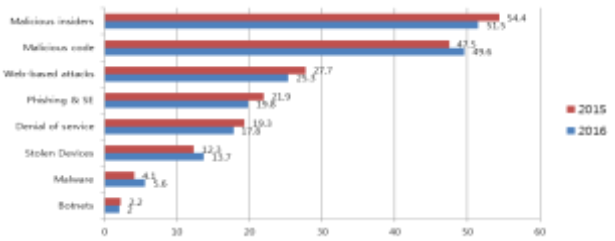


Figure 1. Time Required for Resolving Attacks [1]

Received (December 15, 2017), Review Result (March 8, 2018), Accepted (March 12, 2018)
* Corresponding Author

ISSN: 2005-4297 IJCA
Copyright © 2018 SERSC Australia

In order to combat the insider threat, many organizations have introduced a variety of security equipment to monitor user behavior and have heavily invested in security monitoring system. Representative equipment is SIEM (Security Information and Event Management). It is more intelligent and advanced than ESM (Enterprise Security Management) to cope with new security threats, and it can respond to latest security threat trend such as APT attack through big data analysis. Since SIEM manages information and events of all resources occurring in the organization, it can cope with various security threats if it is used well. However, despite the fact that we can use such equipment and know the importance of log analysis, it is difficult to determine exactly which logs to analyze to handle with insider threat. According to a survey of approximately 700 IT professionals, 88% of respondents indicated that insider threat is an important security problem, but they are struggling with what log to monitor [2]. While there have been a number of studies on this area such as scenario-based threat detection and machine learning-based threat detection, the most important 'log data' has not been studied in detail.

Log data is the basis for insider threat detection. Regardless of the detection method used, it is necessary to collect massive amounts of logs from insider PCs, servers, databases, network equipment, heterogeneous security solutions, etc., and process them into meaningful information after preprocessing. Recently the CERT (Computer Emergency Response Team) insider threat center at Carnegie Mellon University conducted an insider threat research project and published log dataset for insider threat studies [3]. Since then, several research teams made use of CERT dataset to study the insider threat issues and CERT dataset has become a de facto standard for insider threat dataset. [4] conducted a comparison study of the insider threat detection rates according to the CERT dataset version. [5] also used CERT dataset to detect abnormal behavior using outlier detection. [6] proposed an abnormal behavior detection scheme using RNN (Recurrent Neural Network) algorithm, one of the machine learning algorithms.

However, we found that though Carnegie Mellon University's CERT dataset is widely used for insider threat studies, there are many things that need to be supplemented. In this paper, we study actual insider threat scenarios and propose log dataset should be collected to detect threats according to each scenario. The rest of this paper organized as follows. First, we present related researches. We then describe which logs can be detected on which equipment according to each scenario. The logs presented in this paper will help organize learning data for big data analysis or machine learning, and it is expected that more accurate insider threat detection will be possible. Finally, we present the conclusion and future work.

2. Related Works

This section introduces previous insider threat researches with regard to detection schemes.

A. Behavior-based Detection

Behavior-based detection schemes should be based on insider job / role and predefine abnormal behavior. For example, we have blocked USB from being used by our internal security policy, but if an insider downloads certain data using USB, it can be considered a threat. [8] defined anomalous behavior in advance. They also studied abnormal behavior detection method by combining previously defined abnormal behavior elements with machine learning algorithms.

B. Scoring-based Detection

Scoring-based detection techniques define legitimate, normal and abnormal behaviors in advance based on the job / role of the insider, and assign weight to each abnormal

 Copyright © 2018 SERSC Australia

behavior item. Then, if the result obtained from the predefined formula exceeds a predetermined threshold value, it is regarded as a threat and detected. For example, suppose that a total of 5 points or more is defined as malicious behavior. It is assumed that the act of conducting work in the non-business hours is 1 point. It assumes 5 points for downloading files using unauthorized USB. If you download files using USB during non-business hours, you get 6 points and are classified as malicious. Scoring based detection can set weights and thresholds based on the company's internal situation. However, if weights or thresholds are not set properly, there is a possibility that it is not detected even though it is a serious threat. It is also an advanced technique of behavior-based detection because it is a technique that gives weight to abnormal behaviors.

C. Machine-learning based detection

Recently, the most detection researches are focusing on machine learning based technique. There are two types of detection methods based on machine learning: map learning method and non-supervised learning method. Many detection schemes using non-instructional learning have also been studied. The use of anomaly detection technique has the advantage of detecting new types of threats compared to existing statistical based detection. [9] demonstrated that the non-supervised learning model has a higher detection rate than the supervised learning model. [10] conducted a study using HMM (Hidden Markov Model) This was a novel attempt. [11] conducted a research to combine and detect several models to solve the disadvantage that machine learning based detection can be more false than other detection methods.

There have also been studies using various machine learning techniques such as clustering. A study based on machine learning learns normative behaviors of employees from machine learning models and takes into account deviations between abnormal behaviors and normal behaviors. A variety of logs are required to learn the normal behavior of an employee. The accuracy of detection can change depending on what logs are detected. It is therefore important to define appropriate detection factors and to collect appropriate logs.

D. ETC.

The above three detection categories are classified by the detection methodology. In addition, [12] emphasized that human behavior and personality are important for insider detection, and research on how to improve the accuracy of detection by combining human psychological profile information with detection models is underway. And [13] pointed out the importance of insiders on personal information leakage accidents that can occur in inter-company transactions.

3. Outflow Path and Scenarios

This section introduces the way in which internal information is leaked and scenarios of actual malicious behavior. The scenarios that we introduce are based on actual cases and scenarios mentioned in various studies.

A. Outflow path of insiders

Before presenting the scenarios, we examined the most common leakage paths based on actual events. Malicious insiders can use various tools to achieve their goals. The most frequently used tool is 'Thumb drive'. A Thumb drive refers to a portable storage device that can be connected to a computer USB port, such as a USB drive. Recent products are very small in size and have various shapes. Malicious insiders can use this feature to secretly leak large amounts of secrets.

Copyright © 2018 SERSC Australia

E-mail is also one of the most frequently used information leaks path. Many malicious insiders leak information using the attachment features of email. This is because the amount of attachments that can be attached to emails has increased compared to the past, and more information can be leaked at once.

'Cyber attacks' cannot be overlooked. In the past, it was a cyber attack that hackers attacked the system for their own purposes. However, in recent years, a variety of hacking tools have been easily downloaded from the Internet, allowing insiders to leak information using hacking tools that are appropriate for their purposes.

There are various 'Web storages' on the Internet these days such as cloud, blog, and private homepages. These spaces can be the path of the outflow. Employees sometimes upload company information to these Web storages for the convenience of their work. This is not the intention, but it is also the route through which the company's secrets are leaked.

In addition, there are a variety of outgoing routes such as CD / DVD storage devices, messenger attachments, and P2P services. A malicious insider takes some action to steal company information. For example, an insider accesses an information DB to download large amounts of files to steal the organization's secrets. Thereafter, it may be compressed into a single file using a compression program, copied to a USB drive, and then exported to the outside of the company.

If we examine the actions performed by the insider, we can find behavioral indicators performing by the insider when maliciously taking information. The following threat detection factors are based on actual cases in our study.

a. Overtime working
b. Upload files to external web storage
c. Output more documents than usual
d. Frequently copy files to USB
e. Send mail with large attachments
f. Remote access
g. Malicious program download (hack tool)
h. Access to employment site
i. Attach file with messenger

B. Insider threat Scenarios

In this sub-section, we present six possible threat scenarios based on the above-mentioned threat behavior indicators and frequently used outflow routes. We include as many different scenarios as possible and reconstruct them based on actual events. Figure 2 shows the relation of insider threat scenarios and threat detection factors.

A. The employee copies the company secrets via USB over the off-hours and attempts to leak them.

B. The insider prepares to move the company. When the insider decides to move the company, insider tries to leak the company's confidential information to the external E-mail.

C. An employee whose contract has expired installs a malicious program (back door, key logger *etc.*,) just before leaving the company and leaks the company data after leaving the company.

D. For the sake of convenience, several employees who perform the same job upload a file containing customer information to a web space (dropbox, blog, *etc.*,) that employees can share with each other, and this information is leaked.

 Copyright © 2018 SERSC Australia

E. The insider transmits the confidential file of the company by using the messenger, and downloads it from the outside.

F. The insider comes out to the company on the weekend and prints out documents containing the company's key information and leaks them out. During this process, more documents were output than usual.

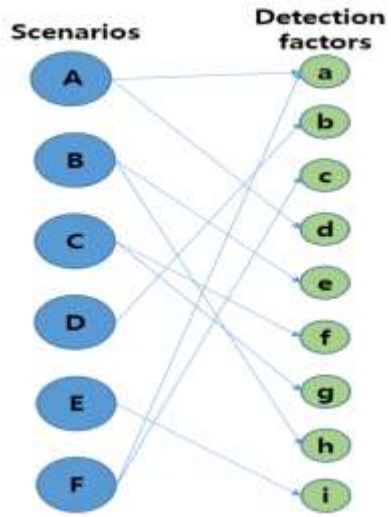


Figure 2. Relation between Scenarios and Detection Factors

4. Proposed Log Dataset

As described in previous section, there are various methods to detect insider threats. Most of the detection methods use the user's log data to determine the user's behavior through a series of processes (formula, machine learning, *etc.*,) and detect the threat. Therefore, whatever detection method is used, proper log data collection is essential for better detection performance.

First, we describe the log sources and logs needed to detect the above scenarios. Later, we will explain the differences between our proposed dataset and the CERT dataset of Carnegie Mellon University.

A. Log data for combating insider threat

This sub-section describes the log sources and logs needed to detect the proposed insider threat scenarios. There are various fields in the original logs, but in our study, we only describe the key fields for insider threat detection. Windows logs and Linux-based OS logs are presented separately, but some logs are presented together.

i. Overtime Working

We can detect overtime working by monitoring logon and logoff events from OS.

Table.1. Windows OS Log on/off Event Log

Fields	**Description**
Source	Microsoft Windows security.
Log Name	Security.
Keyword	Audit success.
Event ID: 4624	Successful Logon Completed.
Event ID: 4634	Account Logoff Completed Log.

Copyright © 2018 SERSC Australia

Table 2. Linux OS Log on/off Log (Kali-linux 2016.1)

Source	/var/log/wtmp
Description	Binary file with login information recorded system. Includes on / off / reboot history
Fields	Account, terminal device name, remote access IP, remote host name for remote login, user login / logout time

ii. Upload Files to External Web Storage

Uploading a file means transmitting it to external storages such as an external blog, a cafe, or a personal homepage using the Internet. Web upload can be detected by using network traffic logs. However, most storage services support SSL encrypted communications. That is, the network packet is encrypted so that it can not be determined whether the file is being uploaded or downloaded, and it is not possible to detect which file is transmitted. Some DLP (Data Loss Protection) solutions provide detection and blocking capabilities for file uploads to instant messengers or web hard drives. The principle is that its encrypted packets are decrypted using HTTPS/FTP proxy.

Figure 3. SSL Communication Visualization using Proxy

Table 3. Features Detected when Uploading Data

Fields	Description
Traffic flow	Insider IP → Uploading server IP
HTTP Method	POST / PUT Method
FTP Protocol	FTP Port(20/21) Ex) ftp://ftp.aaa.com/test.zip
URL signature	File extensions included in url : .xls / .doc / .pdf / .hwp / .zip / .egg etc.

HTTPS/FTP Proxy should be used to ensure the visibility of encrypted traffic as shown in Figure 3. All data is passed through the HTTPS/FTP Proxy before the service requested by the insider is delivered to the Web server. The Proxy decrypts the encrypted communication, examines the leakage of personal information, encrypts it again and transmits it to the web server. To use HTTPS/FTP Proxy, we need SSL certificate of HTTPS/FTP Proxy on the insider PC.

iii. Output more documents than usual

Recent network printers are able to leave output logs. For multifunction printers, not only document printing logs but also copy logs are left.

 Copyright © 2018 SERSC Australia

Table 4. Fuji Xerox Printer Log

Fields	Description
Type of work	Printer / Copy
File name	Document file name to output. Extension.
User name	PC Host name.
Number of output pages	
Result	Success / Fail
Start time	00/00 00:00AM
Finish time	00/00 00:00+M AM

iv. Frequently copy files to USB

Table 5. Windows OS USB Log

Fields	Description
Source	DriverFrameworks-UserMod.
Log name	Microsoft-Windows-DriverFrameworks-UserMode/Operational
Keyword	-
Event ID: 2101	When USB connection, driver installation completed log (USB manufacturers can be identified).
Event ID: 1008	When disconnecting USB, driver process termination log can be checked.
Event ID: 2102, 2100	

Table 6. Linux OS USB Log (Kali-linux 2016.1)

Source	/var/log/kern.log /var/log/syslog
Description	A log file containing device connection and driver installation information. Various logs other than USB connection log.
Fields	Connect: Vendor name, Product name, connect time. Disconnect: Volume label, Disconnect time.

v. Send email with large attachments

OS does not leave any mail related logs, so we should refer to the mail server's log. Below is a log entry that can be referenced by the mail server.

Table 7. Microsoft Exchange Server 2003 Mail Server Log

Fields	Description
Subject(Mail subject)	Field where the mail subject exists.
To (Recipient)	Recipient mail address exists.
From (Sender)	The sender email address is present
Reception / Sending Time	Reception / Outgoing Time (Time registered on the mail server).
File Type (Extension type of attached file)	Identification of attached file name and extension of attached file.

Copyright © 2018 SERSC Australia

vi. Remote access

When a remote connection is made from the outside to the business PC, information leakage becomes possible. Conversely, attempts to connect to the outside from a business PC should also be detected. This is because if you try to access a server PC or DB from a business PC and succeed, you can bring sensitive information. In the case of the Windows OS, a log of the connection attempt to the destination IP, connection success / failure, connection termination, *etc.*, is left. In the case of Linux OS, IP information which tried remote connection (telnet, SSH, FTP) is recorded in wtmp file.

Table 8. Windows OS Remote Connection(RDP) Log

Fields	**Description**
Source	TerminalServices-ClientActiveXCore.
Log name	Microsoft-Windows-TerminalServices-RDPClient/Operational.
Category	Connection sequence.
Event ID: 1024	RDP ClientActiveX tries to connect to server (x.x.x.x). (It is possible to detect connection attempt from insider → outside).
Event ID: 1026	RDP ClientActiveX is disconnected (RDP connection failure or detection at connection termination log).

Table 9. Linux OS Remote Connection Log

Log path	**/var/log/wtmp**
Description	Binary file with login information recorded Includes system on / off / reboot history.
Fields	Account, terminal device name, remote access IP, remote host name for remote log-in, user log-in / log-out time.

vii. Malicious program download (hack tool)

Malicious program has intention of harmful actions such as compromise, alteration, falsification, leakage, or destruction of computer data or programs. The types of malicious programs vary depending on the hacking method such as backdoor program, key logging, and brute force.

There are two types of cases where insiders download malicious programs. First, malicious files are downloaded to the insider's PC by hacker's attack. Secondly, an insider intentionally downloads a malicious program for information leakage. Both cases can be regarded as threats in terms of information leakage. As a basic function of the OS, malicious program download detection is impossible. It should be detected in the form of receiving a log from the vaccine after installing the specific vaccine. Windows / Mac / Linux all leave the same type of logs. Table 10 shows the log detected by the vaccine.

 Copyright © 2018 SERSC Australia

Table 10. Vaccine Log (Windows / Mac OS / Linux OS)

Fields	Description
Threat name	Threats Defined by Vaccine Category (Eg Trojan Malware, Phishing SW, Ransome APP, etc.).
File size	If a malicious file is detected, the size of the detected file.
User's name	PC Host name.
Detected User IP	IP of detected insider.
Actual work	Malicious file processing result. *Left alone: Detected malicious files are not deleted and are present on the insider PC. *Deleted: Detected malicious files are removed by the vaccine. *Guaranteed: Detected malicious files are guaranteed by the vaccine, etc.
MD5/ SHA256	MD5 and SHA256 values of detected malicious files.
File path	The path where the detected malicious file exists (in the insider PC).
Detection time	Wed JUL 00 00:00:00 KST 2017.
OS type	Windows / Mac OS / Linux OS

viii. Access to employment company site

The problem of information leakage from employees who are leaving is known to be serious. In fact, insiders who decide to turnover can find patterns of behavior that are frequently accessed on job sites. In other words, if the frequency of access to the job search site is frequent, it can be judged as an insider who has decided to turnover. In the case of insiders who want to change jobs, it is necessary to pay close attention because they often try to leave the company in exchange for complaints or leakage of internal information.

That is, the act of accessing the job searching site itself cannot be judged as a malicious act but it can be judged as a behavioral indicator which is likely to lead to malicious action. Logs for access to the employment searching site will be logged in the Firewall log on Windows OS. Windows Firewall logs can be detected after registering IP of known employment searching site because it saved on IP basis.

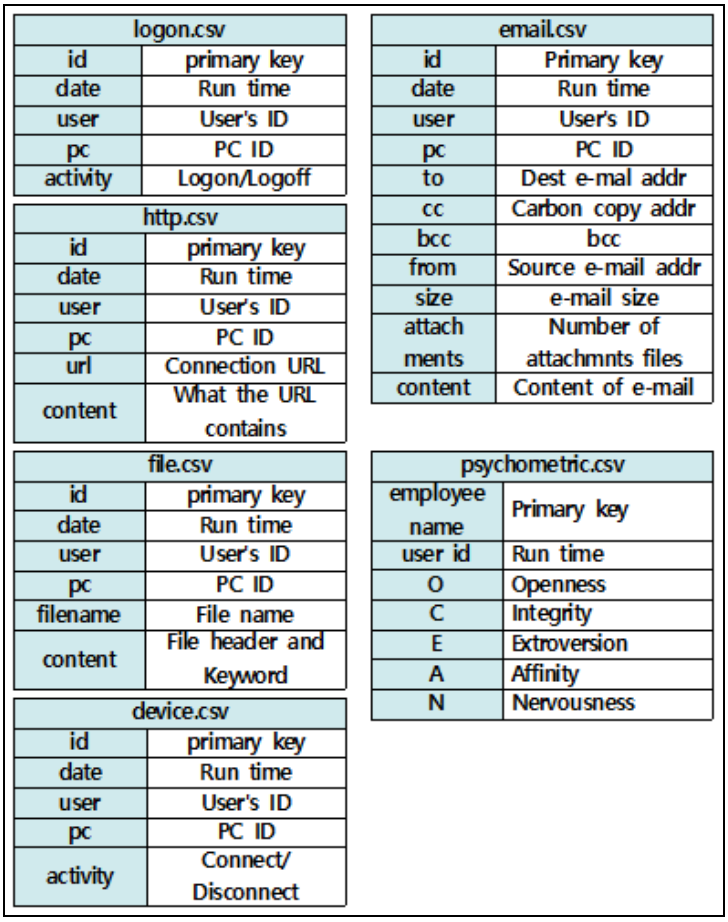
Table 11. Windows OS Firewall Log

Fields	Description
Software	Microsoft Windows Firewall.
Date & time	2000-00-00 00:00:00
Action	Allow / Drop Action log indicating connection status.
Protocol	TCP / UDP
SrcIP / Dest IP	Source / Destination IP information.
Src Port / Dest Port	Source/ destination Port information.
TCP Flags	TCP SYN / ACK / SYNACK / FIN Flags information in decimal representation.
Information	Send / Receive

Copyright © 2018 SERSC Australia

ix. Attach file with messenger

The internal information can be leaked by the messenger's file transfer function. There are two types of messenger communication. The first is text communication and the second is file transmission / reception communication. First, it is found that most messengers send and receive text data through P2P connection. Next, in case of file transmission/reception communication, when a sender transmits a file to a messenger server, the file is stored in the messenger server and the receiver downloads the file from the messenger server. In case of the file transmission/reception log, the same log as the above-mentioned item B can be used.



logon.csv	
id	primary key
date	Run time
user	User's ID
pc	PC ID
activity	Logon/Logoff

http.csv	
id	primary key
date	Run time
user	User's ID
pc	PC ID
url	Connection URL
content	What the URL contains

file.csv	
id	primary key
date	Run time
user	User's ID
pc	PC ID
filename	File name
content	File header and Keyword

device.csv	
id	primary key
date	Run time
user	User's ID
pc	PC ID
activity	Connect/ Disconnect

email.csv	
id	Primary key
date	Run time
user	User's ID
pc	PC ID
to	Dest e-mal addr
cc	Carbon copy addr
bcc	bcc
from	Source e-mail addr
size	e-mail size
attachments	Number of attachmnts files
content	Content of e-mail

psychometric.csv	
employee name	Primary key
user id	Run time
O	Openness
C	Integrity
E	Extroversion
A	Affinity
N	Nervousness

Figure 4. Configuration of CERT Dataset

B. Comparison with CERT dataset

We compare the difference between our proposed dataset and CERT dataset of Carnegie Mellon University.

First, we can compare log dataset format. Our proposed dataset format is described in previous section. In case of CERT dataset, Figure 4 shows the format of CERT dataset. Dataset consists of logon, file, email, device, http, and psychometric. Each dataset log file is stored in .CSV format.

Secondly, we can compare the coverage of insider threat cases. Table 12 shows the coverage difference between the proposed dataset and CERT dataset. While CERT dataset has only 5 scenarios, the proposed dataset has 11 scenarios. Therefore, the proposed dataset defines more insider threat cases.

 Copyright © 2018 SERSC Australia

Table 12. Comparison of Proposed Dataset with CERT Dataset

	Proposed dataset	CERT dataset
1	Overtime Working(logon/logoff)	Logon.csv(logon/logoff)
2	Upload files to external Web storages	
3	Output more documents than usual (Print and Copy log)	
4	Frequently copy files to USB (USB install/disconnect log)	Device.csv (USB Connect/disconnect)
5	Send e-mail with large attachments(e-mail server detection log)	
6	Remote access (RDP detection log)	
7	Malicious program download (Vaccine program detection log)	
8	Access to employment site (Windows OS firewall log)	http.csv
9	Attach file with messenger	
10		File.csv
11		Psychometric.csv

5. Conclusions

In modern society, organizations have to produce and store the confidential information such as trade secret information, critical technical information, nation secret, and so on. The confidential information is the core competitiveness of organization and can cause enormous damage to the organization if it leaks to the competitor. Hence, insider threat is now the biggest security issue in any organization. There are a number of researches on various insider threat combating techniques including threat detection scheme. However, there have been few researches on log dataset for combating the insider threat, which is the most important part of insider threat detection.

In this paper, we proposed some possible actual insider threat scenarios and insider's log dataset based on proposed actual scenarios. In doing so, we investigated actual cases and related researches, and extracted nine behavioral factors, and suggested six possible scenarios of insider threats. We proposed log dataset that can be collected from various devices such as mail server and printer as well as insider's personal PC. We compared the proposed dataset with Carnegie Mellon University's CERT dataset, which is the most widely used in the field of insider threat research. In the future, we will verify the proposed log dataset by using it in the insider threat detection methods.

Acknowledgement

This work was supported by 2017 Research Fund of Myongji University.

References

[1] Penemon Institute, "2016 Cost of Cyber Crime Study & the Risk of Business Innovation", **(2016)**, pp. 11.
[2] Penomon Institute, "Privileged User Abuse & The Insider Threat", **(2014)**, pp. 1.
[3] The CERT Insider Threat Center at Carnegie Mellon University, https://www.cert.org/insider-threat/.
[4] K. Kang, "Study on Big Data Log Analysis for Insider Threat Detection", Proc. 2017 Summer Conference of Korea Institute of Information Security and Cryptography, **(2017)**.

Copyright © 2018 SERSC Australia

[5] J. Kim, M. Park, H. Kim, S. Cho and P. Kang, "Development of Insider Threat Detection Method Using Outlier Detection", Proc. 2016 Korea Industrial Technology Association Fall Conference and Regular General Meeting, **(2016)**, pp. 33.
[6] D. Ha, K. Kang and Y. Ryu, "Detecting Insider Threat Based on Machine Learning: Anomaly Detection Using RNN Autoencoder", Journal of Korea Institute of Information Security and Cryptography, vol. 27, no. 4, **(2017)**, pp. 763-773.
[7] S. Kim, J. Kim, D. Ha and Y. Ryu, "Carnegie Mellon University's CERT dataset Analysis and suggestions", Proc. 5th International Conference on Interdisciplinary Research Theory and Technology, **(2017)**.
[8] Y. H. Lim, J. S. Hong, K. H. Kook and W. H. Park, "A Study on Insider Behavior Scoring System to Prevent Data Leaks", Journal of the Information and Security, vol. 15, no. 5, **(2015)**, pp. 77-86.
[9] P. Parveen and B. Thuraisingham, "Unsupervised Incremental Sequence Learning for Insider Threat Detection", Proc. IEEE International Conference on Intelligence and Security Informatics, **(2012)**.
[10] T. Rashid and I. Agrafiotis, "A New Take on Detecting Insider Threats: Exploring the use of Hidden Markov Models", Proc. 8th ACM CCS International Workshop on Managing Insider Security Threats, **(2016)**, pp. 47-56.
[11] P. Parveen, J. Evans and B. Thuraisingham, "Insider Threat Detection Using Stream Mining and Graph Mining", Proc. IEEE Third International Conference on Privacy, Security, Risk and Trust, **(2011)**.
[12] O. Bradiczka, J. Liu, B. Price, J. Shen, A. Patil, R. Chow, E. Bart and N. Ducheneaut, "Proactive insider Threat Detection through Graph Learning and Psychological Context", Proc. 2012 IEEE Symposium on Security and Privacy Workshops, **(2012)** May, pp. 142-149.
[13] D. Lee and J. Jeong, "A Case Study of Employee Privacy leaks and Fraud during B2B transaction - Focused on Man in the Middle Attack Case", Journal of Security Engineering, vol. 12, no. 5, **(2015)**, pp. 501-514.

 Copyright © 2018 SERSC Australia