Big Data Analysis System Concept for Detecting Unknown Attacks

Sung-Hwan Ahn*, Nam-Uk Kim*, Tai-Myoung Chung**

*Department of Electrical and Computer Engineering, Sungkyunkwan University

** College of Information and Communication Engineering, Sungkyunkwan University

shahn@imtl.skku.ac.kr, nukim@imtl.skku.ac.kr, tmchung@ece.skku.edu

Abstract— **Recently, threat of previously unknown cyber-attacks are increasing because existing security systems are not able to detect them. Past cyber-attacks had simple purposes of leaking personal information by attacking the PC or destroying the system. However, the goal of recent hacking attacks has changed from leaking information and destruction of services to attacking large-scale systems such as critical infrastructures and state agencies. In the other words, existing defence technologies to counter these attacks are based on pattern matching methods which are very limited. Because of this fact, in the event of new and previously unknown attacks, detection rate becomes very low and false negative increases. To defend against these unknown attacks, which cannot be detected with existing technology, we propose a new model based on big data analysis techniques that can extract information from a variety of sources to detect future attacks. We expect our model to be the basis of the future Advanced Persistent Threat(APT) detection and prevention system implementations.**

Keywords— **Computer crime, Alarm systems, Intrusion detection, Data mining**

I. INTRODUCTION

According to the Gartner report, sophisticated hacking attacks are continuously increasing in the cyber space[1]. Hacking in the past leaked personal information or were done for just fame, but recent hacking targets companies, government agencies. This kind of attack is commonly called APT(Advanced Persistent Threat). APT targets a specific system and analyses vulnerabilities of the system for a long time. Therefore it is hard to prevent and detect APT than traditional attacks and could result massive damage[2].

Up to today, detection and protection systems for defending against cyber-attacks were firewalls, intrusion detection systems, intrusion prevention systems, anti-viruses solutions, database encryption, DRM solutions and etc. Moreover, integrated monitoring technologies for managing system logs were used. These security solutions are developed based on signatures and blacklist. However, according to various reports, intrusion detection systems and intrusion prevention systems are not capable of protecting systems against APT attacks because there are no signatures. Therefore to overcome this issue, security communities are beginning to apply heuristic and data mining technologies to detect previously unknown attacks.

In this paper, we propose a new model based on big data analysis technology to prevent and detect previously unknown APT attacks. We compared previous researches which are based on data mining technology for predicting or analysing correlation between attack behaviours and explained its limits. Furthermore we list various sources and their details that can be collected and explain attack predictions earned from applying big data technologies such as classification, text mining, clustering, and association rules. Finally, we propose new attack reaction model based on big data technologies and evaluate the model. We expect this research to be the basis for future implementation of APT attack detection and prevention systems based on big data analysis technologies.

II. RELATED WORK

In the current cyber-space, sophisticated and intelligent threats are increasing. These previously unknown attacks cannot be detect or mitigated using existing pattern matching methods such as signature, rule, and black list based solutions.

For this reason, the anti-virus industry and research groups are combining technologies such as heuristic detection techniques and data mining techniques to detect attacks that are not detected with existing pattern-matching techniques.

In this chapter, we explain the definition and characteristics of an APT attack and existing security solutions as well as big data analysis which is arising as a future solution for detecting unknown attacks.

A. APT Attacks

APT attack is a special kind of attack that use social engineering, zero day vulnerabilities and other techniques to penetrate into the target system and persistently collect valuable information. It can give massive damage to national agencies or enterprises[2].

Recent APT attacks tend to target core industrial control systems instead of ordinary desktops or servers. Moreover, APT attacks are used as cyber weapons between nations. Cyber security is becoming a core aspect of national safety. Attacking industrial systems and causing malfunctioning of theses infrastructures can cause public chaos to the nation.

Examples of recent APT attacks are Stuxnet, RSA Secure ID hacking and the Night Dragon. Stuxnet was a very intelligent malware that was developed to attack Iran's nuclear facilities and make them malfunction. It is known to be acting in the wild for a few years until it was discovered by security researchers.

APT attack is usually done in four steps: intrusion, searching, collection and attack. Figure 1 describes the attack process in detail.

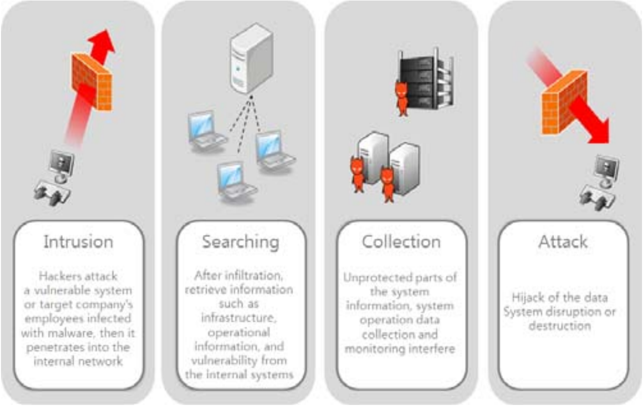
Intrusion
Hackers attack a vulnerable system or target company's employees infected with malware, then it penetrates into the internal network
Searching
After infiltration, retrieve information such as infrastructure, operational information, and vulnerability from the internal systems
Collection
Unprotected parts of the system information, system operation data collection and monitoring interfere
Attack
Hijack of the data System disruption or destruction

Figure 1. The sequence of APT attacks

In the intrusion step of an APT attack, the hacker probes for information about the target system and prepares the attack. To get the access to the system, the attacker searches for users with high access privileges such as administrators and use various attack techniques such as SQL injection, phishing, farming and social engineering to hijack their accounts

Searching is done after the hacker gained access to the system. Hacker analyses system data such as system log for valuable information and look for security vulnerabilities than can be exploited for further malicious behaviours.

In the next step, after the hacker has located valuable information in the system such as confidential documents etc, then, he installs malwares such as rootkits, backdoors to collect system data and maintain system access for the future.

In the final step, the hacker leaks data and destroys target system using acquired privileges. Leaked information can be used for developing other additional security vulnerability exploits. Because APT exploits use zero-day vulnerabilities and obfuscation methods, Anti-Virus program, IDS and IPS are difficult to detect such exploits.

B. Existing Information Security Technologies

Security researchers developed various security technologies to protect the system from evolving attacks. Typical solutions are firewall, IDS/IPS, WAF(Web Application Firewall), ESM(Enterprise Security Management).

1) Firewall: Firewall is a regulation device that controls the network traffic between separated networks and hosts. It is security technology based on access control. It decides whether to allow an access to the internal IP addresses and port numbers. Administrator sets up this access control rules in advance[3].

Initial firewall is located at border of the network and can be used as a protector for the inner network. Also, firewall is used as a primary security solution to this day. Firewall has a simple rule system that allows administrators to control firewall easily. However, firewall cannot detect and analyse threats in the network but just blocks accesses according to IP addresses and port numbers defined by administrators. Therefore a firewall only provides minimum protection from attacks.

2) IDS: IDS is a report system that searches and reports threats according to defined rules and captured traffic. IDS observes and analyses network traffic and detects malicious traffic and unprivileged file access. IDS can be divided into NIDS(Network-based Intrusion Detection System) and HIDS(Host-based Intrusion Detection System). NIDS can observe and analyse traffic in a more higher level than firewall using sniffing technology[4].

HIDS can observe system states such as file modification, file access, process in the host. In contrast to NIDS, HIDS is installed on each host. It detects abused resources and unprivileged access and reports to administrator.

IDS detect and alerts abnormal actions by pre-defined rule. These rules are based on normal users' behaviours and statistic information from system logs.

3) WAF: A website is good hacking target due to publicly opened services. Web servers use a fixed port number on each service. Therefore, a security administrator cannot just block these port numbers to stop malicious access because it needs to provide service to legitimate connections. Furthermore, it is difficult for IDS to analyse encapsulated application data in the packet. Therefore, existing security solutions cannot effectively detect threats in application level in the network stack.

WAF detects and blocks attacks using both positive and negative access control. Positive access control is a technology that blocks everything except defined safe patterns, and negative access control blocks only predefined malicious patterns[5].

Security technologies such as firewall, IDS, WAF basically use pattern matching techniques that are based on pre-defined rules. Therefore, they may not detect or prevent attacks that are encrypted or obfuscated. Also, they have some problems regarding low-performance and false-positives by duplicated rules.

C. Big Data Analysis

Big data has been a great issue in the IT industry for the last couple of years. It defines huge, shortly created and atypical data in digital environment such as text, music, video, and so on. Big data analysis is a technology that searches useful information such as a relation rule, a hidden value from huge data[6].

Big data analysis uses various existing analysis techniques such as machine-learning, artificial-intelligence, data-mining and etc. Among various techniques, we focus on four techniques – prediction, classification, relation rule, atypical data-mining. We think that these techniques are useful to detect unknown new attacks.

First, prediction is a technique that predicts the future possibility and trend. Regression analysis is a representative prediction technique. Researchers can predict attack possibilities using regressing analysis. Regressing analysis can predict similar behaviours from collected attack logs.

Second, classification is a technique that predicts the group of new attack from huge data. Classification helps security administrator to decide direction of protection and analysis. Most used classification techniques are logistic regression analysis and SVM(Support Vector Machine).

Third, relation rule is a technique that discovers hidden relations among data. The action of discovering relation rule is named association analysis or link analysis. The relation from time flow is named as sequence rule. This analysis technique can determine abnormal behaviour by analysing user or process behaviours.

Lastly, atypical data-mining analyses data that cannot be expressed in numbers such as picture, video, audio, text and etc. Typical atypical data-mining techniques are text-mining, web-mining and social-mining.

III. Big Data Analysis System Model

Previously unknown attacks such as APT are evolving to bypass existing security measures. These attacks are impossible to detect or prevent with current technologies which are explained in the Chapter 2. Therefore security incidents constantly occurs using state-of-the-art attack technologies. New security paradigm to react to these attacks is in need. The new paradigm requires big data analysis techniques as a core and integration of defence technologies, central security management, incident prediction technologies. We propose a system model that uses big data analysis technology for extracting data from various sources to react to previously unknown attacks.

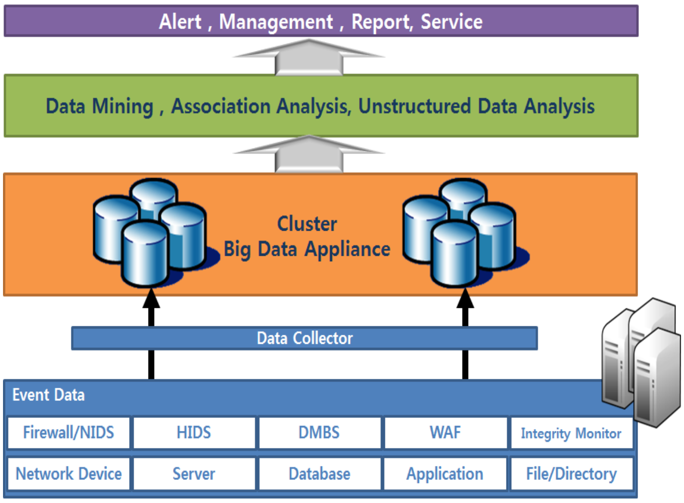

Figure 2. Big Data Analysis System Architecture

As seen in figure 2, entire system is divided into 4 steps.

- Data Collection: Data collection step collects event data from firewalls and log, behaviour, status information (date, time, inbound/outbound packet, daemon log, user behaviour, process information etc.) from anti-virus, database, network device and system. Collected data is saved in big data appliance

- Data Processing: This step validates whether collected data satisfies certain requirements. Then key value pair is created and classified using No-SQL, Hadoop, Mapreduce and etc. It is known that approximately 80 precent of time required for collecting and processing data using data mining is needed. For faster processing, we introduce cloud or distributed system.

- Data Analysis: Pre-processed data from previous step is analysed using prediction, classification, association analysis, and unstructured data analysis to decide user behaviour, system status, packet integrity and misuse of file or system. Used mining technologies are explained in chapter 2.

- Result: If attack or abnormal behaviours are detected, it alarms the administrator and terminates. Moreover, we provide dashboard, management tools to monitor results in real time. Prediction information of analysed system is summarized and reported to the manager. Also configuration update, rule manipulation and deletion, analysis pattern updates are done both automatically and passively.

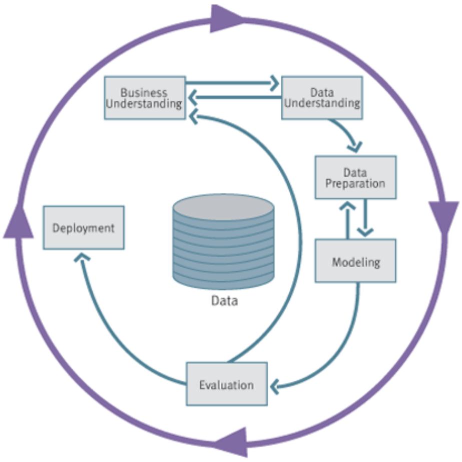

Figure 3. Phases of the CRISP-DM Process Model

Standard process for data mining techniques which are used for big data analysis was proposed by Chapman et al. in 2000[7]. This is called "CRISP-DM Process Model". Proposed standard process is composed of business understanding, data understanding, data preparation, modelling, evaluation, deployment. We analyse and give feedbacks according to this standard process.

We can predict ‘nowcast’ rather than forecast which a near future using big data analysis model we proposed. We expect to extract relation between behaviours and normal/abnormal patterns using big data analysis of logs

TABLE 1. BIG DATA ANALYSIS BASED APPLICATIONS

Applications	Description
Real-Time Monitoring	From a variety of sources, data collection, management, and the state of the system in real time And, the attack by the application to track and predict or monitor users' behaviour
Threat Intelligence	Anomaly detection of threats and attack patterns to be able to date information on management
Behaviour Profiling	Observe the behaviour of the system and the user Tracking and investigating for suspicious packets or behaviours
Data & User Monitoring	Continuous monitoring for the protection of users and sensitive data Prevent misuse of data and computing resources
Application Monitoring	Continuous monitoring for the behaviours of applications and system processes Process behaviours can be an important factor in detecting malicious behaviour
Analytics	Linkage analysis for a variety of monitoring information Infers the possibility of an attack

However for applying multi-source data monitoring analysis technologies in the security area, it requires real-time monitoring, context sensitive behaviour detection, and automatic analysis technologies.

In this paper we are not proposing effective parallel processing algorithm for real time analysis. Instead of using pattern matching or log analysis for predicting cyber-attacks, we believe that we can extract valuable information previously unfound from data and status information collected from various sources by big data analysis.

Moreover, to apply and validate various analysis methodologies using big data, we need professional software and distributed system. In future works, we plan to research on how to implement proposed system and get results using real factors and analysis methodologies.

IV. CONCLUSIONS

Recent unknown attacks easily bypass existing security solutions by using encryption and obfuscation. Therefore new detection methods for reacting to such attacks are in need.

In this paper we proposed big data system model for reacting to previously unknown cyber threats and researched on the deduction of practical technologies. In future works, following researches must be done:

- Classification of data by context of intrusion detection
- Implementation of data relation analysis methodology and its abnormal behaviour detection strategy
- Quantitive and qualitive assessment of proposed model and performance evaluation.

REFERENCES

[1] J. Feiman, “Hype Cycle for Application Security, 2012”, Gartner Group, July, 2012.

[2] “Advanced Persistent Threat: A Decade in Review”, Command Five Pty Ltd, June, 2011.

[3] K. Ingham and S. Forrest, “A History and Survey Network Firewalls”, University of New Mexico, Tech. Rep., 2002.

[4] R. D. Pietro and L. V. Mancini, Intrusion detection systems, in: S. Jajodia (Series editor), Handbook of Advances in Information Security, Springer, 2008..

[5] The OWASP(Open Web Application Security Project) website., Web Application Firewall, Available: http://www.owasp.org/

[6] R. Magoulas and B. Lorica, “Introduction to Big Data”, Release 2.0 (Sebastopol O’Reilly Media), Feb, 2009.

[7] P. Chapman. et al, “CRISP-DM 1.0 – Step-by-step data mining guide”, http://www.crisp-dm.org (2000).