Cyber-attack group analysis method based on association of cyber-attack information

Kyung-ho Son [1], Byung-ik Kim[2] and Tae-jin Lee[3,*]
[1] Division of Liberal Studies, Kangwon University
Republic of Korea
[e-mail: khson@kangwon.ac.kr]
[2] Security Threat Response R&D Team, Information Security Industry Group, Korea Internet & Security Agency
Republic of Korea
[e-mail: kbi1983@kisa.or.kr]
[3] Department of Computer Engineering, Hoseo University
Republic of Korea
[e-mail: kinjecs0@gmail.com]
*Corresponding author: Tae-jin Lee

Received April 25, 2019; revised August 7, 2019; accepted September 16, 2019; published January 31, 2020

Abstract

Cyber-attacks emerge in a more intelligent way, and various security technologies are applied to respond to such attacks. Still, more and more people agree that individual response to each intelligent infringement attack has a fundamental limit. Accordingly, the cyber threat intelligence analysis technology is drawing attention in analyzing the attacker group, interpreting the attack trend, and obtaining decision making information by collecting a large quantity of cyber-attack information and performing relation analysis. In this study, we proposed relation analysis factors and developed a system for establishing cyber threat intelligence, based on malicious code as a key means of cyber-attacks. As a result of collecting more than 36 million kinds of infringement information and conducting relation analysis, various implications that cannot be obtained by simple searches were derived. We expect actionable intelligence to be established in the true sense of the word if relation analysis logic is developed later.

Keywords: Cyber Threat Intelligence, Clustering, Indicator, Attack Information, Relationship, Cyber-attacker

A preliminary version of this paper appeared in IEEE ICC 2009, June 14-18, Dresden, Germany. This version includes a concrete analysis and supporting implementation results on MICAz sensor nodes. This research was supported by a research grant from the IT R&D program of MKE/IITA, the Korean government [2005-Y-001-04, Development of Next Generation Security Technology]. We express our thanks to Dr. Richard Berke who checked our manuscript.

http://doi.org/10.3837/tiis.2020.01.015 **ISSN : 1976-7277**

1. Introduction

Cyber-attacks are gradually becoming more intelligent. Spear phishing attacks increased 55 percent year-on-year in 2015, but the number of victims per attack mail dropped by 39 percent, indicating that attacks from specific groups of attackers were concentrated [23]. In addition, recently in Korea, ransomware has increased rapidly, and only 247 million ransomware were detected and blocked in the first half of 2016. 54 cases of 'Zero-Day Vulnerability' were found in 2015, which was a 125% increase from the year before [23]. Approximately one million malicious codes are appearing every day, and they are used for cyber-attacks. There is also a growing realization that a method of responding to those attacks at each individual point (network, endpoint) has a fundamental limit.

Thus, CTI (Cyber Threat Intelligence)[35, 36, 37, 38] technology comes into the spotlight as a technology that can analyze the meaning of the attack and support decision making about an object to respond by collecting a large amount of cyber infringement information and carrying out relation analysis. The reason is that malicious code, attack IP, and malicious code distribution domain can be secured, but the collection channel of the information cannot collect all factors making up a single cyber-attack.

Generally, only fragmented information can be collected, so the overall attack situation is not fully understood. That is, the exact cause of a cyber-attack is not clearly identified, so it cannot adequately respond to similar cyber-attacks that may occur later. It is impossible to analyze the overall attack aspect and the meaning of the overall trend instead of an individual infringement attack among a large amount of information collected from a different point of view.

The CTI technology answers the question "What should we do to cope with infringement attacks under the present conditions" by processing the information secured throughout the entire cyber-attack process (start, occurrence of damages, response) in refined form and conducting relation analysis. This paper proposes a method of establishing CTI with focus on malicious code.

The rest of this paper is organized as follows: Section 2 introduces related studies on intelligence analysis of the fragmented information; Section 3 proposes a method of identifying an infringement incident by collecting, managing, and mapping the fragmented information; Section 4 presents an API that can be provided through the implementation and result of semantic interpretation in line with data accumulation; The last Section presents the conclusion.

2. Related Work

Cyber-attacks and attack-groups are analyzed from various viewpoints. Many research studies were conducted on attacker profiling from the viewpoint of malware creation. Mohaisen A, et al classified the malware group through dynamic analysis based on the API behavior that occurs when executing malware and estimated the same attacker [1, 2, 24, 25, 26], whereas Kinable, et al studied the method of malicious code classification through static analysis based on the call graph of malicious code [3, 4, 27].

Regarding attacker profiling from the viewpoint of botnet, Gu, G., M. Feily, et al conducted a study on analyzing the attack resources possessed by the same attacker by detecting botnets and analyzing the command and control channel [5, 6]. H. Choi, P. Sroufe, et al performed

research on the detection of the botnet group infected by the same malicious code by analyzing the spam bot that sends spam e-mails [7, 8, 9].

Regarding profiling from the viewpoint of the cyber-attacker, Watters studied cyber-attacker models from the viewpoint of social and economic relation [10], whereas Kapetanakis performed research on case-based reasoning using characteristics that can identify the attacker such as technical standard, purpose, anti-forensic, and grammatical error [11].

Many studies are underway from the viewpoint of overall cyber-attack occurrence. Cho forecast an attacker group using the similarity characteristics of the domain names used for cyber-attacks [12]. Cova, Chen, Chang, et al detected and analyzed “drive-by-download” on the web as a representative means of spreading malicious code [13, 14, 15, 16].

In addition, Han found that the large-scale cyber terror attack in Korea and the cyber-attack against Sony Pictures were committed by the same attacker group through case-based reasoning [17, 28, 29].

Many activities are also conducted to share cyber-attack information in the standard aspect. STIX, a format for sharing cyber threat information, has been established [18], and TAXII communication protocol is available [19]. MAEC has been established to share malicious code information [20], whereas CVE, CVSS, CWE, and CWSS are utilized with regard to vulnerabilities.

In addition, global enterprises in the industry are trying to secure proprietary CTI technologies. Symantec released DeepSightTM Intelligence (CTI service for enterprises), which provides the reputational information such as malicious IP/Domain/Code and it’s behavior analysis data, behavior history and owner’s information. FireEye acquires iSIGHT partners to provide information on the attacker’s motivation, development environment, and analysis result of security issues. IBM provides the IP’s reputational information and vulnerability information using X-Force Threat Intelligence and other services. Besides those internal development activities, cooperation among global enterprises is also performed actively.

CTA (Cyber Threat Alliance), which is established by the initiative of Fortinet, shares cyber-attack information, conducts joint research, and published reports with the participation of Symantec and Intel [21, 30]. FireEye formed CSC (Cyber Security Coalition) with the participation of HP, IBM, and Splunk, developing complementary technologies between security companies and IT companies and integrating their products [22].

3. Proposed Model

3.1 Cyber-Attack Model

Collecting infringement information is a starting point of responding to cyber-attacks. Infringement information (attack resource) refers to each individual resource used for the cyber-attack and includes the time, IP, malicious code, and vulnerability.

A cyber-attack is a set of attack resources used for attacks start to finish. In other words, cyber-attack analysis involves mapping infringement resources and analyzing the combination effectively. Analysts then analyze attacks and analyze attackers' strategies and intentions.

Although there are several scenarios of cyber-attacks, this paper creates an attack model of infringement attacks using malicious code and presents an analysis process of relationships of infringement information. By initiating a cyber-attack, an attacker can destroy the core system,

disclose a lot of personal information, pursue monetary benefits using DDoS attacks and ransomware.

Attackers should avoid the defensive system of attacking targets for successful cyber attacks. Currently, most systems and PCs use security systems such as IPS / IDS and anti-virus to protect their systems. Therefore, attackers must disable these defense systems and then perform ongoing attacks. The most well-known attack method used for this is the "drive-by-download" attack.

The "drive-by-download" method is used on the Web, so even if there is no user's knowledge, the user's PC is infected with malware. An attacker can use the C&C server to control all operations of an infected PC when protecting infected PCs with malicious code. The attacker can obtain the configuration of the internal network inside the organization of the PC infected with malicious code, status of major systems, and connection information. The attacker then develops the attack plan, installs more malware, and initiates a major infringement attack like the destruction of the system[31].

The next section shows three cluster models based on these attack models from the three viewpoints of propagation, malware, and resources.

3.2 Attack Propagation-based clustering

According to the cyber-attack model defined above, this section provides profiling elements from the perspective of malware distribution.

First of all, an attacker could exploit this vulnerability, penetrate the website frequently visited by the user, and add the address of the selected link. Next, the website runs as originally designed when the user visits the website. However, code of related links that exploit vulnerabilities such as JavaScript, Flash, and Web browsers are executed even if it is invisible to the user, and malware is installed from the malicious code distribution site to the user's PC. The following **Fig. 1** shows an example of how an attacker can spread malicious code to a user through a Web:

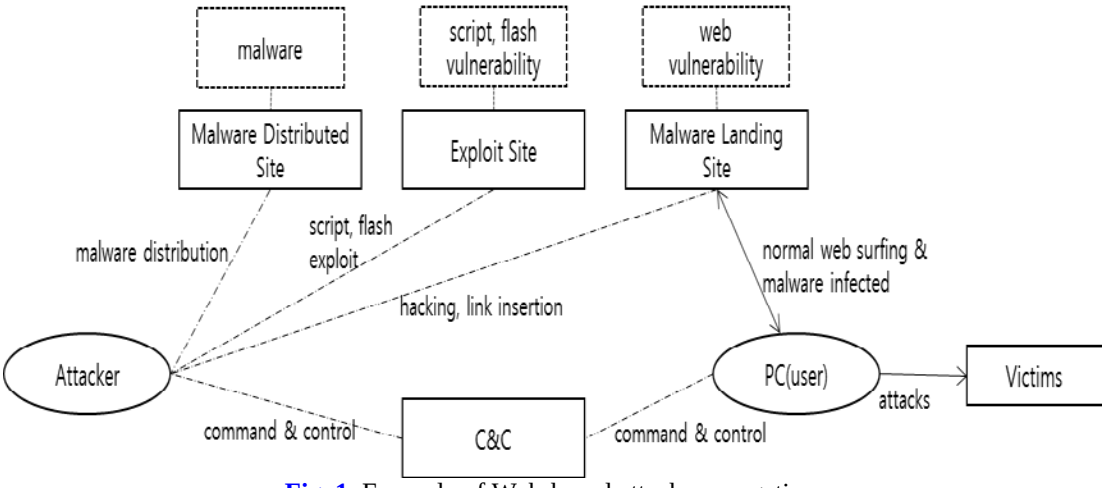

Fig. 1. Example of Web-based attack propagation

In order to propagate malicious code, it is necessary for an attacker to secure infringing resources (malicious code, distributed site, exploit site, exploit code, landing site, attack

vulnerability etc.) at the landing site beforehand[39]. Therefore, for clustering based on these attack radio waves, it is necessary to analyze infringing resources as follows:

A. Malware-based correlation

An attacker would generate malicious code for user infection. These malware are made with the characteristics of attackers. For example, there are compilation methods of malicious code, production environment, frequency of use of specific functions and functions name.

However, various malicious code creation tools have been disseminated, and with these characteristics, it is difficult to cluster attackers or attack groups. Also, most of the malware spread at the early stage of attack is used for the purpose of navigating in advance for a full-scale attack, and it is difficult to show the characteristics of these attacks.

B. Propagation-based correlation

An attacker secures a website with this vulnerability to propagate the generated malicious code. The secured sites are connected to each other by sharing the links related to each other.

Among the websites secured by attackers, websites with many people's connections are responsible for moving users to websites that distribute actual malware without directly distributing malware. Users who access these sites will unknowingly access the malicious code distribution sites through these website links and eventually become infected with malicious code.

This series of connections can be an important clue to identify the attacker. However, most sites related to the distribution of such malicious code are operated in a similar way, making it difficult to guess the exact attacker.

The exploiting code inserted into the exploiting site by the attacker is a key element that creates an environment of running malicious code in the user’s PC, becoming a key asset operated by the attacker. As the intelligent infringement attack exploits the Zero-Day vulnerability[32, 33, 34], existing security technologies cannot handle the attack properly.

3.3 Malware-based Clustering

Most cyber-attacks today use malicious code, and these malicious codes are critical to analyzing attackers. In the case of an attack using the same malicious code, there is a high probability that the attack is mostly caused by the same attacker or attack group. However, an attacker can be the same, even if it is an attack using other malware. It is important to check if the malicious code is the same or not, because some existing malicious code can be changed and reused. Therefore, we are actively studying the similarity of malicious code in various ways and estimating the same attacker based on this.

A. Static-based malware correlation

If you are analyzing malware statically, you can get a variety of evidence for attacker group detection. **Table 1** shows, if you analyze cyber-attacks on June 25th in Korea in 2013 and cyber-attacks on Sony Pictures in 2014, you can see that the names of system-destroying malware are very similar. According to the cyber-attack analysis, on June 25 attacks, malicious codes named taskhosts.exe, taskchg.exe and rdpshellex.exe were used.

In case of Sony Pictures attack as **Table 1**, taskhosts64.exe, taskchg16.exe and rdpshellex32. The malicious code named exe was used. The malicious codes described above have very

similar file names and perform very similar actions (system destruction). In addition, the source code similarity between malicious codes used in each attack is very high.

Table 1. Malicious code name comparison of cyber-attacks performed by the same attacker

Cyber-Attack	6.25 Cyber-Attack in Korea	Sony Pictures Cyber-Attack
Malicious Codes Name	taskhosts.exe taskchg.exe rdpshellex.exe	taskhosts64.exe taskchg16.exe rdpshellex32.exe

As a result of analyzing the attacks performed by the same attacker group among other cyber-attacks, it is confirmed that the strings (domain address, IP information, specific string, special characters, etc.) included in the malicious code are very similar although the file name and HASH code are different there is.

The various information obtained from the static analysis of the malicious code is an important clue to see if it is a cyber-attack caused by the same attacker. However, malicious codes that have recently been circulated or used for cyber-attacks have hidden the source code of malicious code in various ways to prevent such static analysis.

B. Dynamic-based malware correlation

As we have seen, malicious code static analysis is a very important factor in estimating the attacker of a cyber-attack. However, in order to avoid such a malicious code static analysis, we have disabled the static analysis using techniques such as obfuscation of malicious code and malicious code packing. In this case, it is possible to analyze the similarity of malicious code by collecting various information generated by actual execution of malicious code.

Based on this, it is possible to estimate and identify attacker of cyber-attack. Execute individual malicious code and collect and analyze the API sequence to be called at this time. After repeating the same process for the malicious codes to be analyzed, it is possible to determine the similarity of the API sequence of each malicious code and confirm that it is the malicious code generated by the same attacker. The following **Fig. 2** shows the similarity of information generated when five variants of malicious code are executed.

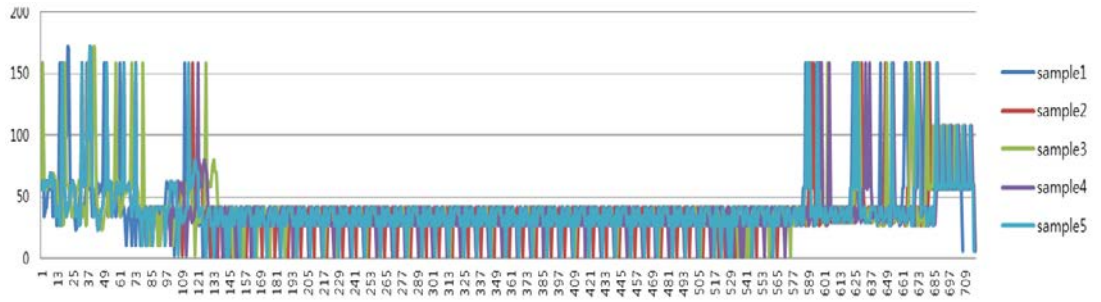

Fig. 2. Example of behavior-based malware mutant detection

In addition, the service name used by malicious code can be used as data useful for identifying the same attacker. 6.25 cyber-attack in Korea and Sony Pictures cyber-attack have been identified as using RasSecrulty and RasMgrp, which is a clue to the similarity between the two cyber-attacks and the attacks from the same attacker group.

C. Network-based malware correlation

When malicious code is executed, the malicious code can gain access to specific websites for downloading additional malicious code from an attacker or receiving an attacker's command. These sites are called 'Command and Control (C&C) Server', which refers to web servers or websites created by certain attackers to perform their cyber-attacks.

If different malicious codes connect to the same C&C server, it is very likely that they are malicious code created by the same attacker. Also, even if the C&C server's IP address is different, if it is a C&C server that uses the same C-Class band, this is also very likely to be malicious code created by the same attacker.

Also, if you check the communication history between the C&C server and the malicious code and use a specific communication protocol of a specific attacker or send data in a similar format, you can confirm that it is malicious code generated by the same attacker regardless of the IP address of the C&C server. When analyzing the communication history between these C&C servers and malicious code, we can classify the cyber-attacks that originate from the same attacker group.

3.4 Attacking Resource-based Clustering

Most cyber-attacks today use malicious code, and these malicious codes are critical to analyzing attackers. In the case of an attack using the same malicious code, there is a high probability that the attack is mostly caused by the same attacker or attack group. However, an attacker can be the same, even if it is an attack using other malware. It is important to check if the malicious code is the same or not, because some existing malicious code can be changed and reused. Therefore, we are actively studying the similarity of malicious code in various ways and estimating the same attacker based on this.

A. Attack IP-based correlation

We can collect cyber-attack IP's that are generally identified as being used for cyber-attacks through the 'Open Source Intelligence (OSINT)' site. In particular, 'Real-time Black List (RBL)' sites provide various types of IP information related to cyber-attacks.

However, since the attacked IP information collected only informs the fragmented IP address, it cannot confirm what role the IP performs in the cyber-attack phase (distribution of malicious code, C&C server, information retrieval server, etc.).

But, if we collect these fragmentary IP information and additional information such as IP's C-Class band, owner, geographical, connected domain, malicious code distributed by that IP, C&C usage, etc. we can identify the attacker of attack cyber-attack based on IP attack. These various information can be obtained through OSINT sites mentioned above. In addition, by analyzing WHOIS service, Geo-Location information, and owner's e-mail information with attacking IP, you can get more meaningful results[39].

B. Attack domain-based correlation

Attackers are using a variety of methods to spread malicious code in cyber-attacks. The most common method is to spread malicious code through the website. To do this, an attacker creates a website that can spread malicious code and uses a domain address to facilitate access.

The domain address is created by using the domain address that modified the website address frequently accessed by the victim, or the address of the online shopping mall address or the field of interest.

In addition, attackers use a detection bypass method such as shortened-URL to bypass domain address based malware distribution detected technology. Most of them use similar domain addresses in the same attack group for cyber-attacks or reuse the same domain address.

Therefore, when analyzing domain addresses, it is possible to identify the background of common cyber-attacks. However, in case of advanced or hidden cyber-attack, it is hard to identify attacker using simple domain address because domain address is used by randomization.

To solve this problem, various information related to the domain should be collected. The additional information collected can be used to identify the same attacker. The information used in this case can be the domain owner, IP connected to the domain, IP used for existing cyber-attacks, history of IP changing, malicious code distributed in domain similar to the past domain, TLD/SLD similar information. Using this information, it is possible to identify similar cyber-attacks with the same attacker group even when the actual domain address is different.

C. Cyber-attack indicator similarity-based correlation

Finally, attackers can be identified based on the similarity between various indicators used in cyber-attacks. This method is a comprehensive analysis of the contents of the previous sections 3.2 and 3.3. All the resources used in cyber-attacks are correlated, and their inherent attribute values are also related to each resource. These associations can be used to identify attackers and to classify cyber-attacks that they have generated. For example, through attack IP, it is possible to collect malicious codes and distributed domain information, but there are difficulties in identifying clear attackers.

Therefore, if the unique attributes of the attacked IP are collected and connected to each other, the association with other existing cyber-attacks can be found. Information that can be collected based on attack IP is IP address, owner and owner email, other connected domain, history of C&C server usage, similar C-Class IP, similar malicious code analysis information, malicious code key string, etc.

Attackers are preparing various resources for cyber-attacks, but they must use specific information to acquire or use these resources. Therefore, when collecting such specific information and analyzing other indicators associated with it, the same attacker can be identified. In addition, although attackers mostly use fake information to acquire or use resources, they can also be the main identifiable elements of an attacker.

However, this analysis can be done by combining a large amount of cyber-attack information and attribute information of the cyber-attack information. For this, big-data processing and graph-based data analysis techniques should be used[40]. In Section 4, we discuss the results of the system implemented using the various clustering techniques described above.

4. Experimental Result

4.1 System Overview

Previously, we have proposed clustering factors from the viewpoint of cyber-attack spread, malicious code, and resources used for the attack to analyze cyber threat intelligence from the

viewpoint of malicious code. In this section, we designed and developed a system to collect a large quantity of cyber infringement resources and cluster those resources based on the relation analysis factors proposed in advance. The following figure shows the system configuration to analyze cyber threat intelligence from the viewpoint of cyber-attack indicator similarity-based correlation:

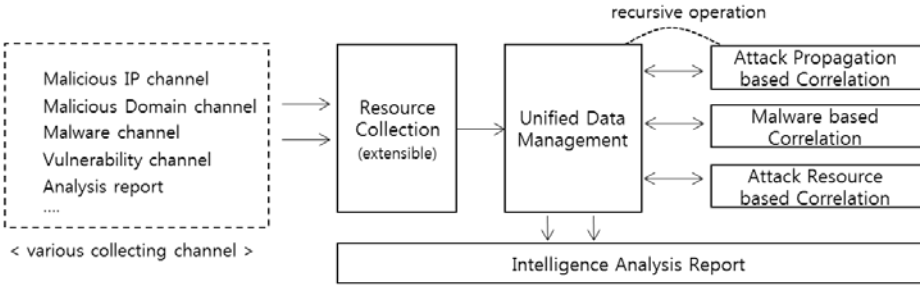

Fig. 3. Suggested system overview

The infringement information can be collected from various sources like as **Fig. 3**. Those sources can be RBL site opened to the public, malicious code sharing site, or various kinds of infringement information internally collected by organization/company.

Since the information is composed of fragmented information, not all information composing the infringement incident, however, a data processing process that reconfigures and manages the information collected by each channel in an integrated manner based on a single standard (e.g., IP, domain, malicious code) is required.

Through this process, the collection channel can be reflected with sufficient scalability when added regardless of the type. The information collected in this manner is reconfigured through three relation analyses proposed previously and is used to detect an infringement attack that seems to originate from the same attacker group. The following **Table 2** shows examples of 'Cyber Threat Intelligence (CTI)' information produced based on similarity of cyber-attack indicators:

Table 2. The cyber-attack intelligence analysis item

No.	Contents	Expected effect
1	Reputation by infringement resource and history information used for the infringement attack	Analysis based on the level of risk, operation of infringement attack blocking policies
2	Malicious code distributed on the same path and connected by the same C&C	Providing correlation among various malicious codes and clues to trace an attacker
3	Infringement incident information having the same C-class band and similar distribution domain name	Correlation among irrelevant infringement attacks can be analyzed
4	Information of malicious code with the same detailed distribution path and infringement incident	Providing a clue to trace the same attacker based on the characteristics of malicious code distribution
5	Infringement incident information wherein the same exploiting code and vulnerability were used	Providing a clue to trace the same attacker who used the same vulnerability

6	Infringement incident information that has the same malicious code installation path on the device	Providing a clue to trace the same attacker based on the characteristics of malicious code behavior
7	Infringement incident information that has the same malicious code file name, compilation time, and debugging path	Providing a clue to trace the same attacker based on the characteristics of malicious code behavior
8	Domain and IP mapping history and information of the infringement incident committed by the same owner	Correlation among irrelevant infringement attacks can be analyzed
9	Information of the malicious code variant that seems to have been created by the same attacker, by analyzing malicious code statically and dynamically	Providing information of the malicious code created and distributed by the same attacker by analyzing variants

4.2 Cyber-attack Indicator based Security Intelligence Analysis

A. Malicious IP-based relationship analysis

We perform association analysis with each cyber-attack using IP information among various information collected from the system. In this case, the information to be compared is the unique attribute information and the history information of each indicator.

Firstly, it is checked whether the same data exists, similar C-Class IP information exists, and owner information is the same or similar. In addition, we analyze the relationship between domain's string and domain owner's information in IP connected domain. Then, analyzed whether the TLD / SLD information of the analyzed domain is similar. Then, we analyze whether the malicious code is distributed in the IP to be analyzed or whether there is an attack command communications. Based on this analysis information, we graphically manage the association of information related to malicious IP. **Fig. 4** and **Table 3** show the elements and criteria for analyzing the association between IP-based collected indicators.

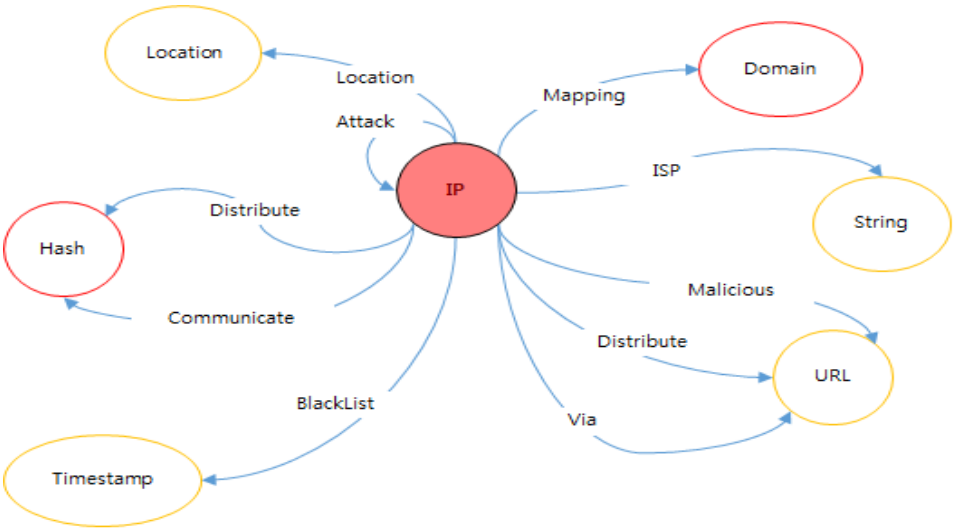

Fig. 4. Concept of malicious IP-based relationship analysis

Table 3. Relationship analysis between IP and indicators

Start	Relation	Properties	Description	Indicator
IP	Attack	Risk, time	Attacker IP and victim IP mapping	IP
	BlackList	channel	BlackList IP detected time	Timestamp
	C&C	time	When IP performs the C&C role, it spreads in URL form	Url
	Communicate		IP-connected C&C communication malware	Hash
	Distribute	Time	Malicious code distributed in IP	Hash
	Distribute	time	When the IP performs the dissemination area, it spreads in the form of URL	Url
	ISP		ISP (Internet Service Provider) information provided by IP	String
	Location	type	Country / Region information for IP	Location
	Malicious	Time, description	Malicious URLs used by IP	Url
	Mapping	Time	IP Reverse Lookup Result	Domain
	Mapping	Time	Hostname used by IP	Domain
	Mapping	time	IP <-> Domain mapping by PRT result	Domain
	Via	time	Mapping malware connecting IP with URL	Url

B. Malicious Domain-based relationship analysis

The association between each cyber-attack can be analyzed by using the domain value used in the cyber-attack, the attribute value possessed by the domain, and the connected information. The information, the string, and the location information used to register the domain are first grasped and the correlation is confirmed. After that, we analyze the association with each cyber-attack by using the history of the malicious code distribution, the IP information of the connected IP, the history information of using the domain as the C & C server, and the like **Fig. 5** and **Table 4** show this malicious domain-based association analysis.

Table 4. Relationship analysis between domain and indicators

Start	Relation	Properties	Description	Indicator
Domain	Admin		Domain administrator name	String
	Admin		Domain administrator email	Email
	Authorized agency		Domain registration agency	String
	BlackList	channel	Time Detected by Black-List Domain	Time-stamp
	C&C	Time	When the Domain performs the C&C role, it spreads in URL format	Url
	C&C		C&C communication malicious code connected to domain	Hash

	Distribute	time	When the domain performs the distribution area, it spreads the URL form	Url
	Distribute		Malicious code distributed by Domain	Hash
	Location	type	Domain Country / Region Information	Location
	Malicious	time	Malicious URLs used by Domain	Url
	Mapping	Time	Domain mapping with IP	IP
	Mapping	Time	Mapping between Domains and malware connecting IP	IP
	Mapping	Time	Hostname that IP used	IP
	New Domain	Type(generate)	Date the domain was originally created	
	Registrant		Domain registrant name	String
	Registrant		Domain registrar email	Email
	Update Domain	Type(update)	Date the domain was last modified	Timestamp
	Via	Time	Mapping a malware connected domain with a URL	Url

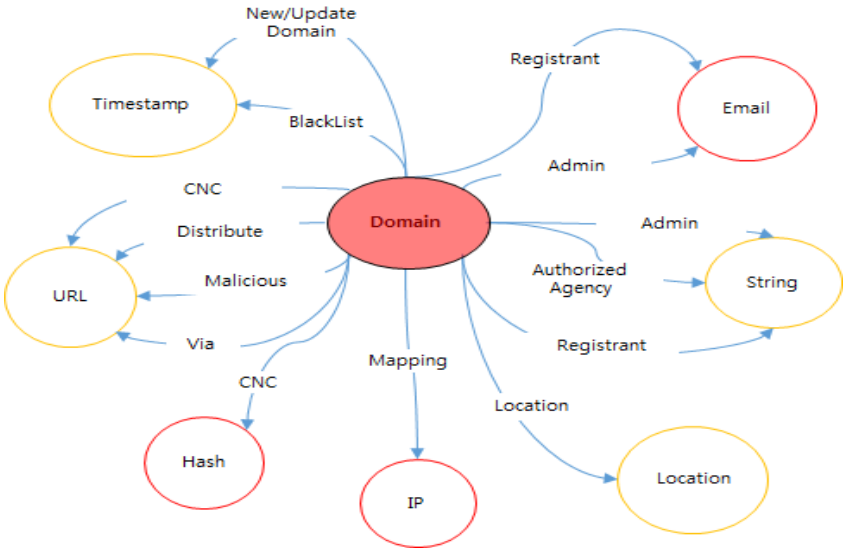

Fig. 5. Concept of malicious domain-based relationship analysis

C. Malicious Code-based relationship analysis

Basically, we analyze the relationship and similarity with the malicious code used in the existing infringement by using static analysis and behavior analysis result of malicious code. After that, C&C server information accessed by malicious code and IP and domain information which distributed malicious code, string and file name of malicious code are compared and analyzed. In addition, it analyzes IP and domain proprietary property information and malicious history information to find another malicious code, IP, domain, etc like as **Fig. 6** and **Table 5**. By using the detected information, it is possible to detect other cyber-attacks related to the cyber-attack to which the analyzed malicious code belongs.

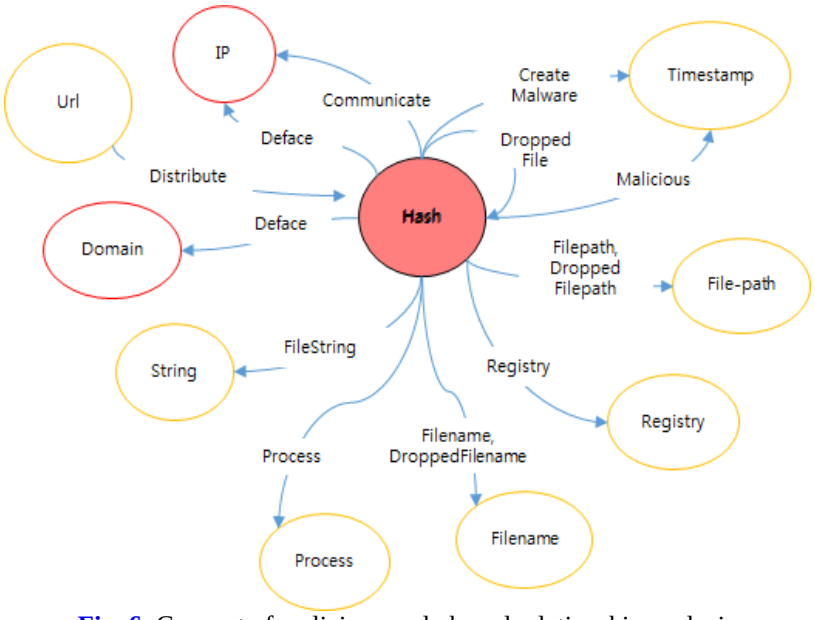

Fig. 6. Concept of malicious code-based relationship analysis

Table 5. Relationship analysis between malicious code and indicators

Start	Relation	Properties	Description	Indicator
Hash	Communicate		Network communication	IP
	Create Malware		Malicious code generation time	Timestamp
	Deface		Malicious Code Modified IP	IP
	Deface		Domains with malicious code	Domain
	Dropped File		Files generated by malicious code	Hash
	Dropped Filename		File name generated by malicious code	Filename
	Dropped File-path		File-path generated by malicious code	File-path
	Filename	time	Malicious file name	Filename
	File-path		Malicious file-path	File-path
	File String		Malicious code internal string	String
	Malicious		Malicious code first occurrence time	Timestamp
	Process		Processes generated by malicious code	Process
	Registry		Registry accessed from malicious code	Registry
URL	Distribute		Distribution URL and malware	Hash
	Distribute		Malicious IP and malware	Hash
	Distribute		Distribution URL	Hash

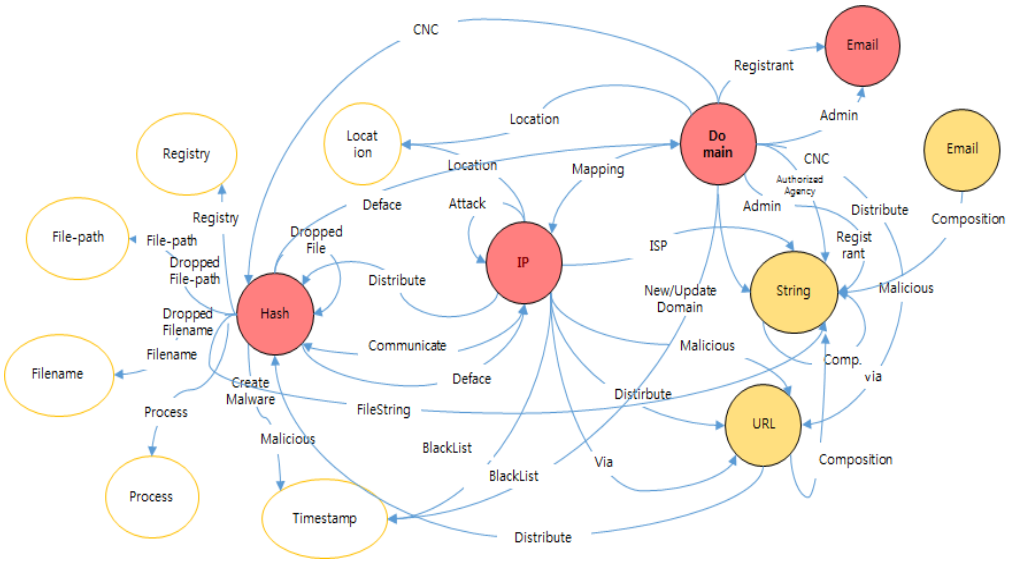

Fig. 7. Total concept of malicious indicator-based relationship analysis

In section 4.2, we investigated the correlation between the indicators used in the cyber-attacks to find the attacks performed by the same or similar attack group among various kinds of cyber-attacks. These associative analyzes are finally constructed as shown in **Fig. 7**.

In section 4.3, we examine the detection of new cyber-attacks, which are caused by the attack group that performed the existing cyber-attacks, through the indicator association analysis through the methods introduced in Sections 4.1 and 4.2.

4.3 Example of Security Intelligence Analysis

Data related to cyber-attacks directly or indirectly were collected for 19 months (October 2015 ~ June 2017) using the developed system, in order to analyze security intelligence. A total number of cyber-attack indicator is 36,743,069 which is 1,073,880 malicious codes, 35,305,058 IPs, and 364,131 malicious domains names information were secured during the specified period, using the system. The secured data was compared with Table 2. Intelligence Analysis Item in Section 4.1 using Section 4.2, and correlation among cyber-attacks was checked, which seemed to be irrelevant superficially.

A. Selecting an analysis target for cyber intelligence analysis

One cyber-attack using ransomware in 2016 and one personal information leak using general malicious code were selected, and two attacks were launched at different times. **Table 6** shows the representative malicious code used for each attack, and **Table 7** and **Table 8** presents the information of each malicious code.

One malicious code used for each attack was selected and analyzed based on **Table 2**. Intelligence Analysis Item in Section 4.1.

Table 6. Analysis target malware information (Depth 0)

Cyber-attack	Malware HASH
Ransomware	9F926B4A0707954EE72631EBC25CA53DE302991A1A…
Information Leak	B789F20A9EA8E28BD3664C9EC2A51CA69A6B12FF16…

Table 7. Ransomware-related information (Depth 1)

Relation Info.	IP	Domain	Attribute
Distribution Domain	192.185.XXX.152	oriinXXXXX.com	
Malware Name			73.exe
			moidh-a.exe

Table 8. Information leak malware-related information (Depth 1)

Relation Info.	IP	Domain	Attribute
Distribution Domain		ieupdate.XXXXX.com	
Malware Name			bundle_ytd_8006.exe
			tbedrs.dll
			tbwal1.dll

B. Intelligence analysis using the correlated information

To check the correlation of two cyber-attacks, the attack information derived from the information of the resources used for those two attacks was examined. Relation among attack information was identified using **Table 2**. in Section 4.1. The information related to **Table 7** and **Table 8** was extracted from the information collected/accumulated in the system, and the result was defined as “depth 2 result,” The extracted depth 2 result was also used to extract another correlated information called “depth 3 result,”

The correlation with “depth 0 (representative malicious code)” was examined as the first analysis target by extracting the derived correlated information in this way, and the point of intersection between two different cyber-attacks was analyzed. **Fig. 8** and **Fig. 9** show the result of intelligence analysis using the correlation analysis standard between **Table 2** in Section 4.1 and cyber-attacks, and **Table 9** presents the information of the detected cyber-attack.

Table 9. Intelligence analysis result of 2-different cyber-attacks

Detection Info.	Domain	IP	Attribute
Malware HASH			B2A43286FF98D5435…
Distribution Domain & Mapping IP	big.p1.XXXXXX.com	114.108.XXX.32	
	oriiXXXX.com	192.185. XX.152	
	myoriXXXX.com	192.185. XX.152	
	goldiXXXX.com	143.95. XX.110	
Distribution IP		192.185. XX.152	
		239.255. XX.250	
		143.95. XX.110	
C&C Server IP		54.254. XX.171	
		82.145. XX 5.39	
		91.203. XX.18	
Malware Name			bundle_YTD_8006.exe
			Moidh-a.exe

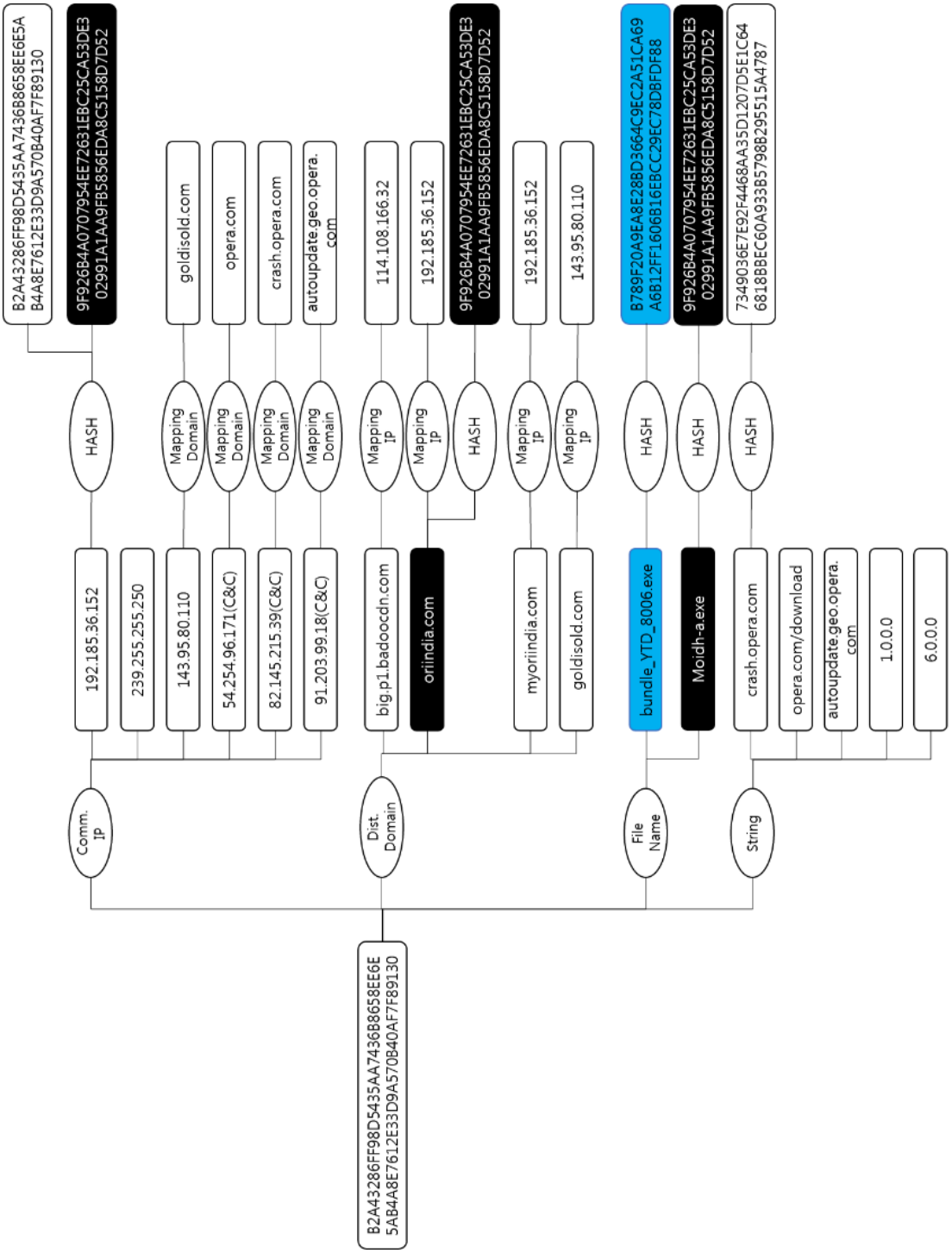

Fig. 8. Two Cyber-attacks relation map, based on the detected new cyber-attack

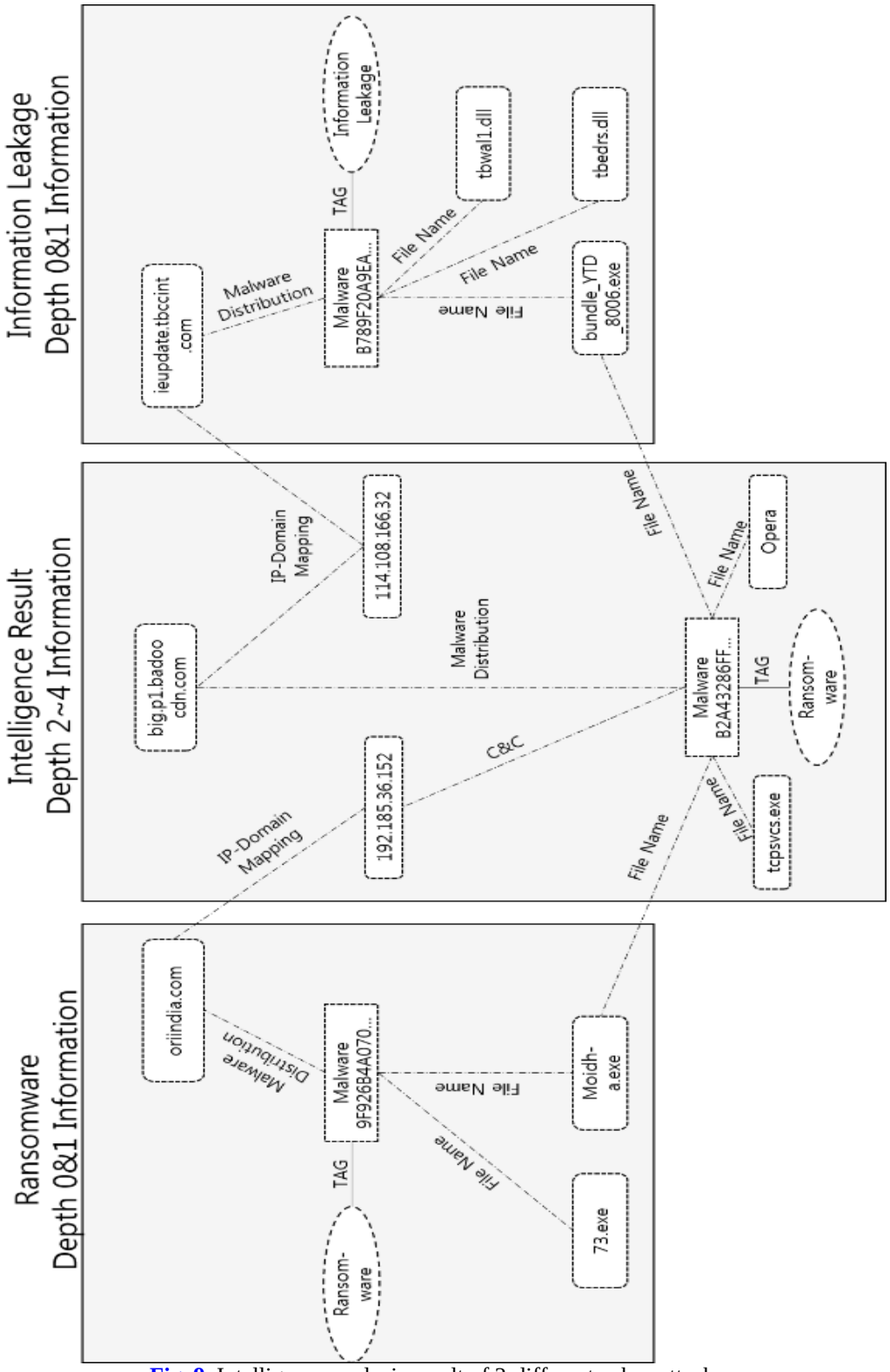

Fig. 9. Intelligence analysis result of 2-different cyber-attacks

The cyber-attack detected by intelligence analysis was found to be another ransomware attack that has occurred previously. Part of the malicious code file name used for that ransomware attack was found to be the same as that of the cyber-attack that has been analyzed, and the malicious code distribution IP and C&C were shared. It was also found that one IP was used for the domain, which has distributed numerous malicious codes, and utilized as a C&C server. It is important to identify similar or identical information in a large quantity of data and analyze correlation in the data quickly.

5. Conclusion

Cyber-attacks are becoming more intelligent. Although the response technology has also made significant advancements to keep up with those attacks, more and more people agree that the individual response method has a fundamental limit. As such, the Cyber Threat Intelligence technology comes into the spotlight as a technology that can collect a large amount of infringement information and support decision making by performing relation analysis on such information. Most global security companies are developing their proprietary CTI technologies and solving the problem through collaboration such as information sharing among various organizations and companies. The CTI technology aims to support decision making about the task that should be performed under the present conditions. This study presented core elements to develop a CTI technology based on malicious code -- which is the cause of most infringement incidents -- among various CTI components, from the viewpoint of cyber-attack propagation, malicious code, and resource. In addition, the system was developed in an environment with multiple collection channels, and the intelligence analysis result was obtained. As a result, we could estimate the cyber-attack launched by each attacker group and list of infringement resources used or possessed by those groups and check the possibility of analyzing the activity details and attack trends and characteristics of those groups. The purpose of intelligence analysis is to support decision making on “What should we do under the present conditions,” which seems to entail lots of work in the future. Currently, the system is designed for the expert analyst to understand the meaning of an attack. Still, the analysis performed by expert analysts is expected to be automated gradually. In addition, the profiling method is now focusing on known elements based on attack characteristics. It seems that the method should be developed such that unknown attack patterns are automatically detected and implications are suggested to the analyst by integrating the machine learning technology later.

Acknowledgement

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2017-0-00158, Development of Cyber Threat Intelligence(CTI) analysis and information sharing technology for national cyber incident response) and supported by 2019 Research Grant from Kangwon National University

References

[1] Mohaisen A, Alrawi O, "Unveiling zeus: automated classification of malware samples," in *Proc. of 22nd international conference on world wide web companion*, pp. 829-832, 2013. Article (CrossRef Link)

[2] Lee, Taejin, and Jin Kwak, "Effective and Reliable Malware Group Classification for a Massive Malware Environment," *International Journal of Distributed Sensor Networks*, vol.12, no.5, 2016. Article (CrossRef Link)

[3] Kinable, Joris, and Orestis Kostakis, "Malware classification based on call graph clustering," *Journal in computer virology*, 7(4), 233-245, 2011. Article (CrossRef Link)

[4] Hu, Xin, Tzi-cker Chiueh, and Kang G. Shin, "Large-scale malware indexing using function-call graphs," in *Proc. of the 16th ACM conference on Computer and communications security. ACM*, pp. 611-620, 2009. Article (CrossRef Link)

[5] Guofei Gu, Junjie Zhang, and Wenke Lee, "Botsniffer: Detecting botnet command and control channels in network traffic" in *Proc. of the 15th Annual Network and distributed System Security Symposium*, 2008 Article (CrossRef Link)

[6] M. Feily, A. Shahrestani, and S. Ramadass, "A survey of botnet and botnet detection," in *Proc. of the 3rd International Conference on Emerging Security Information, Systems, and Technologies (SECURWARE '09), IEEE, Glyfada, Athens*, pp. 268–273, June 2009. Article (CrossRef Link)

[7] H. Choi, H. Lee, H. Lee, and H. Kim, "Botnet detection by monitoring group activities in DNS traffic," in *Proc. of the IEEE International Conference Computer and Information Technology (CIT '07)*, 2007. Article (CrossRef Link)

[8] P. Sroufe, S. Phithakkitnukoon, R. Dantu, and J. Cangussu, "Email shape analysis for spambotnet detection," in *Proc. of the 6th IEEE Consumer Communications and Networking Conference (CCNC '09)*, pp. 1–2, January 2009. Article (CrossRef Link)

[9] Lee, Taejin, et al, "Detection of malware propagation in sensor Node and botnet group clustering based on e-mail spam analysis," *International Journal of Distributed Sensor Networks*, vol.11, no.9, 2015. Article (CrossRef Link)

[10] Watters, Paul A., et al, "Characterising and predicting cyber-attacks using the Cyber-attacker Model Profile (CAMP)," *Journal of Money Laundering Control*, 15(4), 430-441, 2012. Article (CrossRef Link)

[11] Kapetanakis, Stelios, et al, "Profiling cyber-attackers using Case-based Reasoning," in *Proc. of Part of AI-2014 Thirty-fourth SGAI International Conference on Artificial Intelligence, Cambridge*, 2014. Article (CrossRef Link)

[12] Cho, Hyeisun, et al, "The study of prediction of same attack group by comparing similarity of domain," in *Proc. of Information and Communication Technology Convergence (ICTC), 2015 International Conference on. IEEE*, 2015. Article (CrossRef Link)

[13] Cova, Marco, Christopher Kruegel, and Giovanni Vigna, "Detection and analysis of drive-by-download attacks and malicious JavaScript code," in *Proc. of the 19th international conference on World wide web. ACM*, pp. 281-290, 2010. Article (CrossRef Link)

[14] Chen, Kevin Zhijie, et al, "WebPatrol: Automated collection and replay of web-based malware scenarios," in *Proc. of the 6th ACM Symposium on Information, Computer and Communications Security. ACM*, pp. 186-195, 2011. Article (CrossRef Link)

[15] Wang, Gang, et al, "Detecting malicious landing pages in Malware Distribution Networks," in *Proc. of 2013 43rd Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN). IEEE*, 2013. Article (CrossRef Link)

[16] Chang, Jian, et al, "Analyzing and defending against web-based malware," *ACM Computing Surveys (CSUR)*, 45(4), Article No.49, 2013. Article (CrossRef Link)

[17] Mee Lan Han, Hee Chan Han, Ah Reum Kang, Byung Il Kwak, Aziz Mohaisen and Huy Kang Kim, "WHAP: Web-Hacking Profiling Using Case-Based Reasoning," in *Proc. of 2016 IEEE Conference on Communications and Network Security*, 2016. Article (CrossRef Link)

[18] Barnum, Sean, "Standardizing Cyber Threat Intelligence Information with the Structured Threat Information eXpression (STIX™)," *MITRE Corporation*, 2014. Article (CrossRef Link)

[19] Julie Connolly, Mark Davidson, Matt Richard, Clem Skorupka, “The Trusted Automated eXchange of Indicator Information (TAXIITM),” November 2012.
[20] Kirillov, Ivan, et al, “Malware attribute enumeration and characterization,” *The MITRE Corporation, Tech. Rep*, 2010.
[21] CTA(Cyber Threat Alliance), Article (CrossRef Link)
[22] CSC(Cyber Security Coalition), Article (CrossRef Link)
[23] Symantec, “Internet Security Threat Report,” vol. 21, 2016. Article (CrossRef Link)
[24] Fariba Haddadi and A. Nur Zincir-Heywood, “Botnet behaviour analysis: How would a data analytics-based system with minimum a priori information perform?,” *International Journal of Network Management*, Vol. 27, Issue 4, 2017. Article (CrossRef Link)
[25] Gamal A. N. Mohamed and Norafida Bte Ithnin, “SBRT: API Signature Behaviour Based Representation Technique for Improving Metamorphic Malware Detection,” in *Proc. of International Conference of Reliable Information and Communication Technology 2017: Recent Trends in Information and Communication Technology*, pp.767-777, 2017.
Article (CrossRef Link)
[26] W Han, J Xue, Y Wang, L Huang, Z Kong and L Mao, “MalDAE: Detecting and explaining malware based on correlation and fusion of static and dynamic characteristics," *Computers & Security*, vol. 83, pp.208-233, 2019. Article (CrossRef Link)
[27] J Stiborek, T Pevný and M Rehák "Multiple instance learning for malware classification,” *Expert Systems with Applications*, Vol. 93, pp.346-357, 2018. Article (CrossRef Link)
[28] YM Krakovsky, AN Luzgin and EA Mikhailova, “Interval forecasting of cyberattack intensity on informatization objects of industry using probability cluster model,” *Journal of Physics: Conference Series, Mathematical simulation and data processing*, Vol. 1015, 2018.
Article (CrossRef Link)
[29] YM Krakovsky, AN Luzgin and YM Ivanyo, “Cyberattack intensity forecasting on informatization objects of critical infrastructures,” *Materials Science and Engineering*, Vol. 481, Number 1, 2019. Article (CrossRef Link)
[30] Sahrom Abu, Siti Rahayu Selamat, Aswami Ariffin and Robiah Yusof, “Cyber Threat Intelligence – Issue and Challenges,” *Indonesian Journal of Electrical Engineering and Computer Science*, Vol. 10, Number 1, 1. pp. 371-379, 2018.
[31] Gireesh Joshi, R.Padmavathy, Anil Pinapati and Mani Bhushan Kumar, “BrowserGuard2: A Solution for Drive-by-Download Attacks,” in *Proc. of the Second International Conference on Microelectronics, Computing & Communication Systems (MCCS 2017)*, pp. 739-750, 2017. Article (CrossRef Link)
[32] Ziwei Ye, Yuanbo Guo and Ankang Ju, “Zero-Day Vulnerability Risk Assessment and Attack Path Analysis Using Security Metric,” in *Proc. of International Conference on Artificial Intelligence and Security(ICAIS 2019), Artificial Intelligence and Security*, pp. 266-278, 2019.
Article (CrossRef Link)
[33] Ioannis Stellios, Panayiotis Kotzanikolaou and Mihalis Psarakis, “Advanced Persistent Threats and Zero-Day Exploits in Industrial Internet of Things,” *Security and Privacy Trends in the Industrial Internet of Things*, pp. 47-68, 2019. Article (CrossRef Link)
[34] Umesh Kumar Singh, Chanchala Joshi and Dimitris Kanellopoulos, “A framework for zero-day vulnerabilities detection and prioritization,” *Journal of Information Security and Applications*, Vol. 46, pp.164-172, 2019. Article (CrossRef Link)
[35] Ghaith Husari, Ehab Al-Shaer, Bill Chu and Ruhani Faiheem Rahman, “Learning APT chains from cyber threat intelligence,” in *Proc. of the 6th Annual Symposium on Hot Topics in the Science of Security*, pp. 1-2, 2019. Article (CrossRef Link)
[36] Ali Dehghantanha, Mauro Conti and Tooska Dargahi, *Cyber Threat Intelligence*, Springer, Cham, 2018. Article (CrossRef Link)
[37] J Surma, “Cyber Threat Intelligence Systems: problems and challenges,” *Collegium of Economic Analysis Annals, Warsaw School of Economics, Collegium of Economic Analysis*, issue 54, pp. 267-274, 2019. Article (CrossRef Link)

[38] Mauro Conti, Tooska Dargahi and Ali Dehghantanha, "Cyber Threat Intelligence: Challenges and Opportunities," *Advances in Information Security(ADIS)*, Vol. 70, pp. 1-6, 2018. Article (CrossRef Link)
[39] Seulgi Lee, Hyeisun Cho, Nakhyun Kim, Byung-ik Kim, and Jun-hyung Park, "Detection of Similarities in Cyber Threats through OSINT," *International Journal of Innovative Research in Technology and Science*, Vol. 5, Issue 6, pp. 20-25, Nov 2017. Article (CrossRef Link)
[40] Byung-ik Kim, Seulgi Lee, Hyeisun Cho, Nakhyun Kim, and Jun-hyung Park, "Study of Potential Cyber Threat Detection Technology using Big Data and Graph Analysis," *Engineering, IT and Artificial Intelligence*, 2018. Article (CrossRef Link)

Kyung-ho Son received his B.S. degree in received his B.E., M.S., and Ph.D. degree from Sungkyunkwan University in 2001, 2013, and 2015, respectively. He worked at Korea Internet Security Agency from 2001 to 2018 and he has been worked in Kangwon National University since 2018. His research area information assurance, Privacy by Design, Design of Security system, IoT·CPS Security

Byung-ik Kim received the B.S. degree in Computer Science from the University of Ajou, Korea, in 2010. Currently, He is a Deputy General Researcher of Security Threat Response R&D Team at Korea Internet & Security Agency. His research areas include cyber threat analysis, cyber at-tack related data correlation, and sensor.

Tae-jin Lee graduated from Postech Computer Engineering Department in 2003 and graduated from Yonsei University in 2008 and Ajou University in 2017. He worked at Korea Internet Security Agency from 2003 to 2017 and he has been worked in hoseo university since 2017. His research area are artificial intelligence, malware and intrusion detection.