Advanced Science and Technology Letters
Vol.29 (SecTech 2013), pp.30-36
http://dx.doi.org/10.14257/astl.2013.29.06

# Detection of Advanced Persistent Threat by Analyzing the Big Data Log

Jisang Kim[1], Taejin Lee, Hyung-guen Kim, Haeryong Park

KISA, Information Security Group, IT Venture Tower, Jungdaero 135, Songpa Seoul, Korea
jisang@kisa.or.kr, tjlee@kisa.or.kr, nomota@mobigen.com, hrpark@email.kisa.or.kr

**Abstract.** This paper proposes and verifies the algorithm to detect the advanced persistent threat early through real-time network monitoring and combinatorial analysis of big data log. Moreover, provide result tested through the analysis in the actual networks of the deduced algorithms

**Keyword:** Advanced Persistent, APT, big data

## 1 Introduction

This paper proposes the APT detection algorithm, which analyzes the existing incidents in depth, establishes the main factors deduced from the analyzed incidents and detection model to monitor penetration through data leak, and then combines the deduced detection factors and established model. The proposed algorithm is then applied to the actual network to verify its effectiveness.

## 2 Proposed Methodology for Analysis

### 2.1 Deducement of Main Factors through Case Analysis

The analysis of the attack against SK Communications in July 2011, Korea, indicates that the attacker(s) (1) registered the domain and prepared the control (C&C) server for a period of months, (2) borrowed the well-known virus vaccine update method to penetrate the network, (3) installed various remote admin tools (RAT) in the infected PCs, (4) periodically established the connection of the infected PCs to the control (C&C) server, and (5) queried the external DNS instead of internal ones [Com5].

As implications, (1) the organizations should manage their DNS in-house and control the direct query of external DNS, (2) periodic monitoring of communication to the control (C&C) server will help find the infected PCs, (3) transmission of unencrypted text message through the SSL encrypted port should be monitored, and (4) already known call-back domains should be periodically collected and blocked.

1 This work was supported by the IT R&D program of MOTIE/KEIT. [10044938, The Development of Cyber Attacks Detection Technology based on Mass Security Events Analysing and Malicious Code Profiling]

ISSN: 2287-1233 ASTL

### 2.2 Mutual Cross Analysis of Big Data Log

According to many studies, the acquisition, integration, and cross analysis of different logs will increase the accuracy of penetration detection and reduce unnecessary alarms. [Lee S. H.] [Lee H.W. ] Although this paper also seeks to combine the multiple logs to detect the APT, the observation window should be longer than the existing studies to block attacks continuing for a long period.

### 2.3 Consideration for Monitoring for Early Detection

To apply the detection model in an actual network, the following items must be monitored:
Network monitoring assumes that the entire network used by a specific organization can be observed at some bottleneck points. It can be generally executed through packet mirroring at the backbone switch used by the organization; the outbound packet in particular should be observable. Since the Web traffic has traffic that is distributed widely, it has a relatively large analysis target. If normal Web traffic can be induced with a Web proxy (automatically configured using the security S/W used by the organization), the policy should encourage its use. After it is processed, traffic other than that through the normal proxy can be considered the token of malicious code. Such can reduce the amount of analyzed data. If proxy application is not feasible, exception handling can be carried out on the HTTP traffic to obtain similar effect. To understand network penetration and proliferation, a minimum level of system interface log is needed, i.e., log interface to monitor the e-mail download as the main channel for APT attack and system log interface that monitors the privilege boundary test.

Lastly, there is a need to separate the patterns normally used by the organization member and abnormal patterns to narrow the detection range sufficiently. For that, all outbound traffic types within a specific period must be investigated, and acceptance of the investigated traffic should be manually judged within the organization. All outbound traffic except the HTTP/proxy traffic should be investigated for a specific period.

### 2.4 Description of Proposed Algorithm

The configuration needed to run the proposed algorithm is described as follows:

E1 The network packets are collected, and P (t, tcp/udp, ip1, ip2, port2, and payload) for each packet is extracted. It is the data obtained through packet proving/mirroring.

E2 E-mail logs are traced to accept the EL = (t, ip/pc) log (attachment download time/PC address). The e-mail log interface is needed.

E3 The privilege increase logs (Syslog) are traced to accept the ES = (t, ip/pc) log (log for privilege increase). The syslog interface is needed.

E4.C Call-back domain blacklist C = [d1, d2, ...] Already known blacklist IPs are periodically received from external agencies.

 

E5.D Internal DNS server D = [ip1, ip2, ...] It can be defined by setting internally.
E6.S SSL port S = [port1, port2, ...] It uses the well-known port list.

※ The total observation target data can be significantly reduced by handling the HTTP/proxy traffic as exception when the HTTP/proxy concept is applied as a policy in this stage.

The proposed algorithm is run to extract the following data as the result:

D1 List of IPs of suspicious zombie PCs (output) ZIP = [ ip1, ip2, ...] List of IPs of suspicious zombie PCs: The suspicious IP list is found with the reference control data.

D2 List of IPs/ports of suspicious C2 servers (output) CIP = [ip1:port1/confidence1, ip2:port2/confidence2, ...] List of ip:port/confidence of suspicious C2 server: It finds the list of servers suspected to be the clear C2 (Command & Control) server.

D3 Unfamiliar IP/port list (output) UIP = [ip1:port1/confidence1, ip2:port2/confidence2, ...] List of unfamiliar Ip:port/confidence: It is a list of IPs that are modestly suspected of unfamiliar IP connection.

D4 Up/download rate for each IP/port UPDOWNRATE = [ip1:port1:ratio1, ip2:port2:ratio2, ...]: It is used as data for suspecting data leak.

※ The E4.W white list is extracted first for a specific period, and suspicious IPs are then extracted using the following algorithm:

It runs in the following method beginning with the basic blocking algorithm:

**A1 Black White Processing:** If an outbound packet p1 (p1.ip2, p1.port2) belongs to the white list (W), it is passed. If P1.ip2 belongs to the blacklist (B), however, it is blocked, and an alert is generated.

**A2 Blocking of Unauthorized DNS Server Connection:** If outbound packet p1 is a DNS query (udp, port2=53), and if P1.ip2 does not belong to the internal DNS server (D), it is blocked, and an alert is generated.

**A3 Finding Call-back Domain Connections:** If P1.port2 = 53, udp, domain string d is extracted from the payload. If domain string d belongs to the call-back domain blacklist (C), however, it is blocked, and an alert is generated.

Suspicious IPs are found by monitoring the network activities as follows:

**A4 Finding Repeated Connection Attempts:** The recently connected packets (pz, py, pz ...) found from p1.ip2 and p1.port2 of the connection packet p1 are considered periodic external connection if the time differences between px, py, pz... (px.t-p1.t, px.t-py.t, pz.t-py.t, ...) are the same. If p1.p2 and p1. port2 belong to the service white list (W), it is passed. Otherwise, an alert is generated. P1.ip2:P2.port2 is registered as C2 server suspected IP/port/1 (D2), whereas p1.ip1 is registered as zombie PC suspicious IP (D2).

**A5 Finding Scanning:** After recent packets (px, py, pz...) with the same ip1 and ip2 of p1 are found, they are considered to be scanning action if the target port (px.port2, py.port2, pz.port2...) of px, py, pz... is the critical value or higher (ex.: 10 or more). In that case, p1.ip1 is registered as suspicious zombie PC (D2), and an alert is generated

 

**A6 Finding SSL Abuse:** It is considered SSL abuse if port2 of packet p1 belongs to the exclusive SSL port (S) and a text word is contained in P1.payload. In such case, p1.ip1 is registered as suspicious zombie PC IP (D1), and an alert is generated. Ip2:port2/1 is registered as suspicious C2 IP (D2).

**A7 Finding Connection to Unfamiliar IP/Port:** The (ip2, port2) pair is considered unfamiliar IP if p1.ip2:p1.port2 of packet p1 is not in the white service list (W) and port2 is 80 and not from Web proxy. In such case, (ip2, port2, 0) is added to the unfamiliar IP/port list (D3).

The detected suspicious IP values and other logs can be cross-tested to inspect suspicious IP additionally.

**A8 Finding through Correlation Analysis with E-mail Attachment Download Log:** The IPs of the PCs connecting to the suspicious C2 server (CIP) IP/port are searched first. If e-mail attachment download log EL (t, ip) of the PC IP exists within one minute before the time of initial action of A4, A5, or A6, the confidence of the C2 server increases. Subsequently, the IPs of the PCs connecting to the unfamiliar IP/port (UIP) are searched. The confidence of IP corresponding to D2 is increased.

**A9 Finding through Correlation Analysis with Privilege Increase Attempt Log:** If A4, A5, or A6 action occurred on the IPs of PCs recorded in the privilege increase attempt log (ES), and if the IP of the PC belongs to the zombie PC IP (ZIP), the confidence increases. It increases the confidence of IP corresponding to D1.

**A10 Finding Upload/Download Traffic Ratio Turn Around:** The UPDOWN-R values of all observed packets p1 are recorded and managed. IPs whose up/down ratio is turned around (outbound data volume being larger than the inbound data volume) are searched. SRC-IP is then registered as suspicious PC IP/Confidence=0, whereas the DST-IP:port is registered as IP2/port2/Confidence=0 since the IP is suspected of C&C IP (or the confidence value is increased).

**A11 Finding Common Connecting URL and Malicious Code Distributing URL:** URLs are searched from the payload string in the connection records within 1 minute before A4, A5, or A6 action to the suspicious zombie PC IP. If the URL strings overlap (same URL connecting to different zombie PCs), the URL becomes the malicious code distributing URL.

## 3 Result of the Algorithm Test

To verify the effectiveness of the proposed algorithm, the following test environment was constructed in a small office network:

- Period: January 16 ~ February 5, 2013
- Data collection: (1) Packet proving whole traffic from the point of office internal/external connection bottleneck, (2) collection of e-mail download log from the e-mail server, and (3) collection of syslog in the main development server
- Data output terminals: Based on users; 45 users including 34 full-time persons, meetings, and mobile working, total of 100 terminals including mobile phones and smartphones
- Daily traffic volume: Inbound packets - 80,000 ~ 250,000 packets; Outbound packets: 140,000 ~ 370,000 packets; Unique external IPs: 30,000 ~ 80,000.

The network configuration of the testing environment is shown below.

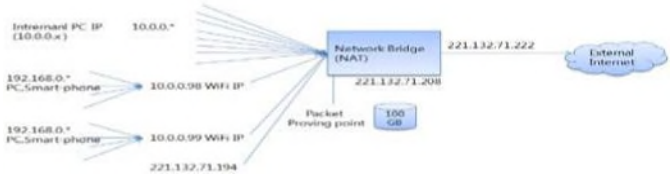


**Fig. 1.** The network configuration of the testing environment

Testing of the above network generated the following results:

a. There were a considerable number of connection attempts within a specific period. For IP:port, more than 300 cases were steadily observed each day.

b. Around 300 stable cases were observed daily. Although there were repeated connecting cases that were newly found, they were observed to be connecting to the existing periodically connecting IP when they are grouped by IP/C-class.

c. Some of the repeatedly connected IPs/ports were known to be used by hacking already. Virus checking of PCs connecting to 211.254.228.46 (update.windowupdate.org) found the malicious code, and IPs are not normally acting as the C2 server. After the malicious code is removed, the traffic is no longer detected. (This algorithm can be considered appropriate.)

d. It is considered scanning if there are 20 or more connections whose SourceIP -> TargetIP is the same but the point number is different. To find scanning of very slow speed, the observation range must be widened to 1 hour or longer. Thus, a big data DB to handle the very large data is needed.

e. We could not find any plain-text packet to 443 (SSL). It will be needed to expand the target to more SSL ports.

f. New IPs occurring daily were found to number more than 4,000 and less than 10,000 (excluding HTTP/80 traffic). The number did not decrease over time. Refer to the figure below.

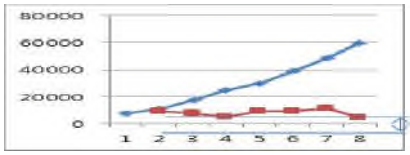


**Fig. 2.** New IPs occurring daily

Investigation of actual traffic indicated 3 types of patterns as shown below.

 

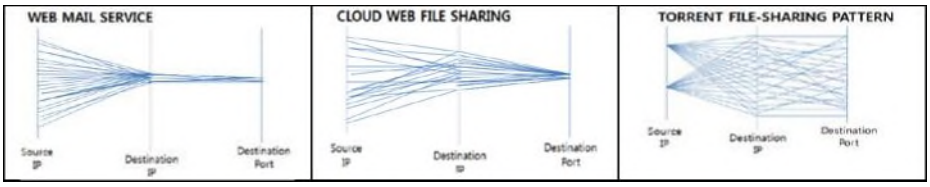


**Fig. 3.** Investigation of actual traffic types

If P2P file sharing services such as torrent cannot be identified because of such pattern distribution, it would be difficult to consider a pure new IP to be suspicious.

g. Removal of torrent connection pattern: Torrent IP connection is a pattern of connection from a single IP to 1,000 or more unique IPs/ports in a short period (less than 1 hour).

h. Removal of well-known cloud services: If the IP is found to be of a well-known SNS service such as Google, Facebook, Myspace, etc., through WHOIS query, the IP can be removed.

i. After removing these two patterns, the new IP cases are reduced to around 300 daily.

j. A8 verification: Searching of connection to the suspicious IPs within 1 minute of Web mail download from the PC IP indicated no significant result. The observation range must be widened longer.

k. A9 verification: Correlation analysis of the privilege increase log and PC IPs connecting to the suspicious IP is performed.

## 4 Observation from Testing in the Actual Network

Extraction of periodically repeating traffic was proven to be effective in finding the actually infected PCs. In the logic of considering the new IPs to be suspicious ones, the service was acceptable only when the cloud and P2P services were effectively removed.

Need for computing power: Although it was a verification of a small network with a small number of users, the volume of calculation was clearly high; even in a small organization, the daily packet distribution exceeded 50GB and 2 million packets.

The parallel-type big data processing system was judged to be the only one that can handle both in and out directions.

## 5 Conclusion

In this paper, the model and algorithm for detecting the latest APT attacks were proposed. The proposed model and algorithm were tested in a small office environment and verified to be somewhat effective. Note, however, that the actual

 

network situation was too complex to analyze with a simple algorithm using the packets.

Future studies may include collection of broader security logs and real-time monitoring and cross analysis of packet traffic simultaneously.

## References

1. [Com5] Command Five Pty Ltd, SK Hack by an Advanced Persistent Threat (2011)
2. Lee S.H.: Study of Penetration Detection Improvement Using Correlation Analysis of the Integrated Security System, Hanbat University MS Thesis (2010)
3. Lee H.W.: Design and Implementation of Integrated Event Log-Based Web Attack Detection System, Korea Society for Internet Information (No. 6, Vol. 11) (2010)

 