Statistical Analysis of Unauthorized-Access Log Data and its Interpretation

Hiroyuki Minami

Abstract We have studied how we analyse unauthorized network access logs and our empirical studies have suggested that we could classify the logs into some typical patterns and tried to develop methodology, to reveal them with aggregated statistical methodologies including symbolic data analysis (SDA). Our motivation is to extract some specific patterns from the logs. Many applications have been already developed to detect anomalies from them, but few are mainly based on statistics. To improve their quality, a mathematical viewpoint is key since most unauthorized actions are based on automatic algorithms. Thus we could apply some statistical (and intensive) model to them. When we develop an intensive statistical analysis for this data, SDA, known as a typical aggregated data analysis method would be applicable. In the study, we discuss how we aggregate the original log data and derive a reasonable classification and interpretation through the analyses.

Information Initiative Center
Hokkaido University, N11W5, Kita-ku, Sapporo 060-0811 JAPAN,
✉ min@iic.hokudai.ac.jp

Archives of Data Science, Series A
(Online First)
KIT Scientific Publishing
Vol. 6, No. 1, 2020

DOI: 10.5445/KSP/1000098011/13
ISSN 2363-9881

1 Introduction

According to the growth of the Internet, we have faced many kinds of cyber-attacks. As a step to attempt a discrimination of these attacks to protect against them, we have studied how we analyse unauthorized network access logs and we consider empirical attack patterns. Our previous analyses and studies have suggested that we could classify the logs into some typical patterns and we have tried to develop methodology to reveal them with aggregated statistical methodologies including Symbolic Data Analysis (SDA) (Bock and Diday, 2000). We introduced an application to analyse the logs based on the idea of SDA (e.g. Minami and Mizuta, 2016).

Our motivation is to extract some specific patterns from the logs including the cases brought by the unauthorized software. Many applications have been already developed to detect anomaly from the logs. Dias and Correia (2019) provide a good summary and references.

To improve the quality, a mathematical viewpoint and domain-knowledge are key ideas (Chen, 2004) since we are sure that most unauthorized actions are based on automatic algorithms from hosts under adversarial control. When we develop an intensive statistical analysis to the data, SDA known as a typical aggregated data analysis method (based on 'Concept', to represent data relations) would be suitable. In the study, we discuss, how we organize a SDA table with the elements of the log data and derive a desirable classification and interpretation on our real data.

2 Background

2.1 TCP/IP

TCP/IP is a popular abbreviated word and its formal name is "Transmission Control Protocol / Internet Protocol suite" (Socolofsky and Kale, 1991). The idea is based on layered network model and the framework consists of 4 layers (see Table 1). The bottom one is "Link" (or called Physical), the uppers are "Internet" (IP is available here) and "Transport" (TCP is used here) and "Application". Each layer is independent from the others. For example, we can use E-mail (in Application layer) regardless of Physical layer (e.g. fibre-optics,

wireless network). In short, we can investigate some specific layers without taking the others into consideration.

Table 1: TCP/IP suite and its elements.

Application Layer	Web (HTTP, HTTPS), E-mail (SMTP, POP, IMAP), etc.
Transport Layer	TCP / UDP
Internet Layer	IP, ICMP, etc.
Link Layer	Fibre Optics, Ethernet Wifi, etc.

The address in Internet layer is called IP-Address. Now, IP-Address version 4 (sometimes abbreviated as IPv4) is more popular in the world than the newer IPv6 (version 6). It consists of 32 bits and we divide them into 4 parts (that is, one part consists of 8 bits) and represent it with dotted decimal notation. For instance, we represent the IPv4 Address `192.168.1.1` as `11000000101010000000000100000001` in binary notation.

2.1.1 CIDR

From the original definition, the total number of IP-Address is 2^{32}, approximately 4.3 billions. Since it is too tough to handle each IP-Address individually, CIDR (Class-less Inter Domain Routing) is introduced as an idea of notation and collective operation. For example, "`192.168.0.0/24`" stands for 256 IP-addresses between "`192.168.0.0`" and "`192.168.0.255`". In the field of computer networks, a subset of IP-Addresses represented CIDR notation, we call it CIDR (address) block.

2.1.2 Port Number

To distinguish a transmission between the same devices, Port Number is introduced in the Transport Layer and two pairs, (Source IP-Address, Source Port) and (Destination IP-Address, Destination Port) designate a so-called "Internet Transmission". Typically, "Destination Port" stands for a sort of the transmission. For example, if it is 80, the kind of the transmission is based on

HTTP, usual (non-encrypted) Web access. Port numbers less than 1024 are known as ‘Well-known’ ports and no one should use them in a non-standard way.

2.2 Popular Protection against Cyber Attacks

We introduce some typical solutions to protect against cyber-attacks.

2.2.1 Filtering and Firewall

The technique of Filtering is to give a selection to pass/deny a transmission and is a basic idea in Firewall. The elements of the two pairs are typically adopted to determine a selection. For example, all transmissions from a certain host which had attacked the device should not pass and we shall add a rule to reject them in our firewall. Recently, more advanced solutions are proposed and implemented, called IDS (Intrusion Detection System) and/or IPS (Intrusion Protection System). The former is to notify us of suspicious network traffic and the latter is to block it automatically. Both of them watch all transmissions and try to identify a suspicious one, according to the elements.

2.3 Our Motivation

Recently, machine learning including deep learning techniques has been adopted in cyber-security applications. It might be effective. However, the trends on cyber-attacks will rapidly change like in a predator-prey game.
We can assume that quite few cyber-attacks run manually since it is too tough to type all instructions by hand. Thus, almost all cyber-attacks would carry out some algorithms, sequentially or simultaneously. This suggests, that it would be effective to detect and identify several patterns through statistical analysis on the attack data. As we see it, if the size is too large, then we seek for some aggregated idea of data analysis.

2.4 Symbolic Data Analysis

Symbolic Data Analysis (e.g. Bock and Diday, 2000) is one of the popular approaches on aggregated data analysis. The principal idea is to aggregate typical data (1st level) into some collective data (2nd level), including Class, Category and Concept. We believe that Internet-related data representation is suitable for SDA. For example, we can regard CIDR, simply as class because it is a set of (pre-defined) IP-Addresses. Our idea is illustrated in Figure 1.

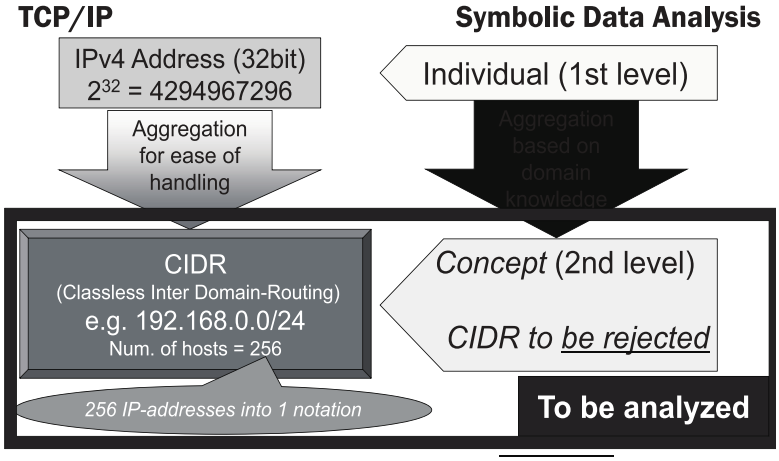

Figure 1: CIDR and SDA concept.

2.5 Suggestion from Pre-Analysis

Through our pre-analysis, we have found that some records have a few specific source port numbers while an operating system usually assigns random integers. Thus, it is straightforward that the records are produced by unauthorized software whose source port number is fixed and hard-coded. If we are able to detect these port numbers and classify the sort of the attack data, we could detect a specific attack pattern.

3 Application

3.1 Target Data

We have collected the data in our server for several years. Figure 2 is an overview of our network environment.

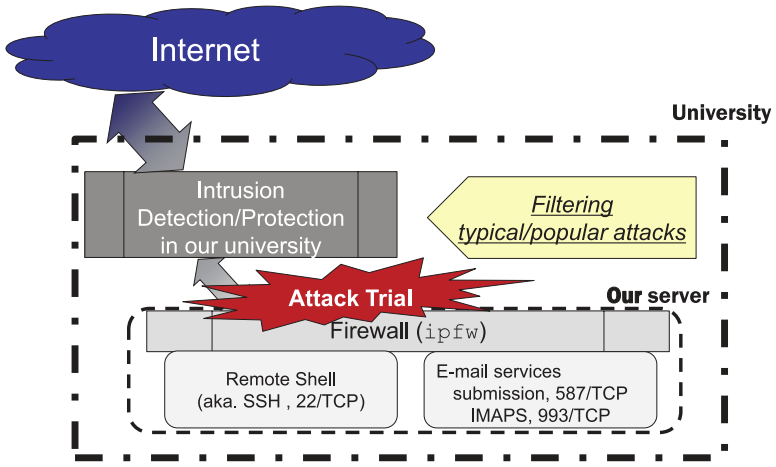

Figure 2: Our network overview.

Usual attack trials are blocked by our front Firewall. If an attack passes the 1st barrier, our next (and final) firewall named `ipfw` rejects and records it. Figure 2 describes the format. We daily add a new (rejection) rule to `ipfw` manually, when we recognize an attack from a new source IP-Address through another information. After that, a new attack from the address is rejected and recorded into the server. Figure 3 explains the elements of a log record.

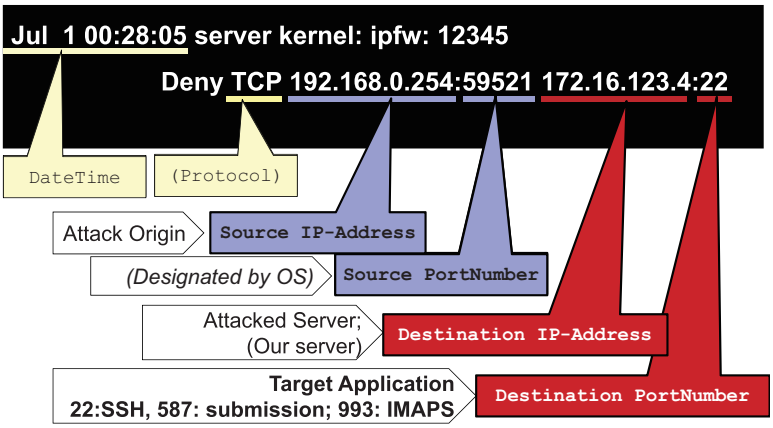

Figure 3: Example of the log data by firewall application `ipfw`.

In the study, we use the log dataset of January 2019, whose record size is approximately 5.1 thousands. The number of the CIDR blocks a source IP-Address is included is 447. Toward practical analysis, we exclude 3 CIDR blocks in advance since the number of the attack data is less than 3.

3.2 How to Classify the Data

Many methodologies have been developed in SDA including PCA, MDS, Pyramidal clustering. In the study, however, we adopt conventional hierarchical clustering to the aggregated data. There, our priority is to reveal the availability of our idea and find some rough trends through the analysis for further steps. In addition, we plan to apply clustering to the data but it might be complicated for an engineer to give an interpretation on a visual output of Pyramidal Clustering.

To apply clustering to the dataset, we need to define a similarity between 2 CIDR blocks (2nd level in SDA). As we described in Figure 1, we regard one CIDR as a datum in SDA then the number of the original IP-Address in one CIDR differs both in definition and practical observations. If an attack would occur randomly, a wide CIDR block would have a disadvantage. Then, we have to take the number of the original elements into account. Just to make sure, we

do not need to take care of destination port numbers since they are specified according to the purpose of a transmission.

To summarize them, we define the similarity s_{ij} between CIDR$_i$ and CIDR$_j$ as follows:

$$s_{ij} = \alpha_1 \times \left| \frac{NBP_i}{N_i} - \frac{NBP_j}{N_j} \right| + \alpha_2 \times \frac{1 - \#(SRCP_i \cap SRCP_j)}{\#(SRCP_i \cup SRCP_j)},$$

where N_k is the number of the IP-Addresses in the CIDR block k, NBP_k is the number of the attack records in k, $SRCP_k$ is the set of the source port numbers in k, α_1, α_2 are the parameters ($k = 1, 2, ...,$ the number of CIDR blocks). In the study, we set α_1, α_2 to 1.

3.3 Results

We implement and carry out our idea with Python and R, since Python is used to handle log data in the IT world, while R is popular for most statisticians.

With the dissimilarity, we apply hierarchical clustering (Ward method) to the dataset and get a dendrogram in Figure 4. Due to security management, we put a mask in the real CIDR notations with pool names from A to F. The leafs with a number in parentheses are the number of CIDR blocks.

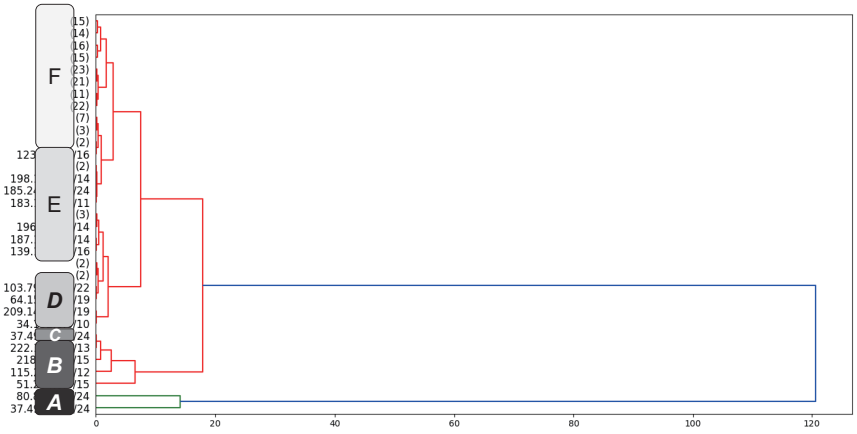

Figure 4: Dendrogram on CIDRs (The real CIDR representations are masked due to security).

From the result, we interpret some classified CIDR blocks:

The blocks in pool A have a tendency that a lot of attacks are observed in bursts (around 3 times in one minute), while we can give an interpretation that the attacks are observed periodically with the same source port number, from pool B. The attacks from pool C (a specific CIDR block) are observed periodically. The interpretation on pool D is different from those of the previous 3 pools since the attacks were targeted to various destination ports, that is, some popular Internet services whether the services are available or not. Then, we are wondering that the attacks are a sort of scan. We cannot give an interpretation in the pool E and F since the number of the attacks is much lower than for the other 4 pools. Then, the feature of the clusters might be just on size.

In total, we can aggregate the cyber-attack records in our server and give reasonable interpretations, compared with previous studies, without IP-Address aggregation. As the study gives us a good perspective, we can extend the analysis. For example, we mask the CIDRs in Figure 4 to hide the original IP-Address assignment. Of course, we have the original CIDR list and investigate which organization or country use a CIDR block. We will consider their commonality based on CIDRs, not based many IP-Addresses.

4 Concluding Remarks

In the study, we offer an empirical study to aggregate and analyse the dataset on cyber-attack as an SDA application. We still have some discussions. In the study, we do not focus on the element “Time”. We should consider that we apply a sort of Time Series Analysis to the dataset. As we mentioned, however, we can hardly assume that all cyber attacks come seasonally like climate. In short, we should discuss an appropriate SDA table configuration, the similarity between Symbolic data and related statistics/estimates in our future work. Also, we are going to use inner variation for further research. For example, the period (from start time to last time), the interval (between 2 attacks) in one CIDR are also considered as elements to be analysed in the SDA world.

Acknowledgements A part of the study is supported by JSPS KAKENHI Grant Number 18H03207.

References

Bock HH, Diday E (2000) Analysis of Symbolic Data. Bock HH, Diday E (eds.), Springer, Berlin. DOI: 10.1007/978-3-642-57155-8.

Chen WW (ed.) (2004) Statistical Methods in Computer Security. Dekker, New York. ISBN: 978-0-824759-39-1.

Dias LF, Correia M (2019) Big Data Analytic for Intrusion Detection System: An Overview. In: Handbook of Research on Machine and Deep Learning Applications for Cyber Security, pp. 292–316. Hershey, Ganapathi P, Shanmugapriya D (eds.). DOI: 10.4018/978-1-5225-9611-0.ch014.

Minami H, Mizuta M (2016) A Study on the Analysis of the Refused Logs by Internet Firewall. In: Proceedings of 2016 International Conference for JSCS 30th Anniversary. URL: `https://bit.ly/3BbGve9`.

Socolofsky T, Kale C (1991) A TCP/IP Tutorial. Network Working Group. URL: `https://tools.ietf.org/html/rfc1180`.