International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 15 (2017) pp. 5318-5323

Estimating Distribution from Data of malware-infected websites

DongYoung Lee[1] and Jinho Yoo [2, *]

[1]Department of Cyber Security, Sangmyung University, 20 Hongjimun 2-gil Jongno-gu, Seoul, Republic of Korea.

*[2]Department of Business Administration, Sangmyung University, 20 Hongjimun 2-gil Jongno-gu, Seoul, Republic of Korea. (*Corresponding Author)*

[2]Orcid: 0000-0003-4359-8009

Abstract

The damages caused by security accidents are increasing day by day. Most intrusions are accidents caused by infection of malicious code. The malicious code attack is presumed to have a certain pattern. In this paper, we try to identify the characteristics of reinfection or re-attack patterns of hackers by looking for statistical distribution, which is most similar to the interval day of infection of malicious code in websites. We found that the most similar distribution to the malware-infected data was the Gamma distribution. Beta and Weibull distribution also showed high similarity. Estimating distribution from data of malware-infected websites will be used for increasing in prediction efficiency and countermeasure development for security policy.

Keywords: Malware Infection; Density Estimation

INTRODUCTION

The number of users using the Internet is continuously increasing, and at the same time, the frequency of attacks with specific purposes such as stealing personal information and cyber terrorism is also increasing. There are many ways to hijack a user's information by infecting a website with malicious code hacking attacks. Malicious code's features and forms have evolved over the last few years, rapidly increasing their damage.

The frequency of malicious code infections continues to increase over time. For instance, infections of mobile devices was highly detected in 2015. In addition, websites frequently infected with malicious codes are abused as phishing sites. [1]

The number of attacks against Web attacks was estimated at 493,000 in 2014, but about 1,100,000 in 2015, a 117% increase. On the other hand, the number of malicious code-infected websites was estimated to be one infection per 1,126 sites in 2014, but by 2015, it is infected once every 3,172. [2] This is an increase in the number of attacks on websites, but the number of malicious code-infected websites has decreased due to the development of technologies to detect and defend them.

Research on existing malicious code mainly focused on detection techniques. Malicious code infections have been known to randomly be infected regardless of time. In this paper, we expect to find a certain pattern of infection of malicious code and try to identify its distribution. By statistically analyzing various malicious code infections, it will contribute to increase the efficiency of web site risk management such as predicting malicious code infection and managing resource allocation.

ANALYZE PREVIOUS RESEARCH AND EXPLORE RESERCH TOPIC

Malicious code-infected websites are divided into Landing Sites and Exploit Sites. Exploit sites are that distributes malicious code, and the Landing sites means an intermediate site that makes malicious code infected. As of January 2013, landing site has a 87% of malicious code infected sites among 180 million websites in Korea. In addition, it was found that 84% of the waypoint sites distribute different malicious codes each time. [3] This is because hackers are distributing new malicious code each time and linking multiple sites.

The malicious code of the exploit site and the landing site have different periods of activity. In the case of landing sites, the same malicious code is often not detected more than once, but the rate of redetection is high in the case of the distribution site. [4] In other words, although many landing sites are constantly changing to spread malicious code, the rate of reuse is high because the management site of the exploit site is managed by the administrator.

Jin-young Lee (2012) proposed a methodology to detect malicious codes by using two kinds of machine learning methods to classify statistical classification and SVM (Support Vector machine) . [5] Yong-Wook Jung (2013) conducted research for reducing false positives of detection by using similarity of malicious codes. He applied weighting to each attribute of malicious code to reflect the AHP (Analytic Hierarchy Process) technique and scored each malicious code. [6]

Personal information leakage accidents are one of the damages caused by malicious code infections. According to a study by Yoon-hee Hwang and Jin-ho Yoo, the average number of accidents has been found to follow the Poisson distribution.

This shows that an average of 12 major accidents occur in one year. In addition, the interval between occuring dates of the personal information leakage accidents was most similar to the exponential distribution, and the Weibull distribution was also similar. [7]

Malicious code can be reinfected at any time. In fact, the proportion of the infected area was higher than that of the landing site. In addition, hackers are using overseas-based exploit sites to make it difficult to delete them even if they are detected. [8] In this paper, we will focus on the transit sites that are infected in Korea because there are many difficulties from detection to action.

Most of the actions that infect websites are automated attacks by hackers using computers as well as passive attacks. It is expected that there will be a certain pattern in these behaviors.

In this paper, we try to identify the characteristics of reinfection or re-attack patterns of hackers by looking for statistical distribution which is most similar to the interval day of infection of malicious code.

RESEARCH METHODOLOGY

Collecting sample data

The sample data was collected by the detection result of malicious code detection system. The malicious code detection system analyzes and detects malicious code hidden on the homepage through static analysis. It then identifies weather the website is exposed to malicious every day. The web crawling function is used to collect suspicious link URLs from the source of a webpage and to store them and classified to find a new pattern. [9, 10]

In this paper, we used the detection results from January 1, 2012 to June 30, 2015. The total number of detections was 115,020, and the top 5 most frequently visited web sites were analyzed. According to the detection results, the landing sites were again infected with another malicious code during the treatment of the infected malicious code. In this study, we assumed that one infection occurred even if it was infected several times a day. The most infected Web sites were infected 124 times in 42 months. <Table 1> shows the URL and detection date of the most infected Site_A.

Table 1: Detection Number of Landing Site_A

Num	Detection Date	Landing Site_A
1	2013-09-20	http://[redacted]net/
2	2014-04-29	http://[redacted]net/
3	2014-05-08	http://[redacted]net/
4	2014-05-12	http://[redacted]net/
5	2014-05-20	http://[redacted]net/
.	.	.
.	.	.
.	.	.
.	.	.
.	.	.
.	.	.
120	2015-04-17	http://[redacted]net/
121	2015-04-22	http://[redacted]net/
122	2015-06-19	http://[redacted]net/
123	2015-06-20	http://[redacted]net/
124	2015-06-22	http://[redacted]net/

Analysis Method

We tried to find the most similar distribution using the 'K-S statistics' for the most infected top 5 landing sites(Site_A, B, C, D, E). In order to find the distribution most similar to the malware infection, we use four distributions of Beta, Gamma, Weibull, and Exponential. And we test the similarity and goodness-of-fitness.

The most frequent landing site_A was infected 124 times in 42 months. <Table 2> shows the interval day when the website was infected with malicious code.

Table 2: Detection Interval of Landing Site_A

Detection Date	Interval day
2013-09-20	-
2014-04-29	221
2014-05-08	9
2014-05-12	4
.	.
.	.
.	.
.	.
.	.
2015-04-17	3
2015-04-22	5
2015-06-19	58
2015-06-20	1
2015-06-22	2

As shown in <Table 3>, about 80% of the total infections of the landing site_A occurred between 1 and 4 days, and the interval with the greatest interval was 221 days.

Table 3: Detection Interval Frequency of Landing Site_A

Interval day	Freq	Per	Cumulative Per
1	49	39.84	39.84
2	27	21.95	61.79
3	6	4.88	66.67
4	16	13.01	79.67
5	9	7.32	86.99
6	3	2.44	89.43
7	3	2.44	91.87
8	2	1.63	93.50
9	1	0.81	94.31
10	1	0.81	95.12
11	1	0.81	95.93
12	1	0.81	96.75
17	2	1.63	98.37
58	1	0.81	99.19
221	1	0.81	100.00
Sum	123	100.0	

In <Table 3>, the 'interval day 1' means either landing sites has not remedied the infected malicious code, or has been treated but reinfected. If the other date interval is 2 or more, it means that the landing sites has an infection caused by an additional malicious code attack.

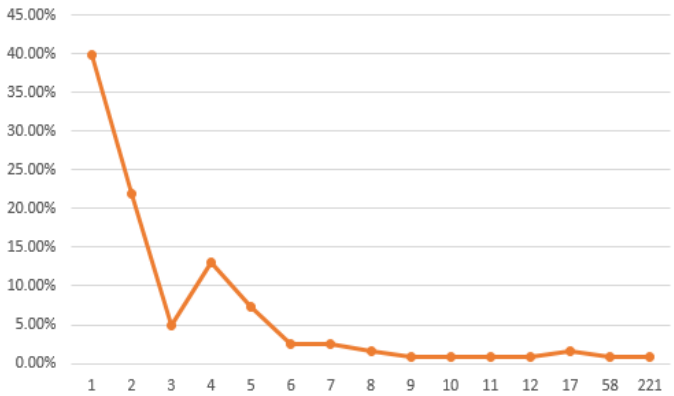

Figure 1: Detection Interval Day Graph of Landing Site_A

Estimation of optimal cumulative distribution function

We compared the cumulative distribution function of the data of the most infected top five landing sites with the statistical distributions in order to find the distribution most similar to the distribution of the interval day when the website is infected with malicious code. In order to find the optimal cumulative distribution function, we used K-S statistics according to the following steps.

(1) Variables (value of interval day) are sorted in ascending order. Next, we obtain cumulative rankings i of x and cumulative distributions of $F_L(i)$ and $F_U(i)$. x is the value of interval day at which the website is infected with malware.

$$F_L(i) = \frac{i-1}{n} \quad , \quad F_U(i) = \frac{i}{n}$$

In the formula, $F_L(i)$ is the lower boundary, $F_U(i)$ is the upper boundary, n is the total number of x, and i is the cumulative rank of x.

Table 4: Cumulative Distribution Value of Landing Site_A

i (rank)	x (Interval day)	$F(i)$ (Cumulative Distribution Value)
1	1	0.008130
50	2	0.406504
77	3	0.626016
83	4	0.674797
99	5	0.804878
108	6	0.878049
111	7	0.902439
114	8	0.926829
116	9	0.943089
117	10	0.951220
118	11	0.959350
119	12	0.967480
120	17	0.975610
122	58	0.991870
123	221	1

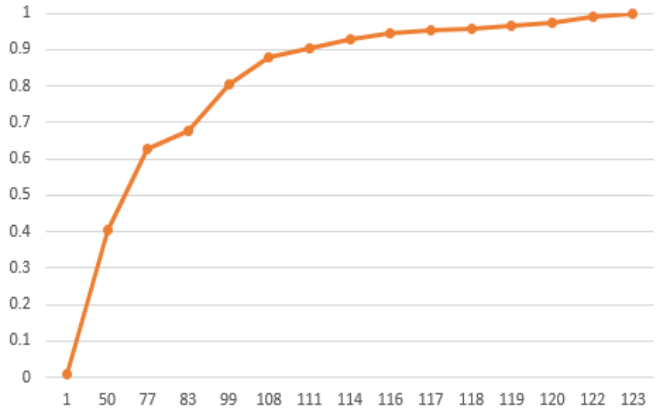

Figure 2: Cumulative Distribution Graph of Landing Site_A

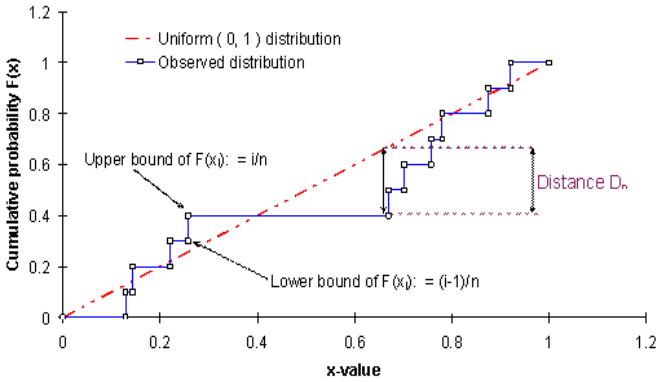

Figure 3: Calculation of the Kolmogorov-Smirnoff Goodness of Fit Statistic

(2) We estimate cumulative distribution values (x) of Beta, Gamma, Weibull, and Exponential and related parameters to find a distribution similar to the interval frequency of <Table 4>.

Table 5: Parameter of theoretical Cumulative distribution

Distribution	Excel Function	Parameter
Beta	BETA.DIST	χ, α, β
Gamma	GAMMA.DIST	χ, α, β
Weibull	WEIBULL.DIST	χ, α, β
Exponential	EXPON.DIST	χ, λ

At this time, it is calculated by applying the 'K-S statistics' methodology. This is used to estimate the distribution function closest to the distribution of the observed data, and quantifies the distance between the observed data cumulative distribution and the theoretical cumulative distribution through the result. [4] And we also compare the difference D_i between the cumulative distribution function $F_L(i)$ and $F_U(i)$ of observed data and the theoretical cumulative distribution function (x).

$$D_i = \text{MAX}(ABS(F(x)\text{-}\ F_L(i)), ABS(F(x) - F_U(i)))$$

In the formula above, D_i denotes the vertical distance calculated and the difference between the cumulative distribution of the actual data. From this the estimated can be known.

(3) We find the cumulative distribution with the smallest D_n cumulative distribution of statistics to be the 'K-S statistics'.

$$D_n = MAX(\{D_i\})$$

D_n means the largest vertical distance value D_i according to the difference of each point, and the smaller the value D_n is, the more similar the two distributions are.

In this case, the parameter with the lowest D_n value uses the 'Solver' function of Excel. 'Solver' changes the variable cell within the specified condition and repeats the calculation until the value of the target cell is minimum.

Table 6: Cumulative Distribution Value of Landing Site_A

Distribution			Value		D_n
Gamma	α	β	2.86	1.06	0.083994
Beta	α	β	2.86	934.69	0.084045
Weibull	α	β	1.93	3.32	0.093756
Exponential	λ		0.2		0.180638

(4) We repeat the simulation procedure (1) to (3) to find the statistical theoretical cumulative distribution most similar to the cumulative distribution of the observed data.

VERIFICATION AND ANALYSIS RESULTS

Simulation results for Site_A showed that the values of (α, β), ('2.86', '1.06') in the gamma distribution were measured at the optimal values, and the maximum difference value D_n was '0.083994'. We found the distribution of the interval days Site_A is infected with malicious code that is most similar to

the gamma distribution. Beta and Weibull distributions showed high similarity, which is much smaller than the exponential distribution.

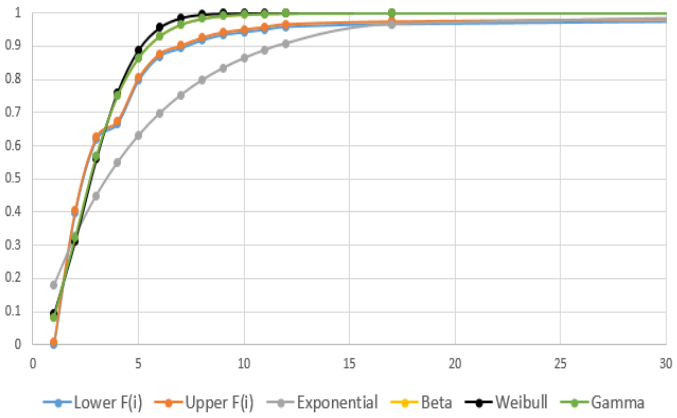

Figure 4: Comparison of Observed Data Cumulative Distribution and Theoretical Cumulative Distribution

<Table 7> shows the result of the cumulative distribution most similar to the observed data using the interval day of infection date of malicious codes Site_B, Site_C, Site_D and Site_E. Site_A, Site_B, Site_C, Site_D, and Site_E are the top 5 most frequently infected landing sites.

Table 7: Cumulative distribution

Site	Distribution			Value		D_n
B	Gamma	α	β	1.79	2.57	0.086291
	Beta	α	β	1.79	386.45	0.086448
	Weibull	α	β	1.44	4.98	0.094322
	Exponential	λ		0.14		0.131606
C	Gamma	α	β	3.88	0.55	0.128327
	Beta	α	β	3.87	1823.59	0.128394
	Weibull	α	β	2.24	2.43	0.141032
	Exponential	λ		0.27		0.235949
D	Gamma	α	β	2.60	1.30	0.78824
	Beta	α	β	1.80	456.64	0.108895
	Weibull	α	β	1.78	3.77	0.089370
	Exponential	λ		0.19		0.175782
E	Gamma	α	β	1.21	4.87	0.119239
	Beta	α	β	1.20	202.99	0.119526
	Weibull	α	β	1.13	6.12	0.122074
	Exponential	λ		0.15		0.137284

The results of the analysis of the top five landing sites with the highest malicious code infections showed that the Gamma distribution was most similar to the observed data cumulative distribution. In addition to the Gamma distribution, Beta distribution, Weibull distribution, and fitness were higher than exponential distribution.

Probability distributions are generally used to statistically predict various situations or data. For example, Gamma distribution can indicate the time that an electrical component fails. Gamma distribution is used to model errors in multi-level Poisson regression models. [11] In meteorology, Gamma distribution is the most suitable for the selection of radar rainfall and surface rainfall distribution. [13] In addition, a rainfall size distribution prediction model is proposed using the extended gamma distribution. [14]

CONCLUSION

There are a number of techniques currently available for malware detection. However, as time goes by, new malicious codes continue to evolve, and there are some parts that are difficult to defend. In this paper, we tried to analyze the patterns of malicious code infections in Korean websites and use them for detection and response. As a result, the gamma distribution has the highest similarity with the observed cumulative distribution in all the top 5 landing sites where malicious code infections are most frequent. Also, the probability that a hacker attempts to infect malicious code again within 7 days is 85.4% on average of 5 landing sites.

The result of this paper can be used as a basic model for establishing preventive countermeasures against detection by predicting the frequency of the occurrence of malicious codes in Web sites and estimating the probability. In the future, we will study the correlation between various features of web sites and malicious code to improve detection accuracy and increase efficiency.

ACKNOWLEDGEMENT

This research was supported by a 2017 Research Grant from Sangmyung University.

REFERENCES

[1] McAfee., “2016 McAfee labs Threats Report: March”, 2016.

[2] Symantec., “Internet Security Threat Report: Voulme 21”, 2016.

[3] Han, Y. I. Lee, T. J. and Park, H. R., “Structural and characteristic analysis of malware network”, Korea Institute of Communication Sciences, 2014.

[4] Yoo, D. H., Kim, J. S., Jo, H. S. and Park, H. R., "Analysis of web-based nature of malware attacks spread", The Journal of the Korean Institute of Communication Sciences, Vol 315, No.5, pp. 15-19, 2014.

[5] Lee, J. Y., "Malicious webpage detection based on the spreading patterns of malware," Dankook University, 2012.

[6] Jung, Y. O., "similarity analysis of malicious codes applied attributes and weights", Jeonnam University, 2013.

[7] Hwang, Y. H, Yoo, J. H., "A Study on the Distibution Estimation of Personal Data Leak Incidents", Journal of The Korea Institue of Information Security & Cryptology, Vol 26, No.3, pp. 799-808, 2016.

[8] Edaily., "Exploit site malicious code, evolving from domestic to overseas IP-based," 2014.

[9] Kim, K. H., "A Study of Techniques in the Web Homepage Malware Detection", The Journal of Electronic Communication Sciences, Vol 8, No.2, pp. 449-452, 2014.

[10] Korea Internet & Security Agency., "A Study on Analyzing the Current Malware Detection Technologies and Planning for the Development Model of Detection & Response System", 2015.

[11] https://en.wikipedia.org/wiki/Gamma_distribution

[12] Lee, D. W., Kang, C. W., "The Statistical Design and Application of control Charts for a Gamma CV", Society of Korea Industrial and Systems Engineering Fall Conference, 2005

[13] Kim, T. J., Lee, D. R., Kang, S. M., Kwon, H. H., "Generation of radar rainfall data for hydrological and meteorological application (I): bias correction and estimation of error distribution" The Journal of Korea Water Resource Association, Vol. 50, pp.1~15, 2017

[14] Park, Y. H., Lee, J. H., Pack, J. K., "Empirical Study on the Prediction of Rain Attenuation in EHF(44 GHz) Band," Journal of the Korea Electromagnetic Engineering Society, 2005