Discovering malware based on co-clustering host-domain graphs

Wenqian Zhang¹*, WeiNa Niu², Jinhong Zhang³ and Jingqi Li⁴

¹The University of Electronic Science and Technology of China, Chengdu, Sichuan, China
²The Sichuan University, Chengdu, Sichuan, China
³The University of Electronic Science and Technology of China, Chengdu, Sichuan, China
⁴The University of Electronic Science and Technology of China, Chengdu, Sichuan, China

*Email: a306810397@gmail.com

Abstract. Malware domain discovery based on passive DNS graph analysis supplements existing methods through DNS request and response data analysis. However, the method does not consider the IP relevance and the high complexity of the suspicious computing process. This paper proposes a malware domain discovery method based on passive DNS and IP relevance. The method calculates reputation score, which combines the shortest path from malicious domains to unknown domain with the malicious IP ratio to determine whether the domain malicious.

1. Introduction

Attackers often use malware to control infected hosts to steal sensitive information. At present, behavior-based malware detection method has high accuracy by running malware in sandbox environment to analyze its dynamic behavior characteristics. However, this method is not only time-consuming but also has huge resource overhead.

In order to improve the recognition rate of Issa’s mothed, a malware domain name discovery method using global domain name and IP correlation is proposed in this paper. Firstly, the request domain name record is filtered according to the domain name whitelist, and the suspicious domain name is retained. Secondly, the passive DNS database is queried to generate the domain name-IP two-part graph. Then, according to domain name and IP blacklist, the selection of seed domain node and the calculation of suspicious IP ratio of unknown domain name are carried out respectively. Finally, the unknown domain name prestige value is inferred by combining the suspicious score of domain name and the suspicious IP ratio of connection, and the domain name larger than the threshold value is identified as malware domain name.

2. Related work

At present, there are many methods for malware domain name identification for DNS traffic at home and abroad, but in general, it mainly focuses on primary and passive DNS data analysis, host-domain name and domain name-IP graph reasoning.
In the enterprise network, the host-domain map reasoning and active DNS data analysis are combined to detect the malware domain name. According to the belief propagation algorithm, the suspicious domain name is searched in the host-domain name map [1]. Xu L proposes an APT malware detection method based on malicious DNS and traffic analysis [2], which extracts features based on active DNS records and combines decision tree algorithms to identify malicious domain names.

Compared with previous studies using DNS analysis to detect malicious domain names, the method in this paper is quite different. Passive DNS analysis method is used, and compared with reference [3], [1], and [2], it is not necessary to obtain large-scale data for characteristics. Compared with the reference [1], the analysis based on the requested domain name record requires a relatively small amount of data and is highly targeted. However, when calculating the suspiciousness of the unlabeled domain name, the method needs to calculate the path of all the seed nodes to the current domain name, and only considers the influence of the domain name on it, and the time complexity and the false positive rate are high [4].

3. Malware domain name discovery based on passive DNS domain name and IP affinity

This section describes the proposed method for discovering potential malware domains based on passive DNS domain names and IP affinity [5]. The flow chart of the proposed method is shown in figure 1. It mainly consists of two parts: 1. The pre-processing stage is responsible for filtering the normal domain name according to the Alexa value and the popular domain list in the monitored network; 2. Using graph reasoning to identify the potential malicious domain name. wherein the domain name suspicious score is determined by the global domain name relevance and the proportion of connected malicious IP.

![Figure 1. Method flow chart.](image)

3.1. Generating a domain name - IP diagram

First, according to the list of suspicious domain names accessed by the host in the enterprise network, the IP list corresponding to the domain name is generated by the A record in the passive DNS database. Secondly, the undirected bipartite graph is used to represent the domain name-IP map, which represents the set of suspicious domain names accessed, indicating the corresponding IP set of the domain name in the passive DNS database, and the matrix represents the connection relationship between the domain name and the IP. For example, the domain name-IP relationship is as shown in figure 2, and its adjacency matrix E is as shown in equation (1).

![Figure 2. Domain name - IP Diagram.](image)
\[ E = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \end{bmatrix} \] (1)

3.2. Mapping domain name association diagram

Because some domain names have the same IP, there is a correlation between the domain names in the domain name-IP map. Building a domain name association graph by single mode mapping \( G(D,W) \) indicates the set of suspicious domain names accessed, and the edge weight matrix to indicate the strength of the association between domain names, as shown in equation (2). The domain name association diagram of the example figure 2 is shown in figure 3.

![Diagram](Image)

**Figure 3.** Illustrates the domain name association diagram.

3.3. Calculating suspiciousness

The suspicious value of each unlabeled domain name is calculated based on the correlation between the domain names and the proportion of malicious IPs corresponding to the domain name. Intuitively, the longer the path to an unmarked domain name to a known malicious domain name, the smaller the association between the domain name and the malicious domain name, or even negligible. Therefore, for each seed node, only the weight of the shortest path is selected and substituted. Assume \( S = \{S_1, S_2, \ldots, S_i\} \) that it is a seed node, that is, a known malicious domain name set; the impact value of the seed node on the unlabeled domain name is obtained by formula (3):

\[ P(d, S) = \frac{\sum_{i=1}^{n} \frac{1}{2^{len}} P(d, S_i)}{2} \] (2)

\( P(d, S) \) indicates the value of the seed node's impact on the domain name, \( len \) indicates the shortest path length from the seed node \( S_i \) to the domain name \( d \). The hypothesis \( Ld, d = \{d, d_1, \ldots, d_i\} \) is the shortest path from the seed node \( S_i \) to the domain name \( d \), which \( P(d, S) \) is the product of the shortest path weight from the seed node \( S_i \) to the domain name \( d \).

The suspicious value of an unlabeled domain name \( d \) is obtained by equation (4):

\[ M(d) = \alpha P(d, S) + \beta R_{IP-M} \] (3)

\( \alpha, \beta \) is the parameter, satisfy \( \alpha + \beta = 1 \); \( P(d, S) \) indicates the impact value of the seed node on the domain name \( d \), \( R_{IP-M} \) indicating the proportion of malicious IP in the IP list corresponding to the domain name \( d \). Assume \( I = \{i_1, i_2, \ldots, i_n\} \) is the domain name corresponding IP list, \( IM = \{i_1, i_2, \ldots, i_n\} \subset I \) is the domain name corresponding IP list, so \( R_{IP-M} = \frac{m}{n} \).
4. Experimental results and analysis

4.1. Experimental environment and assessment methods
Randomly select several hosts to randomly access some blacklisted domain names in the lab environment, replay the traffic generated by the public APT malware running in the lab network environment, collect the network traffic generated every day, and take it as an example. verification. In order to quantitatively compare the recognition effects, the value is selected as the evaluation standard as shown in the formula (4). Indicates the number of malicious domain names that are correctly identified. The non-malicious domain name is determined as the number of malware domain names, and the malware domain name is identified as the number of non-malicious domain names. The machine running the passive DNS-based malware domain name recognition algorithm runs the 64-bit Windows 10 operating system, the programming language is python 3.5, and the development environment is PyCharm.

\[
F1 = 1 - \frac{FP + FN}{2TP + FP + FN}
\]  (4)

4.2. Experimental setup
On a daily basis, the statistical laboratory's domain name access for half a year keeps the records in the top 1/3 of the visit in the list of popular domain names. There are 59 suspicious domain names and 22 malicious domain names, accounting for 37.288%. Then, according to the passive DNS database, obtain a historical IP list corresponding to the suspicious domain name (a total of 833 different IP addresses), and query the malicious IP list to obtain the proportion of malicious IP corresponding to the domain name, compared with the method based on passive DNS graph reasoning. There are four seed nodes in this paper. Next, the shortest path from each seed node to the unlabeled domain name is calculated. The product of the weights of all reachable paths does not need to be calculated, which greatly saves time overhead. Finally, the suspicious calculation is based on the proportion of malicious IP and the degree of association between the seed node and the untagged domain name to identify a malicious domain name that is associated with a known malware domain name. The machine running the passive DNS-based malware domain name recognition algorithm runs the 64-bit Windows 10 operating system, the programming language is python 3.5, and the development environment is PyCharm.

The experimental parameters of the method and Issa method are set as shown in table 1 below:

| Table 1. Parameter setting. |
|-----------------------------|
| Method | threshold | $\alpha$ | $\beta$ |
| Issa Method | 1.8 | - | - |
| Method of this paper | 0.4 | 0.4 | 0.6 |

4.2.1. Comparison of recognition accuracy
In table 2, the two methods judge the non-malicious domain name error as the number of malicious domain names has not changed, and the number of correctly identified malware domain names has changed slightly, but a large number of malware domain names are not recognized, mainly due to This experiment uses the public APT malware dataset.

| Table 2. Comparison of recognition accuracy. |
|-----------------------------|
| method | TP | FP | FN | F1 |
| Issa Method | 7 | 1 | 11 | 53.85% |
| Method of this paper | 8 | 1 | 10 | 59.26% |

Compared with Issa's passive DNS map-based malware domain name discovery method, the
malware domain name detection method in this paper reduces the number of malicious domain names that are misidentified, mainly because there is a domain name that is related to known malicious domain names. Not large, but the domain name returned a large number of malicious IPs in the historical request, and Issa's method based on passive DNS graph reasoning did not consider the IP address of the domain name.

4.3. Comparison of detection time
In the table 3, when verifying the test data, the overall time consumption of Issa's method is nearly twice that of this method, mainly because the method first filters some normal domain names, which reduces the workload; secondly, when calculating the impact of seed nodes on unlabeled domain names, this method uses only the shortest path, further reducing the path traversal time. Combined with the data in table 2, the recognition accuracy of the method is further improved on the basis of reducing the recognition time.

| Method               | Detection time(seconds) |
|----------------------|-------------------------|
| Issa Method          | 17.2                    |
| Method of this paper | 5.8                     |

5. Conclusion
The continued outbreak of malware poses a serious threat to network security, and most of the software uses DNS to establish a connection between the attacked host and C & C. The main recognition method based on DNS traffic is to extract features by analyzing DNS request and response packets, and the domain name recognition method based on passive DNS graph reasoning can effectively supplement it. In order to alleviate the recognition method of Issa, this paper proposes a malicious domain name discovery method based on passive DNS and IP correlation, which combines the proportion of malicious IP when calculating the domain name prestige value. At the same time, the calculation process of known malicious domain name influence value is improved. The experimental results show that the recognition accuracy of the proposed method is improved by nearly 6% and the detection time is reduced by nearly two times compared with Issa's method.

Acknowledgments
This work was supported by National key research and development plan of China under grants No.2016QY04W0800 and Basic research business fees of central colleges of China under grants No.20826041B4252.

References
[1] Zhao G, Xu K, Xu L, et al. Detecting APT malware infections based on malicious DNS and traffic analysis[J]. IEEE Access, 2015, 3: 1132-42.
[2] Manadhata P K, Yadav S, Rao P, et al. Detecting malicious domains via graph inference[C]//European Symposium on Research in Computer Security. Springer, Cham, 2014: 1-18.
[3] Soltani S, Seno S A H, Nezhadkamali M, et al. A survey on real world botnets and detection mechanisms[J]. International Journal of Information and Network Security, 2014, 3(2): 116.
[4] Bilge L, Kirda E, Kruegel C, et al. EXPOSURE: Finding Malicious Domains Using Passive DNS Analysis[C]//Ndss, 2011.
[5] Antonakakis M, Perdisci R, Dagon D, et al. Building a Dynamic Reputation System for DNS[C]//USENIX security symposium. 2010: 273-290.