Developing Apt Attacks Detection System Based on Correlation Analysis Methods

Cho Do Xuan, Tisenko Victor Nikolaevich, Do Hoang Long, Nguyen Vuong Tuan Hiep, Le Quang Sang

Abstract: Advanced Persistent Threat (APT) is an exceptionally perilous attack with a specific target and purpose. It consists of various complex and devious techniques in order to be able to obtain a highly secured trade secret, sensitive information. Currently, the APT attack is tremendously difficult to deal with because of its unique design for each target, which makes prior experiences and rules less accurate in detecting APT attacks. In addition, the APT detection method also must not rely on any single procedures or solutions but to include several phases and technologies. On the other hand, correlation analysis technique is a mathematic one which figures how separate elements affect each other and produces conclusion based on multiple factors mutual properties. Hence, in this report, correlation analysis technique is proposed by the authors.

Keywords: Information Security, APT, unknown domain, attack detection, DNS log, Network traffic, correlation analysis, abnormal behavior, machine learning.

I. INTRODUCTION

According to the APT attack characteristics, typical steps and lifecycle which is presented in [1] and [2], the method has specific target and purpose. Any organizations, individuals, businesses or government agencies, the government can be the victim of this technique. It's showed in [3] that APT attack is significantly harder to detect than any other threat due to its unique traits in an attack scenario For the most part, the lack of public data is the biggest obstruction in detecting an APT attack is. Especially, its victims rarely reveal any details or acknowledge themselves as victims of APT attacks. APT attacks are sophisticated even for new attack methods. However, they are divided into four main phases [1]: reconnaissance, privilege escalation; information extraction and delete traces.Hence, the necessity of correlation analysis technique in detecting APT attacks is unquestionable. This technique requires examination of different phases of the attack in order to make up for attack data inadequacy and improve the chance to detect APT attacks.

II. RELATED WORKS INTRODUCTION TO CORRELATION ANALYSIS TECHNIQUE

The correlation analysis techniques to APT attack are defined in the report [4] as follows: "One method to reduce the number of false positives for bot detection is to require several correlated events before raising an alert. This allows the system to use events that by themselves have a high false-positive rate. However, by requiring multiple events the system is able to filter out the most false positives. The events may be correlated for a single host or for a group of hosts. The advantage of using correlations to detects bots is that there are fewer false positives compared to using just the individual events. At the same time, this can be a disadvantage because stealthy bots, which generate just one or two events, may not be detected". Moreover, the article [5] states a fact about this technique: "Utilize statistical and correlation methods to analyze the latest trends in malware. This is the key that ties all of the other methods together since it meshes rule sets, log examination, and data exfiltration monitoring. Correlation methods are used to examine whatever alerts are currently configured and to look for relationships between each alert that is triggered. These relationships can be with regard to the type of alert, port number or any other type of selector configured by the security analyst. Statistical methods do not rely on prior knowledge of an attack vector, but rather on the time and frequency of a set of alerts". According to the authors, the correlation analysis technique for APT attack detection is a mathematics technique to determine the relationship between the discrete elements in the system so as to conclude whether there is or isn't an APT attack on the system. In short, applying correlation analysis techniques to APT attack detection systems is an increase in its accuracy. For example, identify existence of APT payload in the system rely only on abnormal events from DNS logs, Network traffic or Web logs is arduous. On the other hand, the combination of irregularities on DNS logs or Network traffic or Web logs makes a more certain conclusion about the compromise of the system. There are several methods to analyze mutual connection between events in the system for detecting APT attacks. In this paper, we propose building models of APT attack detection based on correlation analysis of the Domain Name System (DNS) and Network Traffic.

III- SOME CORRELATION ANALYSIS MODELS

In correlation analysis approach to uncover an APT attack, there are two methodologies:

- **Parallel methodology**

The parallel methodology is to use different algorithms to analyze and detect attacks from the input. First, input data are DNS, Network traffic logs or Weblog. Then, the results are yielded by all algorithms. After that,

Revised Manuscript Received on March 30, 2020.
* Correspondence Author

Cho Do Xuan, FPT University Hanoi, Vietnam.
Tisenko Victor Nikolaevich, Peter the Great St. Petersburg Polytechnic University Russia, St.Petersburg, Polytechnicheskaya, 29
Do Hoang Long, FPT UniversityHanoi, Vietnam
Nguyen Vuong Tuan Hiep, FPT UniversityHanoi, Vietnam
Le Quang Sang, FPT UniversityHanoi, Vietnam

a conclusion is returned by the correlation algorithm [6, 7]. Additionally, the conclusion also shows any computers in the network are being attacked by APT if there are any [8]. Also, the paper [8] presented a correlation analysis based on statistic theory to find unusual attributes of the APT attack. Figure 1 model describes the parallel methodology for APT attack detection.

.

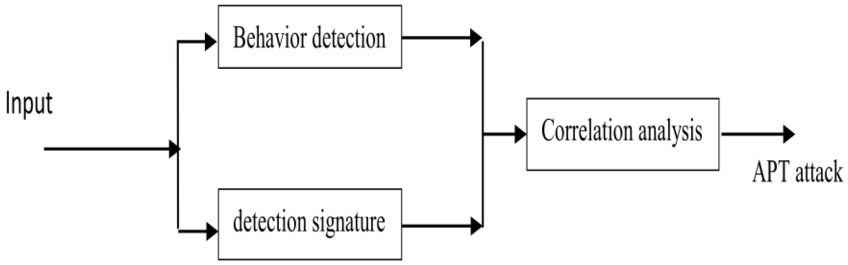


Fig. 1. Model of parallel 2-layer attack detection in APT

- **Synchronous methodology**

The synchronous methodology is complex because it’s a combination of serial and parallel design. For one thing, data from multiple sources are analyzed by serial or parallel elements to detect attacks. Hence, this approach is suitable for a system doesn’t require results from every layer and each result is stricter than its prior. In fact, this tactic depends on abnormal events from different places and times instead of a single strange incident. In this situation, all results from every element are processed by a correlation algorithm in order to generate the final deduction [6, 7, 9, 10]. Figure 2 model describes the synchronous methodology for APT attack detection.

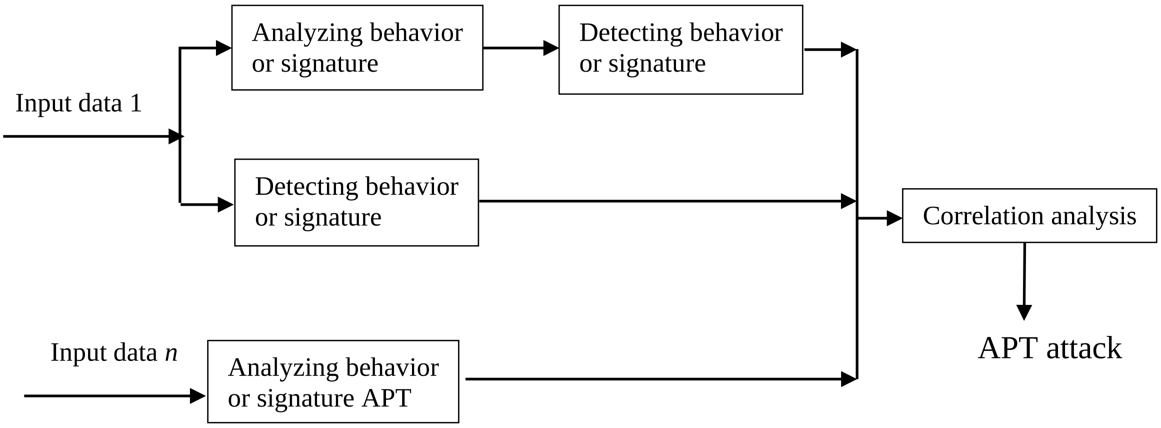

Input data 1
Analyzing behavior or signature
Detecting behavior or signature
Detecting behavior or signature
Correlation analysis
APT attack
Input data *n*
Analyzing behavior or signature APT

Fig. 2. A model complex in APT attack detection

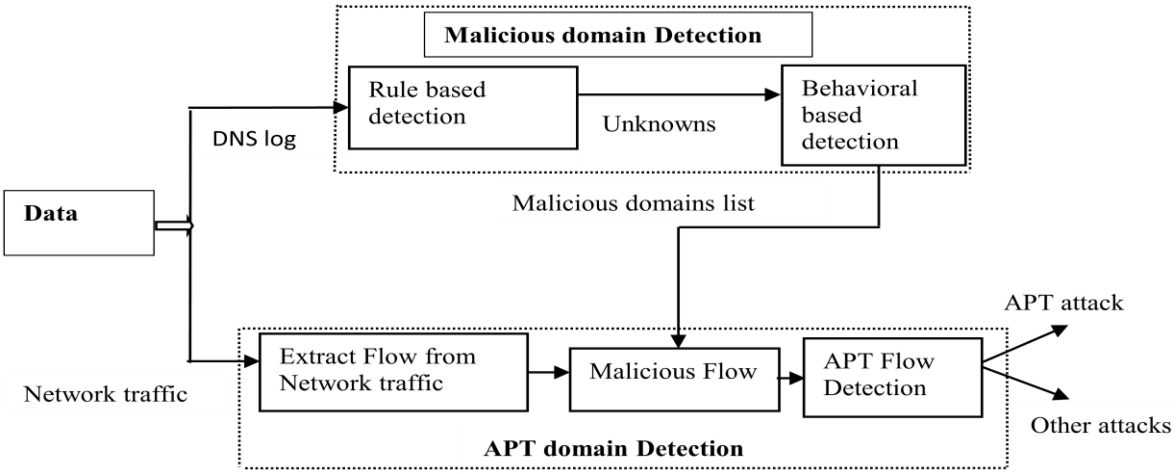


Fig. 3. The process of detecting malicious domain based on Rules

IV- PROPOSING APT ATTACK DETECTION MODEL USING CORRELATION ANALYSIS TECHNIQUE

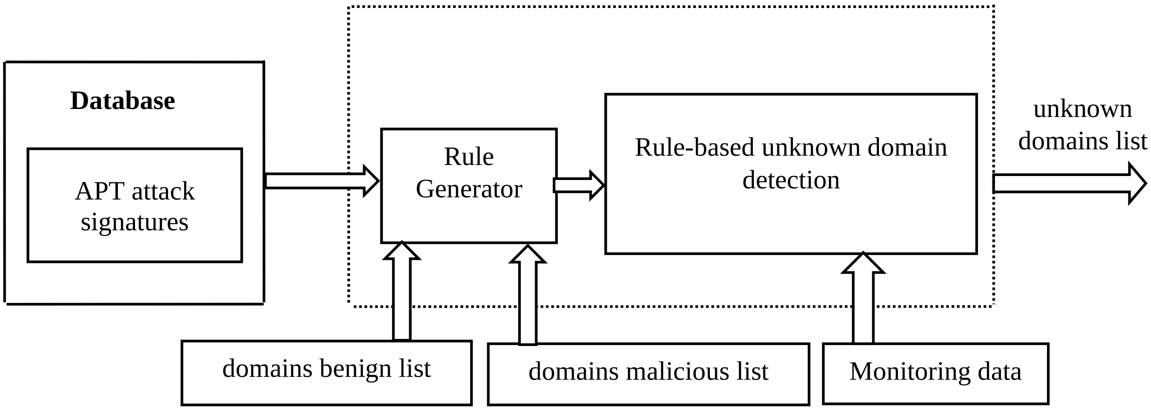


Fig. 4. The process of detecting malicious domain based on Rules

Undoubtedly, as shown in Figure 3, the synchronous approach is chosen among correlation analysis techniques by the authors to detect APT attacks. Particularly, machine learning algorithms are used in this technique to determine the mutual characteristics of DNS and network traffic. As a result, the wrong conclusions about the APT attack are eliminated. Thus, the precision of the detection system is greatly enhanced. To sum up, the implementation steps of this technique are the following:

- Step 1: DNS, Network Traffic and Weblogs collected from several servers are used selectively for input for APT attack detection module which utilizes correlation analysis. In which, DNS logs are examined by a malicious domain name detection module for harmful sites. For this purpose, rule-based analysis and behavior-based analysis are applied cooperatively by the authors.
- Step 2: Network traffic log is categorized into Flows. These flows are important information of the Network Traffic. They would be linked to the respective domain name by comparing the two lists.
- Step 3: List of malicious domains and list of network flows from step 1 and 2 are scrutinized and inspected in order to distinguish APT attack domains from other attack domains. For this problem, the authors suggest the following procedure to study relevant properties between DNS data and Flows data which aid us in concluding whether the system is under APT attack or not:

✓ Phase 1: Finding connections between malicious domains and the entirety of network flow. Afterward, records of malicious DNS and their corresponding flows are Phase 1's result.

✓ Phase 2: Detecting APT flows. The prior phase's records are analyzed so as to indicate APT scenario flows as well as other attack flows.

Additionally, in order to accomplish the said tasks, the APT attack detection model using correlation analysis requires the following components (see Figure 3):

- Data center: The data center stores data and provides information for monitoring and tracking network attacks. The data stored in the data center includes DNS logs, network traffic (Pcap); and all data has to be normalized and pre-processed. Information extracted from the data center is related to behaviors and attributes of attacks.
- Detection malicious domains: The malicious domain detection module sorts all domain in DNS logs into benign domains and malicious domains, using two main techniques: rules analysis and behavior analysis.
- APT Domain detection: The APT domain detection module provides methods, techniques, and algorithms to accurately tell malicious domains of an APT attack campaign apart from other attack scenario malicious domains

V- SEVERAL METHODS APPLIED IN CORRELATION ANALYSIS TECHNIQUE

As shown above, the APT attack detection model based on correlation analysis consists of two main modules: malicious domains detection and APT domains detection.

- **Malicious domains detection**

Malicious domain detection method has 2 stages:

- Stage 1: Detecting abnormal domains based on the defined rules: compare the input data with the set of known APT attack domains set of malicious domain and set of usual domains. Malicious domains detection module is demonstrated in Figure 4).

Explanation to module components:

APT attack Signatures database: The database stores signatures such as domains, IP C&C, of actual APT attacks that have been collected and extracted from reports of well-known security firms and research labs [11,13].

Monitoring data: Data is collected from server systems. This data consists of many components and different structures of DNS logs.

Domains benign list: These domains have been rated as secure domains and are collected from trusted sources [14].

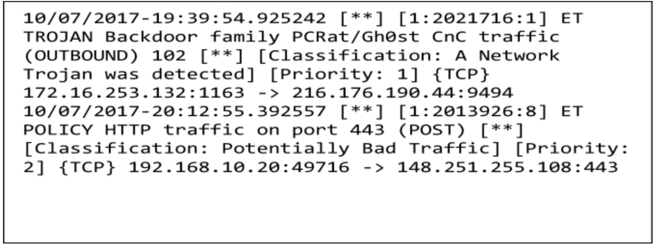
Domains malicious list: are domains that have been rated as unsafe and collected from trusted sources [15, 16, 17].

Rule Generator: This is a technique for automatically generating rules from signatures. Signatures of known malicious APT code, a list of benign domains and malicious domains have been used as input for generating rules used for monitoring and tracking accesses of unknown domains. To create these rules, users can either write their own rules enforcement programs or use open source tools such as Suricata [18], Bro IDS [19].Table 1 shows the results of the rule generation process using the Suricata tool

Table 1. Example of rules generation process from signatures database

Input data	rule obtained correspondingly
AttackName: Attack on Olympic Pyeongchang Domain: thlsystems.forfirst.cz IP: null File: null Port: null Protocol: ip CreatedDate: 08-06-2018 UpdatedDate: 10-06-2018 References: https://securingtomorrow.mcafee.com/mcafee-labs/malicious-document-targets-pyeongchang-olympics/	alert udp $HOME_NET any -> any 53 (msg:"test - Cuoc_tan_cong_vao_Opympic_Pyeongchang - DNS request for thlsystems.forfirst.cz"; content:"|01 00 00 01 00 00 00 00 00 00|"; depth:20; offset: 2; content:"|0A|thlsystems|08|forfirst|02|cz"; flow:to_server; fast_pattern:only; nocase; classtype:trojan-activity; reference:url,https://securingtomorrow.mcafee.com/mcafee-labs/malicious-document-targets-pyeongchang-olympics/; sid:15; rev:1;)
	alert http $HOME_NET any -> $EXTERNAL_NET $HTTP_PORTS (msg:"test - Attack on Olympic Pyeongchang - Related URL (thlsystems.forfirst.cz)"; content:"thlsystems.forfirst.cz"; http_uri; flow:to_server,established; classtype:trojan-activity; reference:url,https://securingtomorrow.mcafee.com/mcafee-labs/malicious-document-targets-pyeongchang-olympics/; sid:16; rev:1;)
	#alert tls $HOME_NET any -> $EXTERNAL_NET $HTTP_PORTS (msg:"test - Attack on Olympic Pyeongchang - Related TLS SNI (thlsystems.forfirst.cz)"; tls_sni; content:"thlsystems.forfirst.cz";flow:to_server,established; classtype:trojan-activity; reference:url,https://securingtomorrow.mcafee.com/mcafee-labs/malicious-document-targets-pyeongchang-olympics/; sid:17; rev:1;)
Input data	rule obtained correspondingly
AttackName: Attack on Olympic Pyeongchang Domain: thlsystems.forfirst.cz IP: null File: null Port: null Protocol: ip CreatedDate: 08-06-2018 UpdatedDate: 10-06-2018 References: https://securingtomorrow.mcafee.com/mcafee-labs/malicious-document-targets-pyeongchang-olympics/	alert udp $HOME_NET any -> any 53 (msg:"test - Cuoc_tan_cong_vao_Opympic_Pyeongchang - DNS request for thlsystems.forfirst.cz"; content:"|01 00 00 01 00 00 00 00 00 00|"; depth:20; offset: 2; content:"|0A|thlsystems|08|forfirst|02|cz"; flow:to_server; fast_pattern:only; nocase; classtype:trojan-activity; reference:url,https://securingtomorrow.mcafee.com/mcafee-labs/malicious-document-targets-pyeongchang-olympics/; sid:15; rev:1;)

Rule-based unknown domain detection: is a unknown domain detection technique based on the set of rules. Multiple open-source libraries, tools are available to compare domain search string, taken from network monitoring tools like Suricata, Snort, or Bro IDS. The message below is an example of a warning to detect C&C traffic in Suricata:



```
10/07/2017-19:39:54.925242 [**] [1:2021716:1] ET
TROJAN Backdoor family PCRat/Gh0st CnC traffic
(OUTBOUND) 102 [**] [Classification: A Network
Trojan was detected] [Priority: 1] {TCP}
172.16.253.132:1163 -> 216.176.190.44:9494
10/07/2017-20:12:55.392557 [**] [1:2013926:8] ET
POLICY HTTP traffic on port 443 (POST) [**]
[Classification: Potentially Bad Traffic] [Priority:
2] {TCP} 192.168.10.20:49716 -> 148.251.255.108:443
```

Fig. 5. An example of detecting unknown domain-based rule sets with the Suricata tool

Unknown domain isn't in neither Domains benign list nor malicious domains list, which is Stage 1 result.

Stage 2: Detect unknown domain with behavior analysis: Machine learning algorithms are used to evaluate the behavior of the unknown domains obtained in the first phase. In this method, two main issues are concerned:

Attribute list: Attribute list is a very important part of determining the abnormal behavior of unknown domains. Research on the detection of unknown malicious domains is various in approach and in using different attributes in order to identify unusual signs of the domains. Some common attributeHHHHHHHHHHHHHHHHHHHHHHHHs are: lexical, link popularity, webpage content, DNS answers, DNS fluxiness, network features, [20]. In paper [10], the authors used 4 main groups of attributes are DNS request and answer-based features, Domain-based features, Time-based features, whois-based features with J48 decision tree algorithm to detect malicious domain. In paper [8], the authors determining unknown domain by using these 3 attribute groups: Domain name lexical features, Ranking features, DNS query features.

Machine learning algorithm: machine-learning algorithms play an important role in labeling objects as normal or abnormal based on attributes and behavior. In this case, Decision Trees, Naïve Bayes Classification; Support Vector Machines (SVM); Random Forests; k-Nearest Neighbor [21] are machine learning algorithms that can detect malicious domains based on suspicious behavior. In a report [8] random forest algorithm is used as the clustering algorithm. In the paper [10], Global Abnormal Forest, KNN are featured.

- **APT Domain detection**

APT domains are extracted from malicious domains list by following steps:

Step 1: In this step, network flows are taken out from the Network Traffic logs by open-source tools (Suricata, Bro IDS). Also, these flows are data stream in a session between 2 end-point devices. Additionally, a Flow's initiated when a device sends a data package, it's discontinued after a defined period of time without any another data package is sent. Recorded data is presented on paper [11, 12, 13], [22].

Step 2: Next, step 1 output is processed by Malicious Domain Detection module. After that, the processed result is monitored by Suricata or Bro IDS.

Step 3: At this stage, suspicious flows are monitored by open-source IDS (Suricata, Snort, etc.). Afterward, machine learning algorithms are used to distinguish between APT attack flows and other scenario flows. For this purpose, 2 properties are taken into account:

Flow behavior: Some abnormal behaviors are defined in the report [23]

Machine Learning algorithm: Decision Trees, Naïve Bayes Classification; SVM; Random Forests; k-Nearest Neighbor [21] is a notable algorithm to detect and categorize APT domains and other attack scenario domains. On the other hand, Deep Learning algorithm such as Neural Network is also viable [24].

The acceptable output of this module is to sort out APT attack domains from other malicious ones.

VI CONCLUSIONS

In this paper, we presented in detail the key issues of APT attack detection model based on correlation analysis techniques including malicious domain detection process and APT domain detection. Accordingly, our APT attack detection system will include many different access layers to optimize the process of finding suspicious signs. Currently, the study and application of correlation analysis model in the problem of detecting APT attacks is an approach suitable to actual needs.

REFERENCES

1. Dr. Eric Code. Advanced Persistent Threat. Understanding the Danger and How to Protect Your Organization. 1st Edition. Amsterdam: Elsevier; 2012.
2. Ping Chen, Lieven Desmet, Christophe Huygens. A study on Advanced Persistent Threats. Communications and Multimedia Security: 15th IFIP TC 6/TC 11 International Conference, CMS 2014, Aveiro, Portugal, September 25-26, 2014. Proceedings, vol. 8735 of Lecture Notes in Computer Science, pp. 63–72, Springer, Berlin, Germany, 2014.
3. Mirco Marchetti, Fabio Pierazzi, Michele Colajanni, Alessandro Guido. Analysis of high volumes of network traffic for Advanced Persistent Threat detection. Computer Networks. 2016: 109; 127–141.
4. APT Detection Indicators – Part 3. https://nigesecurityguy.wordpress.com/2014/04/03/apt-detection-indicators-part-3/. Last accessed 26 Oct 2019
5. APT Anomaly Detection – Part 1. https://nigesecurityguy.wordpress.com/2014/03/19/apt-anomaly-detection-part-1/. Last accessed 26 Oct 2019.
6. R. Sommer and V. Paxson. Outside the Closed World: On Using Machine Learning for Network Intrusion Detection. 2010 IEEE Symposium on Security and Privacy. Oakland, 2010. Pp 305 –316.
7. C. Zhou, C. Leckie and S. Karunasekera. A survey of coordinated attacks an collaborative intrusion detection. Computers & Security. Volume 29, Issue 1. 2010. pages 124-140.
8. Do Xuan Cho; Ha Hai Nam. A Method of Monitoring and Detecting APT Attacks Based on Unknown Domains. Procedia Computer Science, 2019, 150, 316-323.
9. Jiazhong Lu, Kai Chen, Zhongliu Zhuo, XiaoSong Zhang. A temporal correlation and traffic analysis approach for APT attacks detection. Cluster Computing (2017). pp 1–12.
10. Weina Niu, Xiaosong Zhang, GuoWu Yang, Jianan Zhu, Zhongwei Ren. Identifying APT Malware Domain Based on Mobile DNS Logging. Mathematical Problems in Engineering Volume. 2017. pages 1–10.
11. APT & CyberCriminal Campaign Collection. https://github.com/CyberMonitor/APT_CyberCriminal_Campagin_Collections/blob/master/README.md. Last accessed 26 Oct 2019.
12. DeepEnd Research: List of malware pcaps, samples, and indicators for the Library of Malware Traffic Patterns. https://contagiodump.blogspot.com/2013/08/deepend-research-list-of-malware-pcaps.html. Last accessed 26 Oct 2019.
13. APTNotes - Website https://aptnotes.malwareconfig.com/ Targeted. Last accessed 26 Oct 2019.
14. OpenDNS public domain lists of domain names for training/testing classifier. https://github.com/opendns/public-domain-lists. Last accessed 26 Oct 2019
15. Malware Domain List. http://www.malwaredomainlist.com/. Last accessed 26 Oct 2019.
16. Alexa - Top Sites for Countries. https://www.alexa.com/topsites/countries. [Ngày truy cập 1/4/3018]
17. Majestic. https://www.merriam-webster.com/dictionary/majestic. Last accessed 26 Oct 2019.
18. Suricata User Guide. http://suricata.readthedocs.io/en/latest/index.html. Last accessed 26 Oct 2019.
19. Signature Framework. https://www.bro.org/sphinx/frameworks/signatures.html. Last accessed 26 Oct 2019
20. Doyen Sahoo, Chenghao Liu, and Steven C.H. Hoi. Malicious URL Detection using Machine Learning: A Survey. arXiv:1701.07179v2 [cs.LG] 16 Mar 2017.
21. Smola, A.; Vishwanathan, S.V.N. Introduction to Machine Learning, Cambridge University Press, 2008.
22. Intrusion Detection Evaluation Dataset (CICIDS2017). http://www.unb.ca/cic/datasets/ids-2017.html. Last accessed 26 Oct 2019.
23. Elaheh Biglar Beigi, Hossein Hadian Jazi, Natalia Stakhanova and Ali A. Towards Effective Feature Selection in Machine Learning-Based Botnet Detection Approaches. 2014 IEEE Conference on Communications and Network Security. Pp 247- 255.
24. Ang Kun Joo Michael, Emma Valla, Natinael Solomon Neggatu, Andrew W. Moore. Network traffic classification via neural networks. https://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-912.pdf. Last accessed 26 Oct 2019.

AUTHORS PROFILE

Dr. Do Xuan Cho is currently a lecturer at the Faculty of Information Technology at Posts and Telecommunications Institute of Technology in Vietnam
In 2008, received a bachelor's degree in the Saint Petersburg Electrotechnical University "LETI" on a specialty "Computer science and computer facilities", Russia. In 2010, graduated a masters from the Saint Petersburg Electrotechnical University "LETI" on a specialty "Computer science and computer facilities", Russia. In 2013, received a PhD in the Saint Petersburg Electrotechnical University "LETI", on a specialty CAD. Russia. Area of scientific interests - modeling, control systems, algorithmization
Email:chodx@ptit.edu.vn and ***chodx@fe.edu.vn***

Tisenko Victor Nikolaevich, My position is the professor of Institute of computer sciences and technologies in Peter the Great Saint-Petersburg Polytechnic University. I have received the degree Doctor of Technical Sciences in 1998 in accordance of scientific speciality "Systems of automatic Design" in SPbPY. The area of scientific interest is use of new type of fuzzy logics in different applications. I think that we could cooperate intensively in future.Email: ***v_tisenko@mail.ru***

Do Hoang Long, Nguyen Vuong Tuan Hiep, Le Quang Sang, Nguyen Quoc Hoang, are fourth-year students majoring in information security at FPT University. These students have over 2 years of experience working with APT attack detection issues.
Email: longdhse05220@fpt.edu.vn, hiepnvtse05065@fpt.edu.vn,
sanglqse04676@fpt.edu.vn, hoangnqse06012@fpt.edu.vn