International Journal of Pure and Applied Mathematics
Volume 118 No. 20 2018, 3761-3768
ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue

ijpam.eu

Advanced Persistent Threat Analysis using Splunk

Harikrishnan V N,Gireesh Kumar T
TIFAC-CORE in Cyber Security, Amrita School of Engineering, Coimbatore,
Amrita VishwaVidyapeetham, India.
vnharikrishnan@gmail.com, gireeshkumart@gmail.com

Abstract— **Advance persistent threats (APT) are the recently emerged hazard against the modern organization. It is the most challenging attack to detect because of its unique and complex nature. APTs are network specific attacks which span over a long period of time to ex-filtrate sensitive information of various targets. Unlike automated mindless threats, they are skilled, motivated and organized attacks conducted by Cyber Crime Organizations. Present systems use either network analyzer or some host based insider threat detection mechanisms to detect APTs. In this work we analyzeAPTs using splunk-Security Information and Event Management (SIEM). SIEM is the most modern security solution which gave the ability to monitor all the events that are happening inside each system and represent each of the logs in a well-mannered form. 100 of samples of APT logs are compared against the normal event logs to identify the behaviour of an APT. Based on the analysis we developed a set of features that every APT contains. We are also proposing an efficient approach to identify the exfiltration of this APT by using the modern technology called machine learning tool kit.**

Key words: APT, Splunk, Machine learning

I. INTRODUCTION

The malware detection and capturing systems has become an essential component of a secure infrastructure. There are various methods available for security enhancements, one of them is SIEM. The other methods that are used for malware detection includes Firewall, IDS etc., Thus malware capturing and detection has its importance in protecting the network as well as the host machines.Advanced Persistent Threats (APTs) are the most dangerous malwares of modern era. It can be defined as a network attack in which an unauthorized person gains access to a network and stays there undetected for a long period of time.

The main objective of this type of attack is to ex-filtrate data rather than causing damage to the organizations. Since it is looking for specific information on a specific target, it will span over a long period of time using multiple targeting methods, tools and techniques. No matter how long it takes APT will continue to achieve its goal at multi stages by using multiple attack vectors. The existing network monitors are able to detect only the common types of threats, but they are also inefficient to identify APTs because an APT attacker always mimics normal behavior of the users and compromises a set of hosts. With that further communication with the Command and Controlserver will be initiated from the organization itself that eliminate the scope of network analyzers. Present systems use either network analyzer or some host based insider threat detection mechanisms to detect the incompletion. So detecting threats in outbound data is perhaps the best way for an efficient APT detection.

This paper proposes an efficient analysis approach for APTusingsplunk SIEM. It also aims to capture both system and network level logs and create data sets with the sequence of features produced during the analysis of those logs. The existing attack vector or activity pattern is not used in any of the APT's means that all the APT's will be different. Only thing common in every APT is that they contain 5 stages which begins with phishing and social engineering then ends by exporting large volumes of stolen data to the attacker's server. The 5 stages are Reconnaissance Phase,Compromise Phase, Compromise Phase, Lateral Movement Phase and Data Ex-Filtration Phase.In this paper we proposed a framework for APT analysis using splunk Security Information and Event Management (SIEM).The main contribution of this paper is as follows:

- Analyze some of the APT malwares using splunkSIEM.Splunk gave the ability to monitor all the events that are happening inside each system and represent each of the logs in a well-mannered form.
- APT malware features by analyzing normal system event logs and APT event logs.
- Proposed a new approach for predicting the external DNS or malicious sites that actually try to perform malicious activities using Splunk new tool called Machine learning tool kit, which will effectively stop the ex-filtration of data.

II. RELATED WORKS

Roman Jasek, Martin Kolarik, Tomas Vymola[1] are described a system which enhance a security solution to detect APT using honeypots. Their work described the various problems associated with APT and proposed a detection architecture using honeypot. Honeypot is a computer security mechanism set to detect and analyze the Endeavour of attackers on computer systems. Honeypot servers are dedicated servers that attract various attackers and intruders into its simulated network and record their activities. The passive nature of honeypot is a major limitation which helps the attacker to notice it and carried out its activities without detection.Therecomes the scope of honeypot agent. A honeypot agent who directs the attacker directly to the system without indicating the presence of the security measures implemented. Most of the APT attacks include the simulation of user behavior, to detect this type of attempts honeypot system use the agent to set traps which clearly separate the continuous user behavior and the malicious user[7]. Most of the APT attacks are intelligent so they can outperform the installed security measures. In such scenario honeypots offers an additional level of security but not a complete detection approach.

Researchers MircoMarchetti, Fabio Pierazzi, Michele Colajanni and Alessandro Guido[2] recently suggested a method for APT detection by analyzing huge volume of network traffic. The ultimate aim is to find the weakest spots related to ex-filtration and other malicious activities done by APT. The end product list down a ranked set of malicious hosts, which help the security experts to trim down their focus on to a small set of hosts out of thousands of machines in the organization. The selected features of these suspicious hosts are extracted and evaluated over all internal hosts. The level of suspiciousness is evaluated for each host by comparing its past and comparing it with other hosts in the system.

Guodongzhao, KeXu,Lei Xu and Bo Wu [3] proposed a system placed at the network edge whose sole aim is to detect APT malwares based on malicious DNS. It uses a combination of anomaly based detection and signature based detection to find out the malicious APT C&C domains.They developed a reputation engine that compute a reputation score for every IP address based on the features of big data that characterize the properties of malware DNS and based on the features which identify the traffic of compromised clients.It is very difficult to detect APT which relies on DNS to locate command and control servers because APT attack does not use malicious flux service or DGA(Domain Generation Algorithm)domains.A system called IDnS is used to detect the infection which is caused by APTs. This system effectively reduces the volume of network traffic which needs to be recorded and analyzed. IDns consist of four main parts starting from the data collector and ends with the Reputation Engine.The data collector placed at the network boundary collect the inbound and outbound traffic of the network.Malicious DNS Detector will analyze the inbound and outbound traffic for detecting malicious APT domains.A combination of signature-based detector and anomaly-based detector act as the Network Traffic Analyzer which analyze the network traffic of the malicious APT domains.The signature-based detector will look for the known set of malwares where as an anomaly-based detector look for unknown or new set of malwares.At the end Reputation Engine compute a reputation score for an IP address to indicate the level of infection.Malicious traffic features occur in the traffic to a suspicious malicious IP do have very low reputation of normal traffic.

Leyla Bilge, EnginKirda,ChristopherKruegel and Marco Balduzzi introduced a new technique Exposure[4], a system that uses a passive DNS analysis techniques to detect domains that are involved in malicious activities.DNS, domain name server is a mapping of domain name with their numerical identifier or IP.This DNS can be used for various malicious activities,such as Bots takes DNS names to locate their command and control servers, and spam mails contains URLs that resolve to malicious servers.The method uses 15 characters that are filtered from the DNS traffic to characterize different properties of DNS names and the ways they are querying in a typical malicious activity.Whenever an attacker gets the sensitive data and infect the end-user system, the machine will become a bot that communicate to remote server called as botmaster and try to infect other systems also.The bots are then used for stealing sensitive user information,to perform DoS attacks and used to send large numbers of spam messages to achieve financial profit. Every attacker needs a malicious infrastructures, reliable and flexible server infrastructure, and command and control mechanism for

performing any kind malicious activities using bots.For better results, attackers make use of DNS name along with the complex-large-distributed infrastructure. And this makes it difficult to identify the domain name of the malicious server. The idea behind Exposure is that,as malicious services are dependent on Good services of DNS it is able to identify malicious domains which help in mitigating any internet threats that stem from botnets, phishing sites, malware hosting services, etc.They can distinguish between good and malicious domains by observing the behavior which they exhibit.

Igor Anastasov and DancoDavcev[5] suggested a method to implement SIEM in Global and Distributed environments.SIEM used to centralize the log management and to provide proactive threat detection and real-time analysis of system activity to increase the data protection and information security in the organization.SIEM will provide the log handling for logs which generated by many sources, such as antivirus, firewalls,IDS,IPS,O.S, workstations and networking devices and applications. The proposed model is called "Hierarchical Managers Model",because the model is using multiple hierarchical SIEM Managers in the SIEM system Hewlett-Packard– ArcSightESM, ThisArcSight consolidate the log data from various devices and correlate it to raise an alarm in case of attack. The architecture of the ArcSight ESM for the Hierarchical Managers Model comprised of three design layers. The first layer generates the source of logs. Second layer is a collection of servers that are used to collect this source logs from the machine and given to the central server. The central server consolidates these logs and stores those in the log storage. Third layer monitor the stored logs and the servers present in the second layer. But the time taken for each search request is more in the case of ArcSight when it compared to Splunk.

Siva Niranjan and A.R. Vasudevan[6] suggested a method for rule generating in splunk SIEM environment.They used RETE algorithm to formulate the rules and store the event attributes in database. An alert is triggered, when the rule and the attack pattern matches.Splunk monitors various devices for heterogeneous logs such as application logs, windows event logs, system logs, etc. Based on these logs and the trained dataset splunk will generate alarms against the suspicious activities.The important things in splunk are Universal forwarder, Indexed Server Deployment Server and Search Head. Universal forwarder used to forward data from the hosts in the organization to the splunk indexer server. Indexer Server are the databases of splunk,which stores the data forwarded by the universal forwarder. Deployment Server is a technique for distributing configurations,apps and content updates to various splunkinstances.Search heads are used to perform the search management functionality of splunk.

III. Problem Formulation

APT's are the most dangerous type of viruses that most of the organizations are afraid of. Its detection becomes the most important thing in the scope of security world because of its high level of secrecy and stealth approaches. Traditional security solutions are still capable of identifying the known types of malwares and mindless piece of malicious codes. But while coming into APT detection, most of the existing security solutions failed in it because no single attack vector or activity pattern used in it. i.e., no two APTs are same, different combination of attack techniques and ex-filtration methods are done by some well-organized cyber-attack organizations.

Based on the studies and literature survey we decided to analyze some of the APT malwares using splunk SIEM. SIEM is the most modern security solution which gave the ability to monitor all the events that are happening inside each system and represent each of the logs in a well-mannered form. Splunk is the best solution for modern malware detection due to its high speed search head which will give the logs within seconds and the ability to produce alert against suspicious activities in real time. It will monitor various devices for heterogeneous logs such as application logs, event logs, system logs, etc. Based on these logs and the trained data set splunk will generate alarms against the suspicious activities. The important parts of splunk include Universal forwarder, Indexed Server, Deployment Server and Search Head. New add-ons in the splunk like Machine learning tool kit make it even more efficient in the case of security; it provides a wide range of machine-learning methods and techniques for anomaly detection, regression, prediction etc. Selected machine learning algorithm includes linear regression, SVM and Random forest. Linear regression can be used as a machine learning as well as statistical algorithm, since we are focusing on event by event approach linear regression is one of the best method to check the relationship between input and output. SVM is a supervised machine learning which can be used to clearly separate the normal events and malicious events.Random forest is a decision tree based supervised machine learning algorithm which can be used as continuous variable decision tree algorithm for the splunk logs

The Fig. 1 shows the complete system architecture of the APT analysis, which is basically a continuous host monitoring system using splunk.Continuous host monitoring includes the log collector, event aggregator and machine learning tool kit.

Splunk collects all the logs from the connected network devices. Splunk Forwarders installed in all devices and send the logs to the central SIEM server. Splunk monitors various devices for heterogeneous logs such as application logs, windows event logs, system logs, etc. Forwarders are manually installed and configured in the system hosts for forwarding the logs. For collecting these logs we are setting up one indexer in the splunk machine. Forwarders can be installed via installer gui or commandline. Both methods we have to specify the indexer machine ip address and the port which is going to use for forwarding. These universal forwarders forward the logs in the sub system to the indexer present in splunk in real time.Along with this we loaded the splunk with 100 sample logs of apt malwares.These system logs will be evaluated against APT logs in splunk. Machine learning algorithms which is provided by the machine learning toolkit is applied up on these logs to predict the malicious activity. Mainly focusing on the firewall logs and APTlogs.

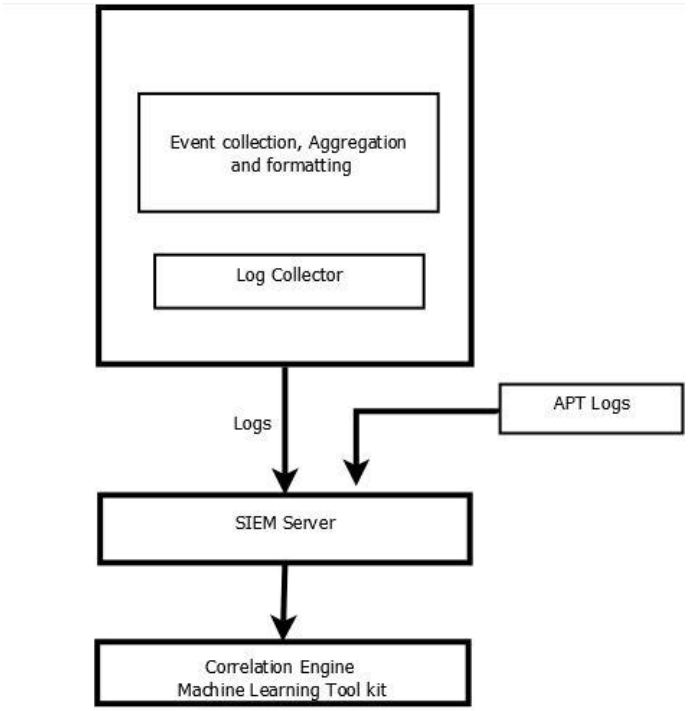

Fig. 1 System Architecture

Splunk SIEM is used for the collection of homogeneous and heterogeneous logs from various devices. The variousmodules in the splunk system are as follows

A. Event Collection and Aggregation Module

Splunk can monitor event log files stored in the local machine. The event log monitor runs as an inputprocessor within the Splunk service. It can monitor 34 different types of events such as Application events, Security events, System events, Setup events, Forwarded events, Internet Explorer events, Key management Service events, etc. The similar events which are generated at different time interval are observed and considered as a single event.

B. Event Formatting Module

Event formatting process helps in converting the logs in different format into a standard CEF format. Here, logs are extracted from Splunk in Comma Separated Value (CSV) format. Then the extracted raw files are normalized with the help of a python CSV reader.

C. Event Correlation Module

Correlating multiple layer events is used to detect the strange behavior of the system.

D. Machine Learning Toolkit

Machine learning tool kit provides a wide range of machine-learning methods and techniques for anomaly detection, regression, prediction etc.

IV. Result and Discussion

The Fig 2.Describes the test bed of APT detection, which actually representing the splunk environment. We are manually installing forwarders in the hosts (sub SIEM in Fig 2) present in the test bed,it includes two windows machine and one Linux machine. Then these forwarders carry logs such as application log, system log, firewall log etc to the central indexer server. Along with this we loaded the splunk with 100 sample logs of apt malwares.

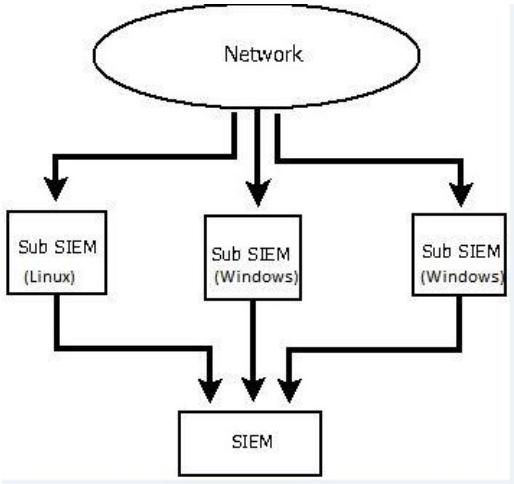

Fig.2 SIEM Test bed for APT analysis

Our analyses were performed upon the logs of the normal activities happening inside the system and the logs of apt events. Analyses of these logs gave some of the attributes that present in every apt activities.They are listed below,

- api_resolve
- file_access
- file_read
- reg_access
- reg_read
- random_logon

Above listed attributesare selected based on the count that they encountered in apt events and normal events. Difference between normal events and apt events are listed in the Table 1

TABLE 1. LOGS OF APT EVENT

Event	api_resolve	file_access	file_read	reg_access	reg_read	random_logon
Apt event1	Yes	Yes	Yes	Yes	Yes	Yes
Normal event1	No	Yes	Yes	Yes	No	no
Apt event2	Yes	Yes	Yes	Yes	Yes	Yes
Normal event2	Yes	Yes	Yes	No	No	no

From the above listed attributes it's clear that an APT activity is running whenever the selected feature are occurring frequently. Our result clearly states that none of the normal events includes all the identified attributes together. But in the case of an APT all these events are performed at different intervals of time in different rounds. Why splunk is the best solution for apt detection is because it can identify these features in real time. Splunk will collect the logs of each activity in real time, based on the count of occurrence of various events we can able to write rules in it for detection. The ultimate aim of every APT is to ex-filtrate data. Identifying this ex-filtration is the hardest part in APT detection. Splunk offers a new tool called Machine learning tool kit, this can be a solution for this.ML tool kit help to predict the external dns or malicious sites that actually try to perform malicious activities. We performed three machine learning algorithms on the firewall logs using splunk machine learning tool kit. Attributes which we concentrated are as follows

- bytes_recevied

- bytes_sent
- packets_recevied
- packets_sent
- src_port
- dest_port
- dst_ip

Algorithms applied upon this are SVM,Linear Regression and Random Forest. Results of these predictions using ML tool kit is shown in the Table 2

TABLE 2 .MACHINELEARNING RESULTS

	Precision	Recall	Accuracy	F1
Linear regression	0.83	0.81	0.81	0.82
SVM	0.96	0.96	0.96	0.96
Random Forest	0.99	0.99	0.99	0.99

From the above results we reached the conclusion that Random Forest has the best result while predicting malwares. So machine learning toolkit is the best solution for identifying the ex-filtration of data.

V. CONCLUSION

The malware capturing and detection system is a very important component of a secure infrastructure. The introduction of advanced malwares such as APT bypasses all the conventional security solutions and became a threat for the organizations and governments. SIEM is one of the best methods among the available security solutions to detect the advanced Malwares, along with other layers of security such as firewalls, IDS and honeypots to implement the concept of defense in depth.This paper proposed an efficient apt analysis using splunk SIEM. From the analysis we obtained some features that will be present in every APT event and a way to identify the ex-filtration using machine learning tool kit. Based on these results we are trying to develop an efficient APT detection approach which incorporates both network traffic analyser and host based internal threat detection in future usingsplunk SIEM.

REFERENCES

1. Jasek, R.O.M.A.N.,M.A.R.T.I.N.Kolarik, and T.O.M.A.S. Vymola. "APT detection system using honeypots." *Proceedings of the 13th International Conference on Applied Informatics and Communications (AIC'13), WSEAS Press*. 2013.
2. Marchetti, Mirco, et al. "Analysis of high volumes of network traffic for Advanced Persistent Threat detection." *Computer Networks* 109 (2016): 127-141.
3. Zhao, Guodong, et al. "Detecting APT malware infections based on malicious DNS and traffic analysis." *IEEE Access* 3 (2015): 1132-1142
4. Bilge, Leyla, et al. "EXPOSURE: Finding Malicious Domains Using Passive DNS Analysis." *Ndss*. 2011
5. Anastasov, Igor, and DancoDavcev. "SIEM implementation for global and distributed environments." *Computer Applications and Information Systems (WCCAIS), 2014 World Congress on*. IEEE, 2014.
6. Raja, M. Siva Niranjan, and A. R. Vasudevan. "Rule Generation for TCP SYN Flood attack in SIEM Environment." *Procedia Computer Science* 115 (2017): 580-587.
7. Ali, P. Dilsheer, and T. Gireesh Kumar. "Malware capturing and detection in dionaea honeypot." *Power and Advanced Computing Technologies (i-PACT), 2017 Innovations in*. IEEE, 2017.
8. "Arcsightesm5.0Software2010."http://www8.hp.com/us/en/softwaresolutions/software.html ?compURI=1340712
9. "Leidos taps splunkitsi for better event management." http://www.splunk.com
10. Chandra, J. Vijaya, NarasimhamChalla, and Sai KiranPasupuleti. "A practical approach to E-mail spam filters to protect data from advanced persistent threat." *Circuit, Power and Computing Technologies (ICCPCT), 2016 International Conference on*. IEEE, 2016.
11. R. J. Rodr´ıguez, X.Chang, X.Li, and K. S. Trivedi, "Survivability analysis of a computer system under an advanced persistent threat attack," in International Workshop on Graphical Models for Security , pp. 134–149, Springer, 2016.

12. Ghafir, M.Hammoudeh, and V. Prenosil, "Disguised executable files in spear- phishing emails: Detecting the point of entry in advanced persistent threat," tech. rep., Peer J Preprints, 2017.
13. "The honeynet project." http://www.honeynet.org/project
14. https://machinelearningmastery.com/linear-regression-for-machine-learning/
15. http://scikit-learn.org/stable/modules/svm.html
16. http://blog.citizennet.com/blog/2012/11/10/random-forests-ensembles-and-performance-metrics