Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 4, April 2015, pg.313 – 316

RESEARCH ARTICLE

Botnet Identification System Using Clustering and Machine Learning C5.0

Ankita Bhaiyya[1], Prof. Sonali Bodkhe[2], Prof. Amit Pimpalkar[3]

[1]Student (M.Tech) CSE, G.H.Raisoni College Academy of Engineering and technology Nagpur, (INDIA)

[2]H.O.D (M.Tech) CSE, G.H.Raisoni College Academy of Engineering and technology Nagpur, (INDIA)

[3]Professor (M.Tech) CSE, G.H.Raisoni College Academy of Engineering and technology Nagpur, (INDIA)

[1] ankita.bhaiyya88@gmail.com; [2] sonali.mahure@gmail.com; [3] amit.pimpalkar@raisoni.net

---

Abstract-- One of the most significant current issues in computer network security is BOTNET. It is an active focus in the research community and industry due to sharp rise of attacks on individual and organizational computers. BOTNET is a massive network of compromised computers used to attack other computer systems for malicious intent. Botnets are one of the most catastrophic threats against the cyber security. Recently, HTTP protocol is frequently utilized by botnets as the Command and Communication (C&C) protocol. In this work, we aim to detect botnet activity based on machine learning approach. To achieve this, we employ. The proposed botnet analysis system is implemented by employing Bisecting K-means clustering and machine learning classifier C5.0.

Keywords--: Botnet Detection, Httpfiltering, and Machine Learning Based Analysis.

I. Introduction

Botnets are and are likely to remain the main vehicle for online crime for the anticipated future. To protect their business models, botnet operators constantly improve their protocols and applications to harden them against detection, analysis and takedown efforts. Our analysis suggests that future botnets will use proper encryption for their protocol messages, rendering them invisible to most deployed network intrusion detection systems. The BOTNET life cycle has several steps from infecting the computer till commanding it for malicious activities. First it transfers malicious code and then automatically executes when computer connects to internet. Supervisor-bot can infected then command and control the infected machines. Botnet program contains server/client code to communicate with the server applications.

There is wide range of the HTTP usage on the Internet, most recent botnets employ HTTP protocol to hide their malicious activities among the normal web traffic. Their C&C channels utilize HTTP protocol to communicate with their bots. Therefore, to investigate the effect of protocol filtering on botnet detection, specifically on false alarm rates, we employed a HTTP filter to select only HTTP related traffic. Citadel and Zeus are the two most powerful botnets that have affected the legitimate Internet realm the most in the past few years.

Machine learning deals with the construction and study of the systems that can learn from data rather than follow only explicitly programmed instructions. It has strong artificial intelligence and optimization. It employs in a range of a computing task.

II. RELATED WORK

Khaleel ahmed and M.A rizvi describes that RSA Algorithm is an effective algorithm which is most secure and less time consuming[1]. It will control all the encryption and the decryption with the help of private and public key of sender and the receiver. But encryption key size is 1024 bit so it required more memory and computing power and more battery.

Jonathan P. Chapman and Felix Govaers used the k means bisecting algorithm for clustering the data[10].first there is a raw data generated on the log than that raw data is gathered through k means bisecting algorithm which convert the data into filtered data according to their relevancy.But its is less accurate it just based on estimation.

WangYang Yu, ChunGang Yan describes a new model called an EBPN and a systematic approach to modeling and validation of ecommerce business processes[7]. The modeling is done by composing control flow and data flow models to build a complete EBPN and can reduce the probability of errors in a modeling process. The proposed model has to be validated for two main properties: rationality and transaction consistency. In this paper, rationality guarantees the structural correctness of an EBPN, and transaction consistency guarantees the transaction properties and protects the interests of trading parties.

Roberto Perdisci, Wenke Lee describes novel botnet detection system that is able to identify *stealthy* P2P botnets, whose malicious activities may not be observable[8]. To accomplish this task, we derive *statistical fingerprint*s of the P2P communications to first detect P2P clients and further distinguish between those that are part of legitimate P2P networks and P2P bots. The parallelized computation with bounded complexity makes scalability a built-in feature of our system. Extensive evaluation has demonstrated both high detection accuracy and great extensibility of the proposed system.

Uma Maheswari and Dr. P.SumathiA data preprocessing treatment system for web usage mining has been analyzed and implemented for log data[5]. It has undergone various steps such as data cleaning, user identification, session identification and clustering. Dissimilar from usual implementations records are cleaned effectively by removing robot entries. This preprocessing step is used to give a reliable input for data mining tasks. Web personalization method and introduce a successful clustering technique using belief function based on Dempster-Shafer's theory. Perfect input can be created when the byte rate of each and every record is found. This algorithm lacks in scalability problem.

III. Observation for botnet detection Techniques

In this section, we are describing what are the measures to detect the botnet. Botnet detection technique used to identify the Botnet activities. Botnet detection technique mainly divided into two approaches which are honeynet-based and Intrusion Detection System (IDS) based.

The earlier informal studies about the Botnet attack is based on setting up honeynet. Most of researchers setting up honeynet to analyze bots, learn tools, tactics and motives of botmaster. However, honeynet is only good for understanding Botnet characteristic and technology but cannot detect bot infection all the times. This situation make the researchers turned to IDS techniques that more useful to identify the existence of Botnet.

Anomaly-based detection technique is a part of behavior based detection. The anomaly-based is divided into DNS based, data mining-based, host-based and network-based.

This techniques attempt to detect Botnet based on several network traffic anomalies such as high network latency, high volumes of traffic, traffic on unusual ports and unusual system behaviour that could indicate presence of malicious bots in the network. Means, it have focuses on normal behaviour to overcome undetected unknown attack.

The DNS-based detection technique has been done by doing the DNS monitoring and DNS traffic anomalies. In order to make this technique successful, it demands for the DNS information that generated by a Botnet [15]. Usually, bots send DNS queries to access bot servers. It is helpful as bot used DNS to find the address of botmaster. At once, the carry out of DNS queries will help to locate in particular bot server.

The data mining-based detection techniques was proposed to improve the accuracy. It is one effective technique for Botnet detection since it can be used efficiently to detect Botnet C&C traffic by using machine learning, classification and clustering approach

The host-based approach will monitor the network traffic for indications of bot-infected machines. The host become worse when bot had been activated lead the changes on system registry and system files. Then, the Botnet makes a series of systems and library calls.

Meanwhile, the network-based approach more focus on monitoring network traffic in; (i) detection of individuals bots by checking for traffic patterns or content that can reveal the command and control (C&C) server or malicious in bot-related activities, and (ii) analyzing the traffic that indicate two or more hosts behave similar patterns as bot to react in the same function. Monitoring in network-based can be done either in active or passive mode.

Similarly to anomaly-based techniques, signature-based detection technique also as a part of behaviour-based detection. This techniques learn and gain knowledge of useful signatures or behaviours from existing Botnet. This solution is useful for detection on known Botnet accurately rather than the unknown bots. In addition, signature-based can make immediate detection and impossibility of false positive. It require less amount of system resource to make the detection.

In hybrid-based detection technique, two or more IDS techniques were combined. It can be the combination of DNS-based with anomaly-based, signature-based with anomaly-based or data mining-based with anomaly-based technique. Due to signature-based, DNS-based and data mining-based that have same capability where it is only able to detect known attack but cannot detect unknown attack. Instead, anomaly-based has this extra capabilities to detect unknown attack compare to other technique. Based on analysis by, the combination of IDS technique will complement each other weaknesses.

IV. PROPOSED SYSTEM WORK

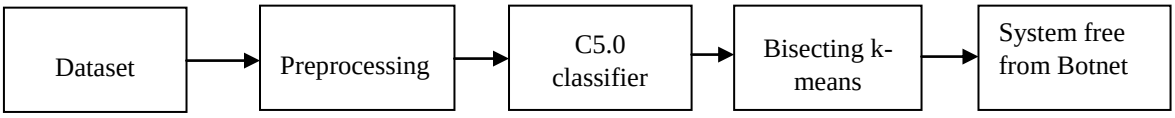

Fig1: Flow of work

In this model we are creating the weblogs. Weblog will be nothing but the entries which will be generated from the machine itself, which may or may not be having the botnets. If botnets are not found then we can use KDD cup for the logs.

The data cleaning main process is removal of outliers or irrelevant data. Analyzing the huge amounts of records in server logs is a cumbersome activity. Therefore initial cleaning is necessary. Botmining is done by the different methods. Botmaster will attack on the group of computers and miner will prevent the system by the attackers.

After the preprocessing the C5.0 algorithm is applied from which the initial botnets are found out. After that Bisecting K-means is applied from which the outlier and actual botnets are identified from the system.

V. CONCLUSIONS

In this study, we have reviewed and summarized the different approaches for existing botnet detection techniques. This will contribute ideas in development of a new Botnet detection technique by finding the gap between these existing Botnet detection techniques.

By applying the C5.0 machine learning classifier on the weblogs generated from machine itself so the filtered packets are generated than by using the clustering algorithm botnet are identified from the system.

REFERENCES

[1] Khaleel Ahmad, M.A. Rizvi Zhiyuan Chen." E-commerce Security Through Asymmetric Key Algorithm". Fourth International Conference on Communication Systems and Network Technologies, 2014

[2] K. Rieck, G. Schwenk, T. Limmer, T. Holz, and P. Laskov, "Botzilla:Detecting the 'phoning home' of malicious software," in Proceedings of the 2010 ACM Symposium on Applied Computing, 2012

[3] W.T. Strayer, D. Lapsely, R. Walsh, and C. Livadas, "Botnet detection based on network behavior," *Advances in Information Security*, vol. 36, pp. 1-24, 2008.

[4] Softflowd. [Online]. http://www.mindrot.org/projects/softflowd/

[5] Uma Maheswari, Dr. P.Sumathi, A New Clustering and Preprocessing for Web Log Mining, World Congress on Computing and Communication Technologies, page-no: 25-29 , IEEE, 2014

[6] G. Gu, R. Perdisci, J. Zhang, and W. Lee, "BotMiner: clustering analysis of network traffic for protocol- and structure-independent botnet detection," in 17th USNIX Security symposium, pp. 139-154, 2008.

[7] WangYang Yu, ChunGang Yan ." Modeling and Validating E-Commerce Business Process Based on Petri Nets". In proceedings of Systems, Man, and Cybernetics: Systems, IEEE Transactions on (Volume:44 , Issue: 3), page-no: 327 – 341, June2013.

[8] Perdisci, R. Wenke Lee." Building a Scalable System for Stealthy P2P-Botnet Detection". In proceedings of Information Forensics and Security, IEEE Transactions on Volume:9 , pages27 - 38,Nov 13.

[9] W.T. Strayer, D. Lapsely, R. Walsh, and C. Livadas, "Botnet detection based on network behavior," *Advances in Information Security*, vol. 36, pp. 1-24, 2008.

[10] Jonathan P. Chapman, Felix Govaers." Network Traffic Characteristics for Detecting Future Botnets". Communication and Information System Conference, Germany, Oct 2012.

[11] NfDump. [Online]. http://nfdump.sourceforge.net/.

[12] E. Alpaydin, Introduction to Machine Learning.: MIT Press, 2004..

[13] T. Karagiannis, K Papagiannaki, and M. Faloutsos, "BLINC: Multilevel traffic classification in the dark ," in Proceedings of the 2005 ACM Conference on Applications , Technologies, Architectures, and Protocols for Computer Communications, 2005.

[14] L. Bernaille, R. Teixeira, and K. Salamatian, "Early application identification," in Proceeding of the 2006 ACM Conference on emerging Networking eXperiments and Technologies, 2006.

.[15] P. Wurzinger, L. Bilge, Th. Holz, J. Goebel, Ch. Kruegel, and E. Kirda, "Automatically generating models for botnet detection," in 14th European conference on research in computer security (ESORICS) , pp.232-249, 2009.

[16]Paul Royal. Maliciousness in Top-ranked Alexa Domains. [Online]. https://www.barracudanetworks.com/blogs/labsblog?bid=2438.

