Advances in Computer Science and Information Technology (ACSIT)
Print ISSN: 2393-9907; Online ISSN: 2393-9915; Volume 2, Number 2; January-March, 2015 pp. 102-106
© Krishi Sanskriti Publications
http://www.krishisanskriti.org/acsit.html

A Survey of Botclouds: Botnet Detection using MapReduce and Big Data Analytics

Sushil Buriya[1], D.S. Bhilare[2] and Arun Singh[3]

[1]School of Computer Science & IT, DAVV Indore, M.P., India
[2]School of Computer Science & IT, DAVV Indore, M.P., India
[3]CIT, UPES, Dehradun Uttarakhand, India
E-mail: [1]sushil.buriya@gmail.com, [2]bhilare@hotmail.com, [3]arunsingh2006@gmail.com

Abstract—*In today's era of Cloud Computing and Big Data, anyone can rent computing and storage power of any size. Security research on the public clouds primarily focuses on the legitimate use of cloud power to fulfill the computing needs of organizations and provide them confidentiality, integrity and availability of data transferred, processed and stored in cloud. Little attention has been paid to use of cloud computing power and fast provisioning to turn it into an attack support by malicious users. Botnets are greatest beneficiaries of this malicious use of cloud computing. For cloud service providers, preventing their cloud infrastructure to being turned into botcloud is very challenging. In this paper, we present an overall overview and analysis of the traditional defense methods against botnets. We also analyze the botnet detection methods using machine learning algorithms for large scale analysis of Netflow data without deep packet inspection. We also present the benefits of using Big Data Analytics for botclouds detection.*

1. INTRODUCTION

The term "moving to cloud" also refers to an organization moving away from a traditional CAPEX model (buy the dedicated hardware and depreciate it over a period of time) to the OPEX model (use a shared cloud infrastructure and pay as one uses it) [1]. According to NIST definition, Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model is composed of five essential characteristics, three service models, and four deployment models [2].

Public cloud infrastructure is provided for open use by general public and exists on the premises of cloud service provider. Public Cloud service providers provide pay-as-you-go computing and storage services, which means anyone can enjoy few hours of supercomputing power [3]. The present availability of high-capacity networks, low-cost computers and storage devices as well as the widespread adoption of hardware virtualization, service-oriented architecture, and autonomic and utility computing have led to a growth in cloud computing. Cloud Service Providers are experiencing growth rates of 50% per annum [1].

Cloud Computing have various security related issues characterized by confidentiality, integrity, availability of data, guest-to-clouds threats, etc [4]. A malicious consumer who uses easily available cloud infrastructure and attacks on other networks or clouds is also a threat for cloud service providers that may ruin the image of CSP. Botnet is a strong technique to launch an attack to other networks. Use of cloud services to implement botnet is known as botcloud. Botclouds are serious threat for Cloud Service Providers [5].

Botnets are built on the very premise of extending the attacker's control over his victims. To achieve long-term control, a bot must be stealthy during every part of its lifecycle. The typical life cycle of a bot has five steps: (i.) Creation, (ii.) Infection, (iii.) Rallying, (iv.) Waiting and (v.) Executing. Once a bot is in place, the only required traffic consists of incoming commands and outgoing responses, constituting the botnet's command and control (C & C) channel. There is two types of topologies to be used by botmasters to build botnets. Figure-1 shows the basic C&C architecture of botnet. Centralized botnets use a single entity (a host or a small collection of hosts) to manage all bot members. Another is P2P model, no centralized server exists, and all member nodes are equally responsible for passing on traffic [6].

The rest of the paper is organized as follows. In the next section literature review of traditional botnet methods, botnet detection using machine learning algorithms and Big Data analytics are presented. In section 3 comparative studies of these techniques with benefits of big data analytics for botclouds detection is described. Section 4 will conclude the paper and section 5 points to future work.

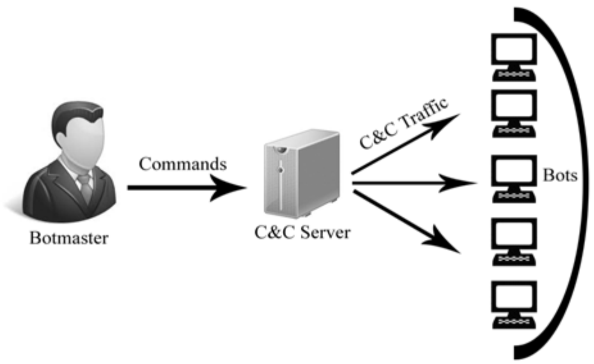

Fig. 1: C&C architecture of botnet

2. LITERATURE REVIEW

In this section we present a brief overview of the existing work related to botclouds. These works are traditional botnet detection methods, botnet detection using machine learning, botclouds and big data analytics for network security.

2.1 Botnet Detection Methods

C&C detection approach is effective because it detects bots before they engage in harmful malicious activities. C&C channels are the medium through which a botmaster commands the bots in botnet to execute attacks like identity theft, phishing, click fraud, and espionage etc. Basil *et al.* presented a C&C detection approach by extracting count-feature sequences (e.g. packet count) from the network traffic and then evaluated the periodogram by applying Walker's large sample test to detect high periodic component. The benefit of approach is that it does not require a signature of botnet behavior. The limitation of method is that it does not effective for botnets with aperiodic behavior [8].

BotGAD (Botnet Group Activity Detector) presents a botnet detection model that reveal both unknown domain name C&C server and IP address of hidden infected hosts. It defines group activity as a key feature of botnets and provides a metric to measure the feature. BotGAD focuses on behavior property of botnet. It measures a similarity coefficient to check a uniformity of groups to detect botnet groups. If bots avoid using DNS, BotGAD cannot detect the bots [9].

Seiichiro Mizoguchi *et al.* proposed the bot detection method which focuses on communication patterns of bots and normal IRC clients. Bots are preprogrammed software, so their communication patterns have mechanical features. This approach uses data transmission intervals to detect the difference between normal clients and bots using clustering analysis. Advantage of this method is that it does not inspect the payloads in the packets. This method is limited to IRC bots and does not work for other type of bots using other protocols [10].

BotHunter implements an event based bot detection engine. BotHunter uses five infection events, such as Inboud Scan, Inboud Infection, Egg Download, C&C communication and Outbound Scan [11]. Many of BotHunter events are triggered using deep packet inspection, which becomes blind with encryption. Han Zhang *et al.* demonstrated a method using BotHunter with a entropy detector to enhance the ability of BotHunter for encrypted bots. It is a complex method because it first calculates the entropy of all data and then differentiates low and high entropy traffic using a threshold entropy value after this use BotHunter to detect the bots [12].

2.2 Botnet Detection using Machine Learning

Florian Tegeler *et al.* presented a system BotFinder that detects infected hosts in a network using high-level properties of the botnet traffic without content analysis. It uses machine learning to identify the key features of C&C communication. It extracts five statistical features from NetFlow traces and creates a model using clustering algorithm for known botnet dataset. The models are used to detect the infected hosts in actual detection phase. It works without deep inspection of packets, so it has capability to investigate encrypted traffic also. If a botmaster uses randomization techniques for C&C communication, it degrades the detection quality of BotFinder. Another limitation of BotFinder is high fluctuation of C&C servers IPs, it also degrade the detection quality of BotFinder [13].

Disclosure is a large scale, wide area botnet detection system that uses NetFlow data to detect the C&C servers of botnet. It reliably distinguishes C&C channels from actual traffic using NetFlow records: (i) flow sizes, (ii) client access patterns, and (iii) temporal behavior. It uses random forest classifier algorithm to build detection models. Authors demonstrate that Disclosure is able to perform real-time detection of botnet C&C channels over data sets on the order of billions of flows per day. Randomization in communication patterns of bots with C&C servers degrades the performance of Disclosure [14].

Fariba Haddadi *et al.* employed two machine learning algorithms, namely C4.5 decision tree and symbolic bid-based (SBB), to generate botnet detection automatically. Two different feature sets are analyzed to check the performance of both machine learning algorithms for different botnet behaviors. Result of analysis describes that SBB performed better than C4.5 in term of the solution complexity [15].

Sherif Saad *et al.* propose an approach for characterizing and detecting using network traffic behaviors. The approach focus on P2P bots during C&C phases (Waiting). Authors extract and analyze a set of features using five machine learning techniques, namely, Support Vector Machine (SVM), Artificial Neural Network (ANN), Nearest Neighbors Classifier (NNC), Gaussian Based Classifier (GBC), and Naïve Based Classifier

(NBC). 17 features that can be extracted from network flows and host communication patterns analyzed using 5 machine learning techniques. Four metrics are used to evaluate each machine learning techniques. The SVM, ANN, and NNC are top three machine learning techniques that can be used to build a botnet detection framework. None of these techniques satisfy the requirement of online botnet detection framework [16].

2.3 Big Data Analytics and Botclouds

Hammi Badis *et al.* have addressed botcloud issue by tackling the characterization of the DDoS attack problem in a public cloud. An experiment performed to understand the operational behavior of a botcloud. System metrics are collected for all phases of botnet. These system metrics are CPU, Memory, Bandwidth Send and Bandwidth Received. The result of experiment show that in DDoS attack, the CPU activity of a botcloud is very low, the memory consumption and the I/O activity are clearly by a bi-modal distribution. Research study concludes that the botcloud activity for idle and active state is clearly identifiable [17].

Jerome Francois *et al.* describe a scalable method for detecting P2P botnets regarding the relationships between hosts. Dependency graph algorithm is applied to NetFlow records that work as a input for PageRank algorithm to Fig. out hosts which are well interconnected such as within a P2P network. The PageRank algorithm is executed using Hadoop for distribution of adjacency matrix of dependency graph among all the data nodes [18].

Big data refers to information that can’t be processed or analyzed using tradition tools or processes. Three characteristics define Big Data: volume, variety, and velocity. Volume of data being stored is exploding. Variety represents all types of data. Velocity describes the speed of data processing. Increased speed of data processing also increases the rate of data generation. The Data which have all 3Vs characteristics is known as Big Data [19]. Big Data Analytics is the process of applying advanced analytical techniques to large datasets to uncover hidden patterns, unknown correlation and useful information. The advanced analytical techniques are predictive analysis, data mining, machine learning, statistical analysis, artificial intelligence and natural language processing [20].

Big Data technologies can be divided into two groups: batch processing and stream processing. Batch processing are analytics on data at rest and stream processing are analytics on data in motion. Hadoop is most popular technologies for batch processing. Pig, Hive and RHadoop are tools to create complex queries and run machine learning algorithms on top of Hadoop. Stream processing does not have single dominant technology like Hadoop. Particular Models for stream processing are InfoSphere Streams, Jubatus, and Storm [21].

3. COMPARATIVE STUDY

We have reviewed various research studies related to botnet detection methods, botnet detection using machine learning, botcloud and big data analytics in section 2. Now we present a comparative study of various techniques in this section.

3.1 Comparative Study of Botnet Detection Methods

Research studies pose several mechanisms for botnet detection. Botnet detection has been classified into bot detection, C&C detection and botmaster detection. Most previous literature refers to single infected machine as bot detection. Another approach is detection of C&C channels of botnet can help in understanding botnet behavior. Botmaster detection is very complicated task, So many botnet detection techniques do not target botmaster. Botmaster detection is a very complicated mechanisms and bot detection approaches only feasible for single bot detection, So C&C detection is an important aspect of botnet detection [7].

Table-1 shows a comparative study of different botnet detection methods. Botnet detection techniques use various types of network traffic data to analyze network traffic on particular detection parameters. We studied 4 botnet detection techniques. All techniques are based on bot detection. Benefits and limitations of these techniques are described.

Our aim is to use an efficient botnet technique in cloud environment to detect botclouds. Botnet using cloud have rapid scalable power of cloud computing. Botmaster may increase the number of bots any time and it is not a predictable number. So botnet detection using the infected host mechanism has a big challenge. Botmasters use C&C servers to give command to bots through C&C traffics. Detection of C&C server is equivalent to detection of all bots communicating with that C&C server. This approach paralyzes the all bots related to that particular server.

Table 1: Comparison of botnet detection methods

Method	Type of Traffic	Detection Class	Benefits	Limitations	Detection Parameters
Using C&C Traffic [8]	IP Packet Traffic	Bot Detection	Does not require a signature of botnet behavior	Does not effective for botnets with aperiodic behavior	Uses periodogram feature of periodic behavior
BotGAD [9]	DNS Traffic	Bot Detection	Enable to detects unknown botnets	If bots avoid using DNS, BotGAD cannot detect the bots	Uses group activity of botnets
Based on DTI [10]	IRC Traffic	Bot Detection	Does not inspect the payloads in the packets.	Limited to IRC bots	Uses data transmission intervals

BotHunter for Encrypted Traffic [12]	Encrypted Packets	Bot Detection	Ability to detect encryptted bots.	Complex method	Uses entropy detectors

3.2 Comparative Study of Machine Learnig Algorithms for Botnet Detection

Machine learning is a scientific discipline that describes the construction and study of algorithms that use data for learning and prepare a model based on inputs to make predictions or decisions. Machine learning algorithms have crucial role in classification of botnet C&C traffic from other benign network traffic. Table-2 explains a comparative study of 4 botnet detection approaches using machine learning algorithms.

Network flow is a sequence of packets from a source to a destination. Most of botnet detection methods using MLAs use network flow data. This analysis does not use the packet payload information and do not violet privacy of user's data. User's privacy is a big issue in cloud computing. So, network flow data is very useful for C&C traffic analysis of botnet without violating the privacy of users.

Table 2: Comparison of botnet detection techniques using machine learning

Method	Type of Traffic	Machine Learning Algorithm	Features	Benefits	Limitations
Bot-Finder [13]	NetFlow Traffic	Clustering	Statistical features of traces	Works without deep inspection of packets	Degrades quality for randomized and fluctuated C&C traffic
Disclosu-re [14]	NetFlow Traffic	Random forest classifier	Flow size, client access patterns, temporal behavior	Large scale analysis without deep packet inspection	Degrade performance for randomized C&C traffic
Using GP & C4.5 [15]	HTTP Traffic	C4.5 and SBB	8 Numeric attributes and 6 flag attributes of network flows	Generates botnet detection models automatic-ally	Botnets do not employ flag attributes
P2P through NBA [16]	NetFlow Traffic	SVM, ANN, NNC, GBC, NBC	Flow based and host based	Detection of botnets before execution phase	Does not support online botnet detection

All the botnet detection techniques using MLAs mentioned in table-2 follow a sequence of steps. First is feature extraction and then feature analysis using a machine learning algorithm. Efficiency of technique depends on both extracted feature and machine learning algorithm that used for analysis. Feature is a statistical measure related to network flow. All the techniques have two phases: one is training phase in which a known C&C pattern is used to prepare a model and another is actual detection phase in which unknown flow traffic analyzed to classify. Support Vector Machine (SVM), Artificial Neural Network (ANN), Nearest Neighbors Classifier (NNC), and symbolic bid-based (SBB) perform well for botnet detection.

3.3 Advantage of Big Data Analytics for Botcloud Detection

Botmaster purchases infrastructures as a service of public cloud and build a botnet by installing bot software in various virtual machines of one or more tenants. Attack is performed to other networks through C&C servers via internet. Fig. -2 shows a basic architecture of botcloud. Traditional botnets require long time to scale up their network but botclouds use rapid scalabilty and distributed computing advantage of cloud computing. So we need a rapidly scalable and distributed solution for botcloud detection before its execution phase.

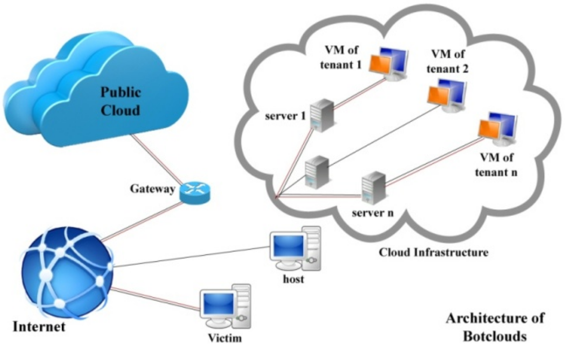

Fig. 2: Architecture of botcloud

New Big Data technologies, such as batch processing and stream processing, are enabling the storage and analysis of large heterogeneous data sets at an unprecedented scale and speed. These technologies have potential to detect botclouds through proactively monitoring of public clouds. Big Data Analytics have following advantages:

(a) Collecting network flow at a large scale.

(b) Performing deeper analytics on the network flow data.

(c) Providing a consolidated view of classification results.

(d) Achieving real-time proactive monitoring of public clouds.

4. CONCLUSION

Proactive monitoring is necessary to detect botclouds. We have presented a comparative study of traditional botnet detection techniques. This study precise that Detection of C&C server is equivalent to detection of all bots communicating with that C&C server. This approach paralyzes the all bots related to that particular server. We have also analyzed the various botnet detection techniques using machine learning algorithms. These techniques analyze network flow data. User's privacy is a big issue in cloud computing. So, network flow data is very useful for C&C traffic analysis of botnet without violating the privacy of users. Feature is a statistical measure related to network flow. Efficiency of technique depends on both extracted feature and machine learning algorithm that used for analysis. Finally, we have described the benefits of Big Data Analytics for botcloud detection.

5. FUTURE WORK

Cloud Service Providers need a proactive monitoring approach to detect botcloud activity in the cloud premises without disturbing service and violating privacy constraints of legitimate consumer of public cloud. For future work, we plan to design a model for botcloud detection by proactively monitoring the network flow data without violating the privacy of packets for CSPs.

REFERENCES

[1] http://en.wikipedia.org/wiki/Cloud_computing

[2] Peter Mell and Timothy Grance, "*The NIST Definition of Cloud Computing*", in *NIST Special Publication 800-145,* September 2011, Gaithersburg, USA.

[3] http://www.qubole.com/big-data-cloud-database-computing/

[4] Huiming Yu *et al*., "Cloud Computing and Security Challenges", in ACM SE, March-2012, Tuscaloosa, AL, USA.

[5] Kassidy Clark *et al*., "Bot Clouds", Nov. – 2010, Netherlands.

[6] John R. Vacca, "The Botnet Problem" in *Computer and Information Security Handbook*, Burlington, MA 01803, USA: MKP, 2009, ch. 8, pp. 119-128.

[7] Sheharbano Khattak *et al*., "A Taxonomy of Botnet Behavior, Detection and Defence", in IEEE communications surveys & tutorials, vol. 16, no. 2, second quarter 2014.

[8] Basil AsSadhan *et al*., "Detecting Botnets using Command and Control Traffic", in *Eighth IEEE International Symposium on Network Computing and Applications*, MA, USA,2009.

[9] Hyunsang Choi *et al*., "BotGAD: Detecting Botnets by Capturing Group Activities in Network Traffic", in *COMSWARE*, Dublin, Ireland, 2009.

[10] Seiichiro Mizoguchi *et al.*, "Implementation and Evaluation of Bot Detection Scheme based on Data Transmission Intervals", Fukuoka, Japan, 2010.

[11] G. Gu *et al.,* "BotHunter: Detecting malware infection through ids-driven dialog correlation" in Proceedings of the 16th USENIX Security Symposium (Security'07), August 200.

[12] Han Zhang *et al.,* "Detecting Encrypted Botnet Traffic", in 16th IEEE Global Internet Symposium, Turin, Italy, 2013.

[13] Florian Tegeler *et al*., "*BotFinder: Finding Bots in Network Traffic Without Deep Packet Inspection"*, in *Co-NEXT'12,* December 10-13, 2012, Nice, France.

[14] Leyla Bilge *et al*., "DISCLOSURE: Detecting Botnet Command and Control Servers Through Large-Scale NetFlow Analysis", in *ACSAC* '12 Dec. 3-7, 2012, Orlando, Florida, USA.

[15] Fariba Haddadi *et al*., "*On Botnet Behavior Analysis using GP and C4.5"*, in *GECCO'14,* July 12–16, Vancouver, BC, Canada, 2014.

[16] Sherif Saad *et al.,* "Detecting P2P Botnets through Network Behavior Analysis and Machine Learning", in Ninth Annual International Conference on Privacy, Security and Trust, Montreal, QC, Canada, 2011.

[17] Hammi Badis *et al.*, "Understanding Botclouds from a System Perspective: a Principal Component Analysis", Troyes Cedex, France, 2014.

[18] Jerome Francois *et al.*, "BotCloud: Detecting Botnets Using MapReduce" in WIFS'2011, Iguacu, Brazil, 2011.

[19] Paul C. Zikopoulos *et al*., "*Understanding Big Data Analytics for Enterprise Class Hadoop and Streaming Data*", New York Oxford University Press, 2013.

[20] Gaurav Dua *et al*., "Big Data: The Next Big Thing", NASSCOM, New Delhi, India, Tech. Rep., 2012.

[21] Alvaro A. Cárdenas *et al*., "Big Data Analytics for Security Intelligence", CSA, Las Vegas, US, Tech. Rep., Sept. – 2013.