Detecting P2P Botnets Network Traffic Behaviors Using Feature-Based Learning Techniques

Aye Aye Thu, Khin Than Mya
University of Computer Studies (Yangon), Myanmar
suchiq13@gmail.com, khinthanmya@gmail.com

Abstract

Botnets have become one of the major threats on the Internet. They are used to generate spam, carry out DDOS (Distributed Denial of Service) attacks and click-fraud, and steal sensitive information. Nowadays, many researchers interest to analyze the botnet technology and emphasis the botnet behaviors. It is needed to classify communication network traffic which is important fact to study the botnet behaviors. In this paper, we proposed an approach to detect botnet activity by analyzing and classifying network traffic behaviors due to P2P (Peer to Peer) based botnets. This system represents the important and most challenging types of botnet currently available that based on classifying P2P botnets. The classification techniques used in detection framework are RF (Random Forest) and SVM (Support Vector Machine). The performance evaluation of the two popular classification techniques is also presented. According to the experiments, proposed system has promising accuracy even with small time window by comparing two machine learning algorithms.

Keywords: *Botnets, Machine Learning, HTTP, IRC, P2P, Waledac, Storm, RF, SVM.*

1. Introduction

Computer network have become essential part of human life because of development of information and communication technologies. Online shopping, e-banking and stock trading became very useful applications for human. In the network environment, many hackers try to use various methods to steal sensitive information.

A botnet is a collection of software agent or robots that run autonomously and automatically [1]. Basically, the composition of a botnet includes: the server programs used to control the infected computers, the client programs installed on the infected computers waiting for the control instructions, and the malicious software to infect computers to become zombie computers. Botnet also have a variety of type, including IRC based, HTTP based and P2P based bots.

The first botnet appeared in 1993 in the Internet Relay Chat (IRC) networks, and become popular after 1999. In New Zealand, a 19-year-old hacker controlled 150 million computers through the Internet, which is the largest well known botnet and also another Chinese hackers controlled 60,000 computers to attack a music website which is causing the website out of service even though its server is being transferred to Taiwan or the USA. Finally, the two hackers were arrested. IRC and HTTP based botnets are vulnerable because they are based on centralized architecture.

In a P2P botnets, any zombie computer can be a client or a server, and it connect to the botnet according to the peer list .Therefore, a P2P botnets doesn't need any particular server to download programs or receive instructions, and the hackers can launch attacks from any computers in the P2P botnets.

The objectives of this paper is to establish a P2P botnets detection system to identify abnormal traffic behavior by applying feature-based classification technique. It can solve the problem of high false positive rate in anomaly-based detection. It is find evidence of P2P botnets activity by monitoring passive network traffic and by using data mining techniques. The systems also classify malicious (botnet) and non-malicious traffic.

The structure of the paper is as follows: Section 2 discusses the concept of botnet architecture; Section 3 explains an overview of previous approaches on detecting botnets; Section 4 describes P2P botnets and two machine learning algorithm which use at our system; Section 5 provides classification framework based on network traffic behaviors; Section 6 evaluates proposed framework using existing experimental datasets and by comparing the performance obtained with two data mining techniques and finally Section 7 covers the conclusion remarks and future work.

2. Botnet Architecture

Bot is a new type of malware installed into a compromised computer which can Command and Control (C &C) remotely by botmaster. After the bot

code has been installed into the compromised computers, the computer becomes a Bot or Zombie. Botnets are networks consisting of large number of Bots. Botnets are created by botmaster (a person or a group) to use for malicious activities such as distributed Denial-of-service (DDoS), sending large amount of SPAM mails and Trojans. According to the Command and Control (C &C) channel, we categorized botnet topologies into two different models, the Centralized model and the Decentralized model.

IRC and HTTP based botnets have centralized architecture at their Command and Control infrastructure in Figure (1). The centralized mechanism of botnet has made them vulnerable to being detect and disabled. If it has a single point of failure the C&C server, all bots will lose contact with their bot master. Currently, botnet technology trend lead to decentralized nature and P2P botnets based on decentralized architecture in Figure (2). P2P botnets have no centralized server and bot are connected to each other which act as both C&C server and client.

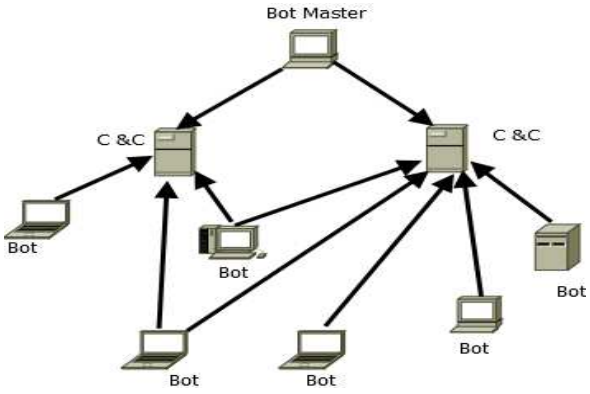

Figure 1. Centralized architecture

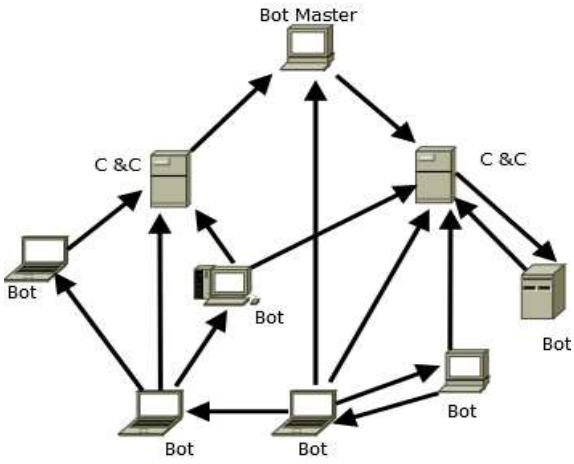

Figure 2. Decentralized architecture

3. Related Work

Many different approaches have been proposed for detection of botnet. There have two approaches for botnet detection. One approach is based on locating honeynet in the network. Another approach is monitoring and analysis of passive network traffic [15].

M.ARajab, E.Cooke.F, C.Schiller, J.Binkley, K. K. R. Choo et al. [12, 5, 4, 9] described how to apply honeynets for botnet detection. Honeynets are functional to understand Botnet characteristics and technology, but cannot detect bot infection all the times.

In [13], M. Rosech presented a signature-based technique uses its knowledge of known malicious characteristics to generate pre-specified signatures, and any execution sequence matching with a signature is flagged as anomalous. Snort, a widely used open source network intrusion prevention system (NIPS) and network intrusion detection system (NIDS), has an ability to perform real-time traffic analysis, protocol analysis, content searching, and content matching. Consequently, this solution is not functional for unknown bots.

B. Saha and A, Gairola proposed an anomaly-based approach that uses its knowledge of what constitutes normal behavior and automatically classifies normal patterns; any deviation from normal pattern is classified as malicious and faulty. A key advantage of anomaly-based approaches is its theoretical ability to detect novel attacks [3].

The work by E.Bloedorn, A.D. Christiansenm and W. Hill et.al in [6] presented such that data mining aims at recognizing useful pattern to discover regularities and irregularities in large data sets. At data mining techniques, include correlation, classification, clustering, statistical analysis, and aggregation can be used for knowledge discovery about network nodes.

G. Gu, R. Perisci, J. Zhang et al. also presented botminer [10] is the most recent approach which applies data mining techniques for Botnet C&C traffic detection. Botminer is an advanced Botnet detection tool which is independent of Botnet protocol and structure. Botminer can detect real-world Botnets including IRC-based, HTTP-based, and P2P Botnets with a very low false positive rate.

G. Gu, P.Porras, V.Yegneswaran et al. BotHunter [8] detects the bots by associating IDS events to a user-defined bot infection dialog model, and it is a passive detection system. Compared with these techniques, BotProbe only requires a shot time to provide a result: at most one round of actual C&C communication.

In this paper, we explore the benefit of anomaly-based approach in security domain. Specifically, we analyze the node traffic behaviors characteristics in P2P botnets. Today, the detection of P2P botnets is more difficult and challenging. It is difficult to detect P2P botnets than a centralized botnet. Many researchers applied Support Vector Machine classifier to detect online botnet. In this system, we apply Support Vector Machine classifier but the proposed

feature sets are differing from other researcher's features. And also use Random Forest classification techniques to compare with Support Vector Machine classifier.

4. Techniques of P2P Botnet and Data Mining

In the field of P2P botnets classification, the characterizations of network traffic behaviors are proposed by using the two data mining techniques.

4.1. P2P Botnet

P2P techniques are becoming popular and it is difficult to trace. New tools and new techniques are required to prevent P2P based botnets. Agobot, pybot, Sinit, Phatbot, Nugache, Peacomm or Storm, onficker, Zeus, Waledac and Wordpress are the name of P2P bots. Among them, Storm and Waledac botents are famous botnets in the World and millions of personal computers are infected by these botnets in the world. They send 1.5 billion spam email messages daily and seriously affect the global network activities. At the moment, we would like to research to analyze the Waledac and Storm botnet traffic flows and pattern behaviors by using data mining techniques.

4.1.1. Waledac Botnet

The waledac botnet is a very famous spam spreading botnet, the packets set corresponds to a period of spamming attack.

Waledac establishes connections mainly through TCP packets and it uses the packets with parameters PSH and ACK to communicate with P2P botnets.

4.1.2. Storm Botnet

The storm botnet spreads quickly in a short time to form a large botnet. It was first discovered in 2007 and used the implementation of Distributed Hash table (DHT) in the Kademelia P2P networks. It utilizes email attachments to induce users to click on them. Because the change of its traffic flows are usually small. The primary protocol the used is UDP. Each bot will use UDP protocol to communicate. Normally, the storm will include a SMTP component to spread the spam email. Storm botnet sends UDP packets to a large number of botnets attempting to establish connections during the connection stage.

4.2. Data Mining Techniques

A wide range of data mining techniques including correlation, classification, clustering, statistical analysis and aggregation can be used for knowledge discovery and network nodes.

4.2.1 Random Forest (RF)

The random forest is an ensemble of unpruned classification or regression trees. Random forest generates many classification trees. Each tree is constructed by a different bootstrap sample from the original data using a tree classification algorithm [11]. Since each tree is constructed using the bootstrap sample, approximately one-third of the cases are left out of the bootstrap samples and not used in training. These cases are called out of bag (oob) cases.

After the forest is formed, a new object that needs to be classified is put down each of the tree in the forest for classification. Each tree gives a vote that indicates the tree's decision about the class of the object. The forest chooses the class with the most votes for the object.

The main features of the random forest algorithm are listed as follows.

(i) It is unsurpassable in accuracy among the current data mining algorithms.
(ii) It runs efficiently on large data sets with many features.
(iii) It can give the estimates of what features are important.
(iv) It has no nominal data problem and does not over-fit.
(v) It can handle unbalanced data sets.

In random forest, there is no need for cross-validation or a test set to get an unbiased estimate of the test error. Forest error rate are based on correlation between any two trees and the strength.

The system can employ the training data to build the forest, and then use the test data to calculate the error rate.

4.2.2. Support Vector Machine (SVM)

SVM is a kernel-based classification algorithm, which blends linear modeling and instance-based learning together. It provides an easy and efficient way of mapping data on to a higher dimensional space. The most important component in a SVM classifier is its kernel function. The most popular kernel functions used in SVM are Polynomial kernel and Gaussian distribution-based Radial Basis Function (RBF) kernel [14].

Advantages:

- Produce very accurate classifiers.
- Less overfitting, robust to noise.

Disadvantages:

- SVM is a binary classifier. To do a multi-class classification, pair-wise classifications can be used.
- Computationally expensive, thus runs slow.

In this study, we implement the SVM classifier based on Polynomial kernel in order to maintain a real-time detection.

5. Proposed Detection Model

The proposed P2P botnets classification framework is shown in Figure 3. The input to the detection framework is network time series packets.

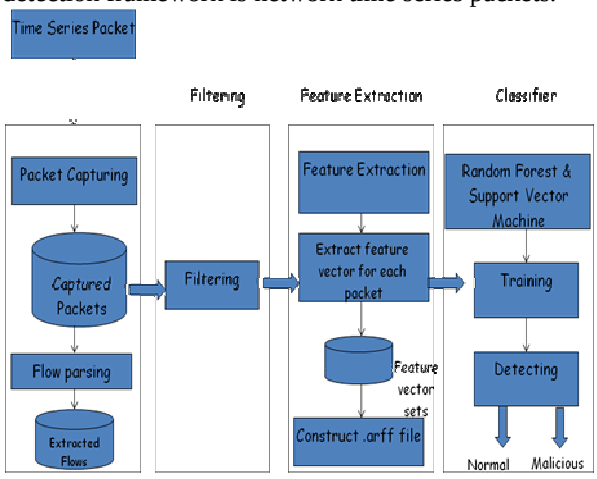

Figure 3.Classification Framework

The procedures of the framework are as follows.

(i) packet capture module
(ii) Filtering module
(iii) Feature extraction module
(iv) Classifier module

5.1. Packet Capture module

The packet capturing phase is responsible for capturing packets on a network interface. The module adopts a sampling strategy based on sliding window technique. The time-window sampling strategy is implemented by capturing packets during each time window (10s, 30s, 60s, 120s).

In the flow parsing, we use captured packets set to construct flows and decide how many different flows existing in a current time window. The output of the packets capture module is a packets set that contains all the captured network packets.

5.2. Filtering module

In the filtering module, the classification process can be speeded up by filtering out non- P2P packets through the well-known ports. The well-known ports are ranging from 0 to 1023, are those recognized and defined by the Internet Assigned Numbers Authority (IANA), but not all of the port numbers are defined. To reduce the processing time and data amount for classification, the system was focused on P2P traffic flows.

The well known Ports are filtered to remove non-P2P packets except port 53, port 23, port 80 and port 443 because P2P applications also communicate through these three ports. In this filtering, we proposed the port association algorithm to filter out non- P2P packets in Figure 4.We also set the length of the interest packet length as [40~159] bytes because of botnet characteristics.

```
BEGIN

    IF    ((1024<SOURCE_PORT<=65535) OR
          (SOURCE_PORT == 53 OR
          SOURCE_PORT==443 OR
          SOURCE_PORT == 23 OR
          SOURCE_PORT == 80))
THEN P2P
    IF    ((1024<DEST_PORT <=65535) OR
          (DEST_PORT == 53 OR
          DEST_PORT==443 OR
          DEST_PORT == 23 OR
          DEST_PORT == 80))
          THEN P2P
    IF     ((SOURCE_PORT <= 1024) AND
          (DEST_PORT <= 1024))
THEN NON-P2P
END
```

Figure 4. Port Association Algorithm

5.3. Feature Extraction Module

Features are extracted from Waledac botnet and Storm botnet. Data mining algorithms require appropriate ‘features’ as inputs in order to train models. For this research, network trace files (pcap) files were obtained. These trace files were then used for feature extraction and extracts ten features for detection .In proposed features we use information about the number of packet byte, the number of PSH and ACK packets per flow and the number of data bytes with maximum threshold in Table 1.

Table 1. Description of Proposed Features

No	Feature	Description of feature
1	SRC_IP	Number of Source IP address
2	SRC_PORT	Number of Source Port number
3	DES_IP	Number of Destination ip address
4	DES_PORT	Number of Destination port
5	Pkt_RATE	Rate of packet per second
6	Byt_RATE	Rate of Packets byte per second

7	NO_DB	Number of data bytes less than 25 and continuously data bytes reach to maximum threshold
8	AVG_LEN	Number of average packet length in a given time interval
9	NO_DUP	Number of duplicated packet length in a given time interval
10	NO_PSH_ACK	Number of times PSH and ACK set at TCP packets travelling in a given time interval

For Packet Rate,

$$\text{Packet rate per Second} = n_p \times \frac{1}{(t_e - t_s)} \quad (1)$$

n_p = number of packets
t_e = end packet sent time
t_s = start packet sent time

For byte Rate

$$\text{Byte rate per second} = b_t \times \frac{1}{t_e - t_s} \quad (2)$$

b_t = total number of bytes
t_e = end packet sent time
t_s = start packet sent time

Equation (1) and (2) are packet rate and byte rate calculation. The output of the Feature extraction module is a feature vector sets and .arff file is constructed based on the feature vector set.

5.4. Classifier Module

In this paper, we select Random Forest and Support Vector Machine techniques as the classification algorithm due to their classification accuracy. The classifier module takes .arff file as input and classifies the normal node and malicious node based on their behavior features. By using proposed framework, this system compared the accuracy of two data mining techniques. RF provides the high classification accuracy and the relatively robustness to outliers and noise among popular data mining techniques.

6. Experimental Results

In this system, we used ISOT dataset which is the combination of several existing publicly available malicious and non-malicious datasets. It is included two datasets containing malicious traffic from the French chapter of the Honeynet project, involving the Storm botnet and the Waledac botnet, respectively. After that contains the labeled datasets from the traffic lab at Erisson research in Hungary.

At ISOT dataset, contains over a million packets of general traffic that ranges from web browsing (HTTP) to peer to peer traffic and gaming such as Quake and World of Warcraft and packets from popular bittorent clients such as Azureus [7]. The ISOT dataset is the combination of several existing publicly available malicious and non-malicious datasets. This process is shown in Figure 5.

The experiments are supervised style, which means a training set and a testing dataset are included. In this system, contain 3500-4500 bot feature vectors and 3500-4500 normal feature vectors. Among the bot feature vectors, half of them are Storm Bot and half of them are Waledac Bot.

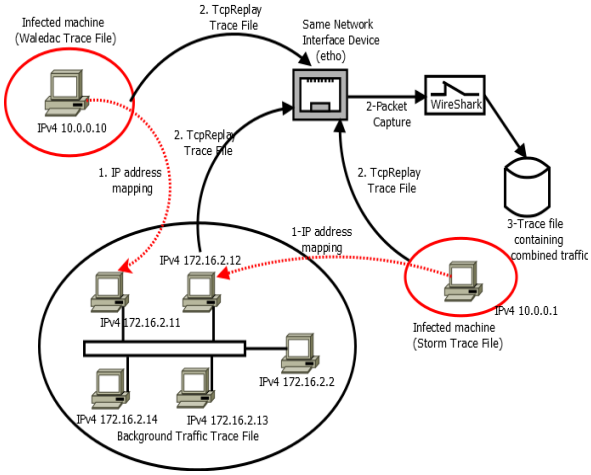

Figure 5. Dataset Merging Process

All of these three datasets are labeled so that we can efficiently conduct the evaluation experiments and summary of these traffic are shown in Table 2.

Table 2. Summary of Experiment Datasets

Source IP address	Type of traffic
172.16.0.11	Waledac & non Malicious
172.16.0.12	Storm (SMTP spam)
172.16.2.2	Non-malicious
172.16.2.3	Non-malicious
172.16.2.11	Storm & non-malicious
172.16.2.12	Zeus & non-malicious
172.16.2.111	Non-malicious
172.16.2.112	Non-malicious
172.16.2.113	Non-malicious
172.16.2.114	Non-malicious

Testing dataset contains 800-900 bot feature vectors and 800-900 normal features. Among the bot

feature vectors, half of them are Storm Bot and half of them are Waledac Bot.

The accuracy of the SVM and Random Forest classifier is measured using following four metrics: False negative rate (FNR), False positive rate(FPR), True Positive rate (TPR), and Accuracy. All of these four metrics are calculated using Equation (3) to Equation (6) respectively.

$$FPR = \frac{FP}{TN + FP} \quad (3)$$

The number of real normal nodes is classified as malicious nodes

$$TPR = \frac{TP}{TP + FN} \quad (4)$$

The number of real malicious nodes is classified as real malicious nodes

$$FNR = \frac{FN}{TP + FN} \quad (5)$$

The number of real malicious nodes is classified as normal nodes

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (6)$$

Accuracy describes the percentage of correct detection of both malicious and non-malicious nodes.

The following experiments compared the accuracy of Support Vector Machine and Random Forest in classifying P2P botnets viruses correctly. The accuracy of Random Forest Classifier (98.88%) is higher than that of Support Vector Machine (97.88%).At previous study [2] researchers presented evaluation works on SVM configuration in which they selected values (1,2,5) for parameter Exponent, and selected values (10^{-2},10^{-1},...10^{5}) for parameter complexity. Based on their study, this system calculated parameter Exponent on values (1, 1.5, 4) and evaluate parameter Complexity on values (10^{-2}, 10^{-1},10^{0}, 10^{1},10^{2}).

Random Forest first selects a small subset of available variables at random. It is actually a bootstrap subsample and typically select about square root (K), where K is the total number of predictors available. The accuracy result of the Random Forest system with the various numbers of trees and the accuracy rate is not significantly different starting from the number of trees 18 to the number of tress 300.

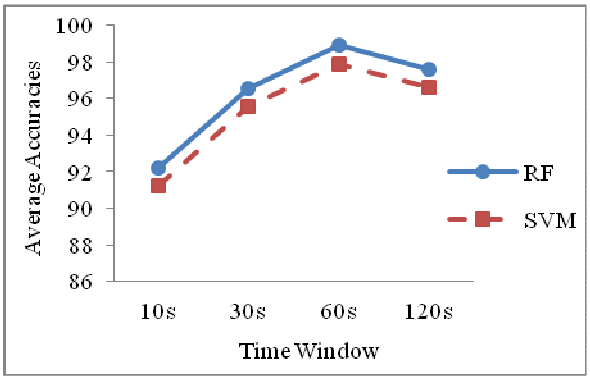

Figure 6. Comparison of Average Accuracies between Two Techniques

According to Figure 6, the average accuracies of the Random Forest are slightly higher than the Support Vector Machine. Evaluating results showed that by using a 60s time window in packets capturing process, the detection system can obtain the high accuracy of detection.

The system also showed that using these two classifiers and the activity of traffic classification with promising accuracy by observing small portions of a full network packets volume.

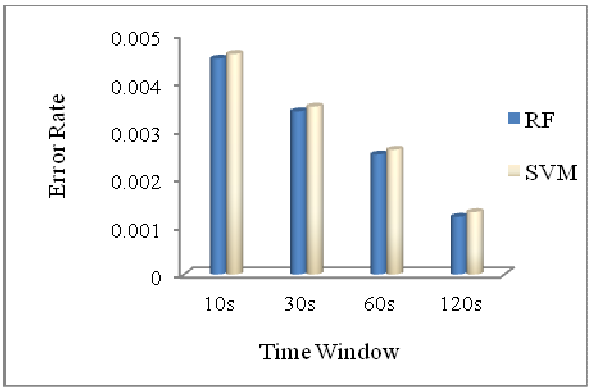

Figure 7. Detection Error Rate

Figure 7 shows the classification error rate in two data mining techniques.

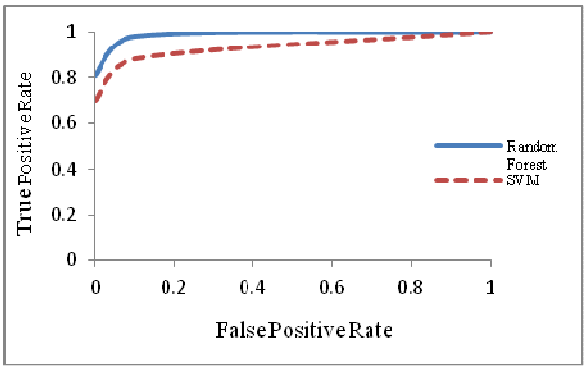

Figure 8. ROC curves of two techniques

Finally, the system show the Receiver Operating Characteristics (ROC) curves of two data mining approaches in Figure 8. The calculated results show that they are close to the area under the ideal curve (AUC=1). The proposed design is to have optimized the use of Random Forest algorithm for classification and it is manage to avoid false alarms during heavy traffic in networks. The accuracy of detection framework achieved in classification by removing unnecessary traffic at filtering module. Therefore, the results can prove the Random Forest approaches are reasonably competitive and practical for botnet detection.

7. Conclusion and Future works

In this paper, a P2P botnets virus detection system is improved based on two data mining algorithm, i.e., Random Forest and Support Vector Machine classifier. The results show that the system can identify normal or malicious flows produced by P2P botnets viruses correctly in a short time to achieve the goal of infection control. Future research should be focused on expanding the types of data mining techniques used in our classification system. We also intend to detect different classes of botnet in future work.

References

[1] A. Ramachandran and N. Feamster, " Understanding the network-level behavior of spammers," in Proc. ACM SIGCOMM, 2006.

[2] Asa Ben-Hur and Jason Weston, *A user's guide to support vector machine*, Methods Mol Biol. (2010), 223-239 (English).

[3] B. Saha and A, Gairola, "Botnet: An overview," *CERT-In White PaperCIWP-2005-05, 2005*.

[4] C.Schiller,J.Binkley and D.Harley (2007)."Botnet: The killer web applications, " Rockland, MA: Syngress Publishing.Feb. 2007.

[5] E.Cooke, F. Jahanianm and D. McPherson, The zombie roundup: Understanding, detecting, and disrupting Botnets," Proc.of workshop on Steps to Reducing Unwanted Traffic on the Internet (SRUTI'05), June 2005.

[6] E.Bloedorn, A.D. Christiansenm and W. Hill et.al; *Data mining for network intrusion detection: How to get started*, Tech. report, The MITRE Corporation, 2001.

[7] French Chapter of Honeynet http://www.honeynet. og/chapters/france.

[8] G. Gu, P.Porras, V.Yegneswaran, M. Fong, and W. Lee, "Bothunter: detecting malware infection through ids-driven dialog correlation, " in *Proceedings of 16th USENIX Security Symposium on USENIX Security Symposium*, (Berkely, CA, USA), pp. 12:1-12:16, USENIX Association, 2007.

[9] G. Gu, R. Perisci, J. Zhang , and W. Lee, "Botminer: Clustering analysis of network traffic for protocol- and structure independent Botnet detection," in Proc. 17th USENIX Security symposium , 2008.

[10] K. K. R. Choo, "Zombies and Botnets, "Trends and issues in crime and criminal justicem no. 333, Australian Institute of Criminology, Canberra, March 2007.

[11] L.Breiman, " Random Forests" , Machine Learning 45(1):5-32,2001.

[12] M.A Rajab, J. Zarfoss, F. Monrose, and A. Terzis, "A multifaceted approach to understanding the Botnet phenomenon, " *6th ACM* SIGCOMM on Internet Measurement Conference, IMC 2006, 2006, pp.41-52.

[13] M. Rosech, *Snort-lightweight intrusion detection for networks*, Proceedings of LISA '99: 13th Systems Administration Conference, The USENIX Association, Novenber 1999.

[14] www.cs.uky.edu/~jzhang/CS689/PPDM-Chapter2.pdf

[15] Z.Zhu,G.Lu,Y. Chen, Z.j. Fu, P.Roberts, K.Han, "Botnet Research Survey," in Proc. *32nd Annual IEEE International Conference on Computer Software and Applications* (COMPSAC '08), 2008, pp.967-972.