# Investigation of peer-to-peer Botnet using TCP control packets and data mining techniques

**Conference Paper** · December 2013

CITATIONS
0

READS
421

**4 authors**, including:

Mohammad Alauthman
Petra University
**42** PUBLICATIONS **546** CITATIONS
SEE PROFILE

Mohammed Alamgir Hossain
Anglia Ruskin University
**156** PUBLICATIONS **2,516** CITATIONS
SEE PROFILE

Rafe Alasem
University of Bradford
**8** PUBLICATIONS **162** CITATIONS
SEE PROFILE

# Investigation of peer-to-peer botnet using TCP control packets and data mining techniques

Mohammad Alauthaman*, Nauman Aslam, and M.A. Hossain

Department of Computer Science and Digital Technologies,
Faculty of Engineering and Environment,
Northumbria University,
Newcastle upon Tyne, NE1-8ST, UK
Email: mohammad.alauthaman@northumbria.ac.uk
Email: nauman.aslam@northumbria.ac.uk
Email: alamgir.hossain@northumbria.ac.uk
*Corresponding author

Rafe Alasem

Department of Electrical Engineering,
Faculty of Engineering,
Imam Mohammad Ibn Saud Islamic University,
Riyadh, Saudi Arabia
Email: rkasem@imamu.edu.sa

**Abstract:** Nowadays botnets are commonly used in cyber-attacks and malicious activities. A botnet is the main way to carry and spread many malicious codes in internet that are responsible for many malicious activities including spam mail, distributed denial of service attack and click fraud. In this paper, we propose an approach to detect botnet's malicious behavior by using data mining classification techniques based on the features of TCP control packet. We study the performance and accuracy of popular classification techniques on existing datasets. Experiment shows that the proposed approach is able to identify botnets with high accuracy rate and high performance in a short time. The evaluation results show that the proposed solution can detect bot hosts with more than 99% accuracy, whereas the average of false positive rate is lower than 2%.

**Keywords:** P2P botnet; data mining; C4.5 algorithm; TCP protocol; command and control; C&C.

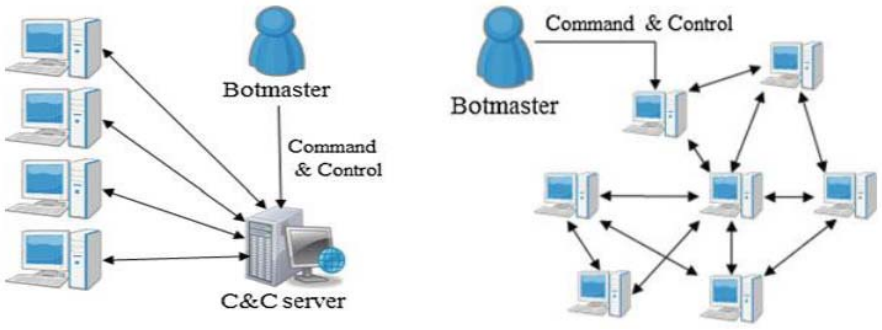
## 1 Introduction

Internet services are increasing in popularity; many online services and activities appear daily. These online services cause a massive volume of online financial transactions and sensitive information to be exchanged via internet converting the attacker's interest from curiosity to financial benefit. Attackers use different malwares to achieve their goals.

Among the various forms of malware, botnet is considered as the most serious mean for conducting online crimes (Silva et al., 2013).

A botnet is a network of compromised computers (bots) remotely managed by an attacker (Botmaster). To effectively administer a botnet, the Botmaster constructs an infrastructure of the communication channel to send a command to the bots, and to receive the result from them (Lu et al., 2011). This communication channel is known as command and control (C&C) channel. The main difference between a botnet and other malware is the infrastructure used in the C&C (Zeidanloo et al., 2010). In contrast with other malware which performs malicious behavior individually, a botnet works as a group of infected host based on C&C communication channel. botnets can be classified into two main categories based on C&C infrastructure: centralized C&C and decentralized C&C (Han and Im, 2012). In centralized botnets, the Botmaster usually uses the C&C server to send a command to the bots as shown in Figure 1.

**Figure 1** Structures of the botnet, (a) centralized structure (b) decentralized structure




(a) (b)

Due to its simplicity and fast connection approach command, centralized botnets are widely used. The IRC and HTTP botnet are among the most famous approaches. However, despite their simplicity, the main limitation of a centralized botnet is a single point of failure of the C&C server. In order to avoid the weakness of a single point of failure in centralized botnets, the attackers have recently started to build botnets based on decentralized C&C infrastructure such as P2P botnet (Felix et al., 2012).

Peer-to-peer (P2P) botnet is a new class of botnet which has replaced the old centralized IRC/HTTP based botnet to avoid a single point of failure and avoid detection in C&C connection. In its architecture as shown in Figure 1 there is no central point for C&C server, so that any host in the network can work as a client and a server at the same time. Thus, if a single client is taken offline, the gaps in the network will be closed automatically and the network shall continue to work (Grizzard et al., 2007). Therefore, attackers are aware of the strengths of P2P botnets without single point failure and increase the creation of botnet based on P2P structure such as Storm bot, Conficker bot and Waledac bot (Davis et al., 2009). Many existing botnets detection systems focus on the following strategies:

1 detect bot during the attack phase

2 require deep packet inspection for signature matching

3 analyzing whole network traffic

4 rely on group botnets activities.

In this work, we divided the TCP flow into control packets and payload packets. Then we extract features from the header of control packets, and it will be used to classify connection as either malicious or normal based on data mining classification techniques. Practically speaking, we estimate the accuracy of the C4.5 algorithm.

The rest of this paper is organized as follows: Section 2 briefly review some related work to P2P botnet detection approach. A proposed approach described in Section 3. Section 4 presents the experimental results. Finally, the conclusion and future work present in Section 5.

## 2 Related work

In recent years, the research on botnet detection has become an important issue. Nevertheless, the detection of P2P botnet has only recently emerged. The work in Han and Im (2012) classify P2P botnet detection system into three general types, data mining, machine learning and network behavior and traffic analysis. Feily et al. (2009) classify botnet detection system into four general categories; signature-based detection, anomaly-based detection, DNS-based detection and mining-based detection. Our focus will be on data mining, network behavior and traffic analysis. The authors in Wen-Hwa and Chia-Ching (2010) use a methodology based on packet sizes to distinguish between P2P botnet traffic and legitimate P2P traffic. They present the following observations. Firstly, P2P Bot tries to update information for other bots more than stay idle. Secondly, Bot transmits data at the minimum level of communication. Also they use WEKA analysis tools to analyze network traffic. It was found that P2P botnet packets size is smaller than any other P2P applications. Zhao et al. (2012) introduced P2P botnet detection system using machine learning techniques based on flow interval of network traffic. They applied Bayesian network and decision tree as a classification method to investigate online P2P bot detection. The main drawback of this technique is its sensitivity to evasion methods such as random connection interval.

Our approach is similar to the method expressed in Wen-Hwa and Chia-Ching (2010), and Tarng et al. (2012). We increase detection accuracy and performance by reducing the amount of input packets. The proposed approach presents features based on control packet only and utilizes data mining techniques.

## 3 Proposed approach

As illustrated in Figure 2 our proposed detection approach can be divided into four phases. Firstly, the system captures network traffic and filters all TCP control packets. Secondly, we generate connection information between hosts based on time interval. Thirdly, we extract knowledge from connections important to detect botnet. Finally, we utilize data mining classification techniques to classify connections status as normal connections or botnet connections. Each phase is explained in detail in the following subsections.

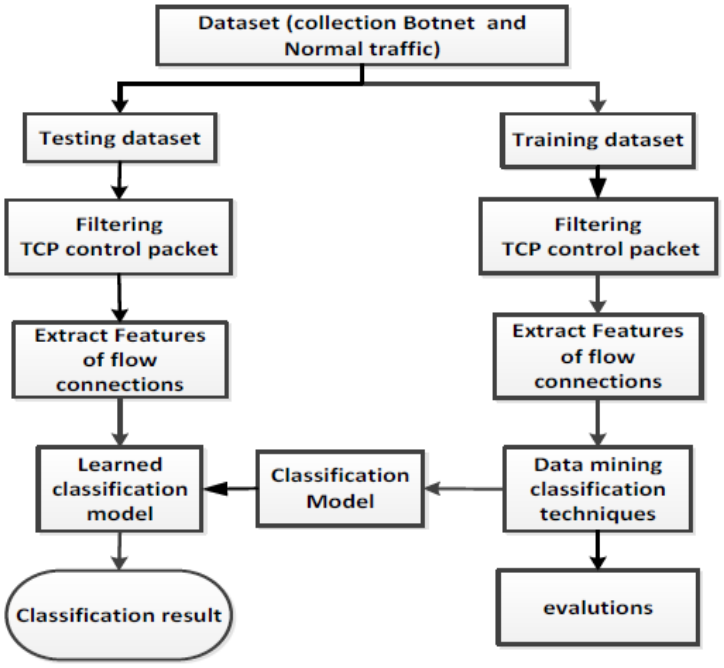
### *3.1 Filtering*

We filter only TCP control packets to reduce the volume of network traffic as well as to increase the performance of proposed approach. In the filtering phase, we split the operation into two steps. First, all TCP traffic is filtered. Second, TCP control packets are extracted.

### *3.2 Feature extraction*

In features extraction phase, we extract features important to detect botnet's malicious behavior; and gather features in 8-tuple attribute every time interval to pass to classification algorithm. We chose features that are commonly used in previous methods (Zhao et al., 2012; Wen-Hwa and Chia-Ching, 2010; Junjie et al., 2011) to detect botnet as shown in Table 1. These features were extracted based on the definition of connection as group of packets exchanged between two different hosts, which are identified by the 4-tuple (source IP address, destination IP address, source port and destination port).

**Figure 2** Proposed approach



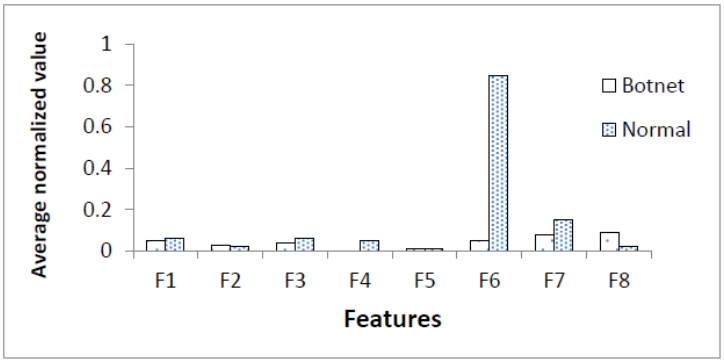

In our proposed method, all features are extracted directly from control packets header. In comparison with previous approach like Lu et al. (2011), where they used a deep inspection payload packet content. As a consequence, it increases the performance of detection, and it reduces the use of system resource such as memory computations in the processor. To estimate the importance of each feature for successful detection, we evaluate the normalized value for each vector of feature by using Min-Max normalization

(Al Shalabi and Shaaban, 2006). Figure 3 shows the average distribution of normalization value for each feature. We found different distributions between botnet traffic and normal traffic. The average length of TCP control packets and the average received of length TCP control packets are a highly significant feature to detect botnet flow connections. The features F6, F7 and F8 are of high importance. For the remaining other features, F5 and F2 have a low significant to differentiate between network traffic.

**Table 1** List of connections features based control packet

| *Features number* | *Features description* |
|---|---|
| F1 | Total number of sent control TCP packets per connections |
| F2 | Total number of received TCP control packets |
| F3 | Total number of TCP control packets |
| F4 | Total length of sent TCP control packets |
| F5 | Total length of received TCP control packets |
| F6 | Average length of sent TCP control packets |
| F7 | Average length of received TCP control packets |
| F8 | Average length of TCP control packets |

**Figure 3** Normalized features comparison




### *3.3 Data mining classification algorithm*

In a classification approach, the classifier tries to learn several features as an input to predict an output. In the case of botnet detection, the classifier tries to classify the behavior of TCP traffic as botnet malicious or legitimate by learning features and knowledge extracted for TCP traffic. We selected popular data mining classification techniques for evaluation based on TCP control packet features, decision tree classification C4.5. The C4.5 algorithm is a popular approach for classification and prediction of decision trees. In constructing a decision tree, a target record will enter the root node, and a branch for each possible value for the targets is built. The same process is applied repeatedly until all the records in a node end up with the same class or the tree cannot be split any further (Quinlan, 1979).

It's worth mentioning that there are several reasons encouraged us to choose the C4.5 classifier. The first reason is that the C4.5 algorithm achieves good accuracy result with minimum computational time. Second, it is simple and easy to implement. Third, more complex data mining algorithm needs longer training and classifying time.

## 4 Experiment and evaluations

### *4.1 Dataset*

We obtained two datasets that contain malicious and non-malicious traffic used in evaluating our proposed system. The source of datasets is provided as follows: first, IOST dataset (Saad et al., 2011) that contains malicious traffic from French chapter of honeynet project involving Waledac botnet and Storm botnet. It also contains non-malicious labeled traffic collected from the Traffic Lab at Ericsson Research in Hungary and from Lawrence Berkeley National Lab (LBNL). Second, ISCX dataset (Shiravi et al., 2012) that involves a normal activity and non-malicious traffic. We merged all dataset packets in one file. In order to evaluate our approach, we split traffic into 10, 30 and 60 second time interval. After monitoring traffic for each time interval, we passed packets to generate connection and to extract features and then save the result into '.arff' file format and then passed the '.arff' file to WEKA.

### *4.2 Evaluation result*

In our work, we used WEKA for data mining purpose. It includes tools for data preprocessing, classification, regression, clustering, association rules, and visualization (Hall et al., 2009). Our implemented program is used to extract features from a PCAP file needed in classification techniques. In order to evaluate the accuracy of detection, we used a ten-fold cross-validation to estimate the error rate for classifiers. The true positive rate (TPR), false positive rate (FPR) and accuracy rates of the C4.5 algorithm are listed in Table 2.

**Table 2** Classification result

| *Time interval* | *Average TPR* | *Average FPR* | *Average accuracy rate* |
|---|---|---|---|
| 10 second | 97.3 | 6.2 | 98.1 |
| 30 second | 99.6 | 1.4 | 99.4 |
| 60 second | 99.6 | 2.9 | 99.2 |

We use three different scales of time interval 10, 30 and 60 seconds. The average performance of each time interval is shown in Table 2. The results show that the 30 seconds time interval has high accuracy rate 99.4% and low FPR of 1.4%. Additionally, we estimate the importance of features by measuring the gain ratio with respect to the class. Figure 4 shows the WEKA result of average importance rank based on C4.5 decision tree classification.

To test the efficiency our proposed approach in detecting botnets, we plot the receiver operating characteristic (ROC) curve to show the trade-off between TPR and FPR. A perfect classifier would have an area under curve (AUC) close to 1.0. The x-axis represents a FPR and y-axis represents a TPR. As shown in Figure 5, the AUC is 0.993.

We find that the proposed approach performs well on classifying botnet connection traffic based on TCP control packets in 30 seconds time interval.

**Figure 4** Evaluation features rank

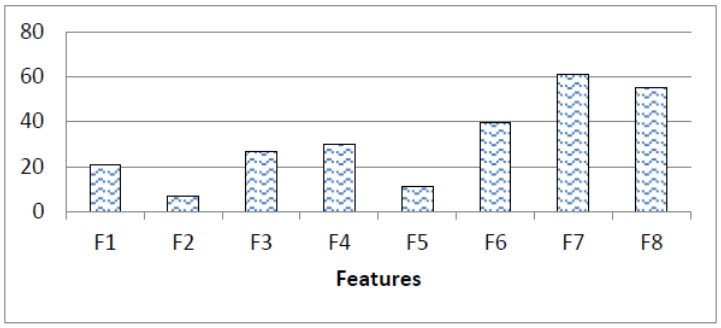



**Figure 5** ROC curves 30-second time interval

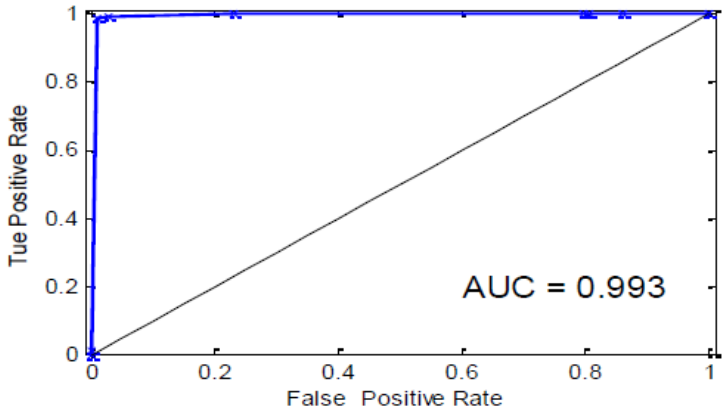



**Table 3** Result comparison

| *Approach* | *Wen-Hwa and Chia-Ching (2010)* | *Tarng et al. (2012)* | *Proposed approach* |
|---|---|---|---|
| Core technique | C4.5 tree, Naive Bayes and Bayesian network classifiers | Bayes classifier and neural network | C4.5 tree |
| Accuracy rate | C4.5: 98%<br>Naïve Bayes: 89%<br>BayesNet: 87% | BayesNet: 95.7%<br>Neural network: 98.7% | Average accuracy: 99.4% |
| Traffic filtration rate | N/A | N/A | More than 50% |

Comparing with other approaches, the proposed approach reduces the amount of processing to increase the performance. It is not easy to compare among different botnet detection approaches because each uses different dataset and botnet samples in the

experiments. Therefore, we compare the proposed approach with another detection approach based on the accuracy of detection and traffic filtering rate. Table 3 shows a comparison of two previous data mining botnet detection approaches. The table also shows that the detection accuracy of the proposed approach is better than the previous solutions. In addition, Table 3 shows a comparison of traffic filtering rate with other data mining approaches. We observe that the proposed approach is not only efficient and robust but also lightweight in memory and computational costs.

## 5 Conclusions

In this work, we investigated the use of TCP control packets to detect P2P botnet malicious behaviour. In this regard, our approach first filtered TCP control packets and translated packets to flow connections. In the next step, important features were extracted from connections to help detect botnets. Finally, we utilized data mining classification techniques to identify P2P botnet malicious connections. We implemented the approach and evaluated the detection accuracy. The experiment results confirm the fact that TCP control packets could be used to detect P2P bots with high detection accuracy and low FPR. In the future, we plan to extend our approach by adding unsupervised learning approach to select relevant features that would further increase the accuracy and performance.

## References

Al Shalabi, L. and Shaaban, Z. (2006) 'Normalization as a preprocessing engine for data mining and the approach of preference matrix', in *Dependability of Computer Systems*, pp.207–214, IEEE.

Davis, C.R., Fernandez, J.M. and Neville, S. (2009) 'Optimising sybil attacks against P2P-based botnets', in *2009 4th International Conference on Malicious and Unwanted Software (MALWARE)*, pp.78–87.

Feily, M., Shahrestani, A. and Ramadass, S. (2009) 'A survey of botnet and botnet detection', in *Third International Conference on Emerging Security Information, Systems and Technologies, 2009, SECURWARE '09*, pp.268–273.

Felix, J., Joseph, C. and Ghorbani, A. (2012) 'Group behavior metrics for P2P botnet detection', in Chim, T. and Yuen, T. (Eds.): *Information and Communications Security*, Vol. 7618, pp.93–104, Springer, Berlin, Heidelberg.

Grizzard, J.B., Sharma, V., Nunnery, C., Kang, B.B. and Dagon, D. (2007) 'Peer-to-peer botnets: overview and case study', *Proceedings of the First Conference on First Workshop on Hot Topics in Understanding Botnets*, USENIX Association, Cambridge, MA.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I.H. (2009) 'The WEKA data mining software: an update', *SIGKDD Explor. Newsl.*, Vol. 11, pp.10–18.

Han, K-S. and Im, E. (2012) 'A survey on P2P botnet detection', in Kim, K.J. and Ahn, S.J. (Eds.): *Proceedings of the International Conference on IT Convergence and Security 2011*, Springer, Netherlands, Vol. 120, pp.589–593.

Junjie, Z., Perdisci, R., Wenke, L., Sarfraz, U. and Xiapu, L. (2011) 'Detecting stealthy P2P botnets using statistical traffic fingerprints', in *2011 IEEE/IFIP 41st International Conference on Dependable Systems & Networks (DSN)*, pp.121–132.

Lu, W., Rammidi, G. and Ghorbani, A.A. (2011) 'Clustering botnet communication traffic based on n-gram feature selection', *Computer Communications*, Vol. 34, No. 3, pp.502–514.

Quinlan, J. (1979) 'Discovering rules from large collections of examples: a case study', in Michie, D. (Ed.): *Expert Systems in the Micro Electronic Age*.

Saad, S., Traore, I., Ghorbani, A., Sayed, B., Zhao, D., Wei, L., Felix, J. and Hakimian, P. (2011) 'Detecting P2P botnets through network behavior analysis and machine learning', in *2011 Ninth Annual International Conference on Privacy, Security and Trust (PST)*, pp.174–180.

Shiravi, A., Shiravi, H., Tavallaee, M. and Ghorbani, A.A. (2012) 'Toward developing a systematic approach to generate benchmark datasets for intrusion detection', *Computers & Security*, Vol. 31, No. 3, pp.357–374.

Silva, R.S.C., Silva, R.M.P., Pinto, R.C.G. and Salles, R.M. (2013) 'Botnets: a survey', *Comput. Netw.*, Vol. 57, No. 2, pp.378–403.

Tarng, W., Chou, C-K. and Ou, K-L. (2012) 'A P2P botnet virus detection system based on data-mining algorithms', *International Journal of Computer Science*, Vol. 4, No. 5, pp.51–65.

Wen-Hwa, L. and Chia-Ching, C. (2010) 'Peer to peer botnet detection using data mining scheme', in *2010 International Conference on Internet Technology and Applications*, pp.1–4.

Zeidanloo, H.R., Bt Manaf, A., Vahdani, P., Tabatabaei, F. and Zamani, M. (2010) 'Botnet detection based on traffic monitoring', in *International Conference on Networking and Information Technology (ICNIT)*, IEEE, pp.97–101.

Zhao, D., Traore, I., Ghorbani, A., Sayed, B., Saad, S. and Lu, W. (2012) 'Peer to peer botnet detection based on flow intervals', in Gritzalis, D., Furnell, S. and Theoharidou, M. (Eds.): *Information Security and Privacy Research*, Vol. 376, pp.87–102, Springer, Berlin, Heidelberg.