Library Progress International *Print version ISSN 0970 1052*
Vol.44 No.3, Jul-Dec 2024: P. 26203-26209 *Online version ISSN 2320*

Original Article *Available online at www.bpasjournals.com*

Feature Extraction Based IoT Botnet Detection Using Machine Learning Technique

[1]Brajesh Mishra, [2]Ravi Shankar Sharma

[,2] Department of Computer Application, Rabindranath Tagore University, Raisen, Madhya Pradesh, India
[1] brajeshmishra2780@gmail.com,
[2] rssharma@aisectuniversity.ac.in

How to cite this article: Brajesh Mishra, Ravi Shankar Sharma (2024) Feature Extraction Based IoT Botnet Detection Using Machine Learning Technique. *Library Progress International,* 44(3), 26203-26209

Abstract
In recent years, there has been a proliferation of IoT (Internet of Things) devices and its enabling technologies in industries, product flow management, healthcare, transportation and other smart environments. The provision of IoT devices with IP (Internet Protocol) address allows for communication between these cyber-physical systems without any intervention. Lack of security on these end devices has led to many attacks like denial-of-service, Botnets, identity theft and data theft attacks. Botnets are one of the most serious threats to the Internet security and one of the most challenging topics within the fields of computer and network security today. Machine learning techniques can combats cyber-attacks by detection and prevention of these Botnets. In this paper, we explore on Botnet attacks that is prevalent in IoT devices as well as this paper presents a new feature combination based approach for botnet detection. Various features like generic features, statistical & subnet features have extracted during feature extraction. Than ML methods has been applied on combination of feature set. The proposed method has achieved average accuracy around 99.8%. The developed method has also been evaluated on various machine learning models.

Keywords: Botnet, IoT traffic, attacks, anomaly detection, machine learning, feature engineering.

Introduction

The growth of the internet in today's world has made our life fast and easy but it has also added a lot of challenges in dealing with its security and privacy. The botnet is one of the most dangerous attacks in the network amongst the various existing malware. The term Bot in a botnet is derived from the word Robot which works according to the scripts or the program written by an attacker (Botmaster). The botnet is the collection of various infected hosts (called bots) that are created and controlled by an attacker remotely through the Command and Control (C&C) channel. Botnets are used in a variety of malicious activities like click fraud, phishing, spamming, malware delivering and DDoS (Distributed Denial of Service) [1], [2], [3]. Internet of things (IoT) are growing exponentially and playing a vital role in our everyday life. IoT nodes can use internet protocol address and connect to internet. These self-configured smart nodes are driving beyond many cutting-edge applications such as process automation, home automation, smart cars, decision analytics, smart grids, health care system, educational development, industrial development and so on [4][5]. Analysts are predicting that there will be a society with more connected devices than people living on this planet. The International Data Corporation (IDC) forecasted that there would be 41.6 billion connected IoT devices producing 79.4 zettabytes (ZB) of data in 2025 compared to the estimated population of 8.1 billion [6].

In IoT systems, heterogeneous nodes are connected to complex network architecture and pose security concerns. The key challenge is to ensure security in resource constraint IoT nodes [7]. Otherwise, these IoT nodes are vulnerable to different types of attacks. IDSs are a rudimentary and powerful security mechanism in maintaining sufficient network protection in any IoT embedded environment [8][9]. They are proficient in monitoring, analyzing, and detecting real-time data packets through passive traffic collection even-if they are intruders or not.

IDSs are traditionally organized into network-based (NIDS) and host-based IDS (HIDS) based on the detection places. Any IDS aims to monitor traffic and recognize different malware activities immediately [10][11]. With the escalating number of anomalies, the upgrading and development of IDSs have become exceedingly important as the main challenge in intrusion detection is to find out the obscure attacks from the routine traffic flow. Due to the dynamic approach of drawing a fine line between malware and benign data with high detection accuracy, shallow Machine Learning (ML) has become the center of attention of many researchers to upgrade the performance of IDSs [11]. Many supervised and unsupervised ML tools are effectively introduced for this purpose i.e., fuzzy logic (FL), support vector machine (SVM), artificial neural network (ANN), K-nearest neighbor (KNN), logistic regression (LR), hidden Markov model (HMM), genetic algorithm (GA), naive Bayes (NB), random forest (RF), decision tree (DT), decision forest (DF), decision jungle (DJ) and compacted hybrid algorithms, etc. [12]. Like worms and viruses, a bot is a program that infects the vulnerable hosts to extend its reach. A defining characteristic of the botnet uses the C&C channel. They can be updated and led through this channel [13]. The attacker may use a wide variety of transport protocols for C&C communication such as TCP, UDP, ICMP, GRE, etc. In essence by controlling all the hosts remotely, this is an effective distributed platform which the adversary can use to send spams, launching DDOS attacks, click fraud, privacy leakage by making it hard for the defender to trace back to the adversary. In order to successfully neutralize such attacks we need to develop an efficient system which provides high accuracy. Machine learning (ML) algorithms have shown promise in detecting anomalies in various domains, and their application to IoT network traffic analysis holds great potential. This research aims to investigate the use of ML techniques for detecting network traffic anomalies in IoT environments.

1. Literature review

Various botnet detection techniques are mentioned by author of [16].

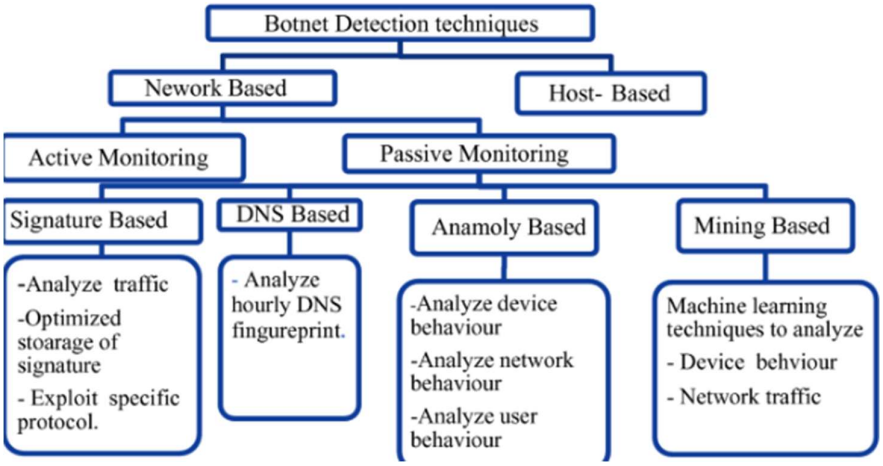

Figure 1: Botnet detection techniques [16]

Taxonomy of machine learning methods for IoT botnet is shown below:

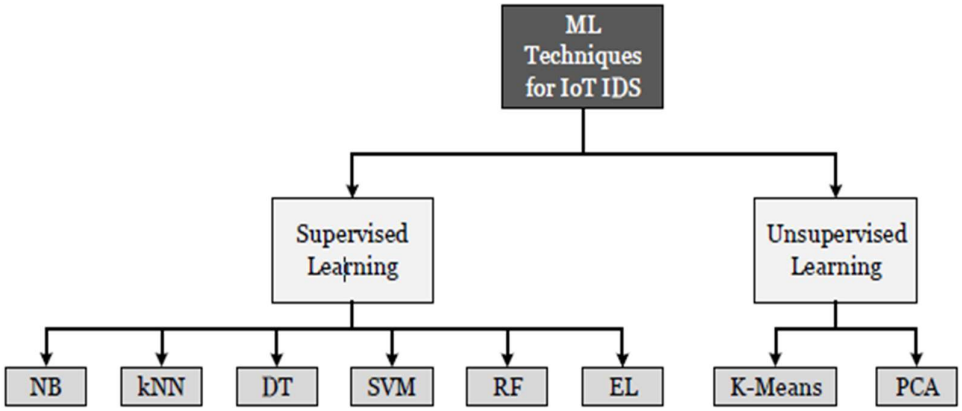

Figure 2: Taxonomy of ML methods for IoT botnet [17]

The paper [14] proposes a multilayer framework for botnet detection using machine learning algorithms that consist of a filtering module and classification module to detect the botnet's command and control server. They have used behavior-based analysis through flow-based features that analyzed the packet header by aggregating it to a 1-s time. This type of analysis enables detection if the packet is encapsulated, such as using a VPN tunnel. They have used unsupervised clustering. In [15], a botnets' detection method, based on machine learning, is formalized and evaluated. This proposal makes use of Splunk, a tool that allowed us to use the Random Forest algorithm to analyze DNS logs in order to detect connections to C&C servers. The resulting procedure complements the use of machine learning with the verification against other data sources for improving the results. The proposed approach of [18] first selects the optimal features using feature selection techniques. Next, it feeds these features to machine learning classifiers to evaluate the performance of the botnet detection. Experiments reveals that the proposed approach efficiently classifies normal and malicious traffic at an early stage. They have used 4 different classifiers such as (SVM, Logistic Regression, Multilayer Perceptron, and Random Forest) out of which random forest produces good results. In the future, further investigation is needed to discover the detection of botnet not just for centralized architecture but also to deal with decentralized architecture which is P2P based botnets. The study of [19] proposes machine learning methods for classifying binary classes. This purpose is served by using the publicly available dataset UNSW-NB15. This dataset resolved a class imbalance problem using the SMOTE-OverSampling technique. A complete machine learning pipeline was proposed, including exploratory data analysis, which provides detailed insights into the data, followed by preprocessing. During this process, the data passes through six fundamental steps. A decision tree, an XgBoost model, and a logistic regression model are proposed, trained, tested, and evaluated on the dataset. In addition to model accuracy, F1-score, recall, and precision are also considered. Based on all experiments, it is concluded that the decision tree outperformed with 94% test accuracy. The authors [20] developed a Host Intrusion Detection and Prevention System (HIDPS) is implemented in a fog computing infrastructure for real-time and precise attack detection. The proposed model integrates NIDS with federated learning, allowing devices to locally analyze their data and contribute to the detection of anomalous traffic. The distributed architecture enhances security by preventing volumetric attack traffic from reaching internet service providers and destination servers. This research contributes to the advancement of cybersecurity in local network environments and strengthens the protection of IoT networks against malicious traffic. This work highlights the efficiency of using a federated training and detection procedure through deep learning to minimize the impact of a single point of failure (SPOF) and reduce the workload of each device, thus achieving accuracy of 89.753% during detection and increasing privacy issues in a decentralized IoT infrastructure with a near-real-time detection and mitigation system.

BotHunter [21] aims to recognize the infection and coordination dialog that occurs during a successful malware infection. A similar approach, BotSniffer [22], focuses on the detection of C&C channels which are essential to a botnet. Therefore it exploits the underlying spatio-temporal correlation and similarity property of botnet C&C (horizontal correlation). The C&C server uses to contact every bot at the same time, then each of them uses to undertake some malicious actions following the C&C commands; these behaviours can be observed simultaneously in a network to spot a C&C channel, thus an underlying botnet. BotHunter and BotSniffer perform their evaluation on their own honey net or on traces authors built by executing malware binaries. However these

traces are not publicly available and [23] highlighted the lack of suitable comparisons for botnet detection algorithms due to the lack of public botnet datasets.

Methodology

The botnet detection system proposed in this paper makes use of features from the Net Flow data set to distinguish the malicious IPs. The system has 2 phases: training phase and testing phase. During the training phase the system identifies different features per destination IP and a classification model is trained based on these feature inputs. During testing phase the classifier is run on the unlabeled data set and the prediction is verified against the ground truth data. At a high level following models the design of the botnet detection system using feature wise supervised machine learning algorithm.

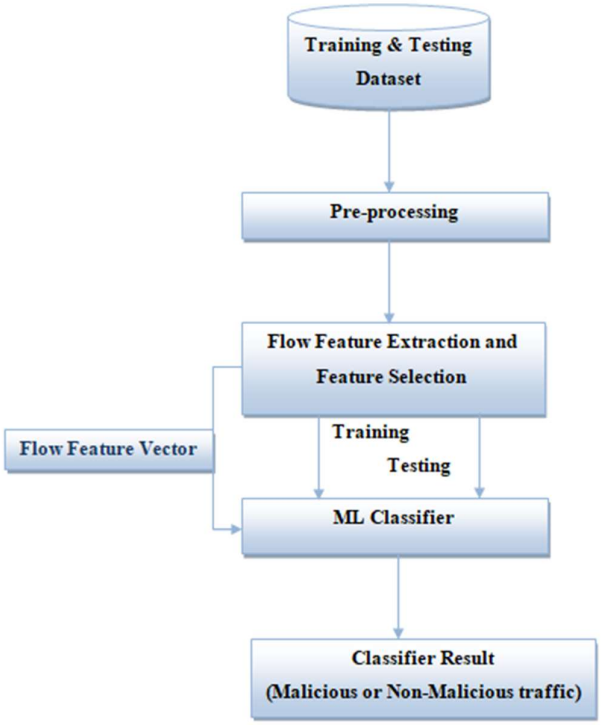

Figure 3: Proposed System

The feature extraction stage extracts the necessary details for each unique destination IP by analyzing the Net Flow dataset and building an intermediate feature dataset. Total 35 key features are identified from the Net Flow dataset. Features are categorized into four:

1. Generic Features
2. Statistical Features
3. Subnet Features
4. Periodic Communication Features

List of features extracted for each destination IP in the Network Flow are as follows:

Table 1: Features extracted

Features /dst IP	**Type**	**Feature Group**
Total Source IPs	Numerical	**F1**
Total Protocols	Numerical	
Total BidirectionalFlows	Numerical	
Total ClientFlows	Numerical	
Total ServerFlows	Numerical	
Protocol Information	Bit String	
Total, Max, Min,Mean, Std Dev,Var of Packets	Numerical	**F2**
Total, Max, Min,Mean, Std Dev,Var of flows	Numerical	
Total, Max, Min,Mean, Std Dev,Var of Bytes	Numerical	
Total, Max, Min,Mean, Std Dev,Var of Source Bytes	Numerical	
Total IPs in each /24 subnet of dstIPs	Numerical	**F3**
Total Flows in each /24 subnetof dst IPs	Numerical	
Total Packets ineach /24 subnetof dst IPs	Numerical	
Total periodic communications	Numerical	**F4**
Ratio of Total Source IPs with Periodic commn. / Total SourceIPs	Numerical	

The feature sets we evaluated are F1, F2, F3, F4, (F1,F2), (F1,F3), (F1, F4), (F2, F3), (F2, F4), (F3,F4), (F1,F2,F3), (F1,F3,F4), (F2,F3,F4) and all features. It is expected that by grouping feature sets proposed system will produce more accurate results rather than individual feature set.

4. Result and Evaluation

For the dataset, we made use of CTU-13 bidirectional Net Flow dataset. The CTU-13 is a dataset of botnet traffic that was captured in the CTU University, Czech Republic, in 2011. The machine learning models and the data visualization are implemented in python using jupyter. For the machine learning library, we used scikit and tensor flow.

Evaluated values for average accuracy are shown in table below:

Table 2: Average accuracy

Feature Set	**Avg Accuracy RF**	**Avg Accuracy NN**	**Avg Accuracy LR**
FeatureSet 1	0.998	0.84365	0.758
FeatureSet 2	0.997	0.54885	0.494
FeatureSet 3	0.962	0.72685	0.696
FeatureSet 4	0.505	0.50905	0.492
FeatureSet (1,2)	0.998	0.8691	0.763
FeatureSet (1,3)	0.998	0.899	0.825
FeatureSet (1,4)	0.998	0.8522	0.768
FeatureSet (2,3)	0.998	0.7137	0.7
FeatureSet (2,4)	0.997	0.5387	0.494
FeatureSet (3,4)	0.962	0.73845	0.7
FeatureSet (2,3,4)	0.998	0.7502	0.704
FeatureSet (1,3,4)	0.998	0.90445	0.829
FeatureSet (1,2,4)	0.998	0.87245	0.772
FeatureSet (1,2,3)	0.999	0.90835	0.828
All Features	0.998	0.91105	0.835

Chart for comparing average accuracy is shown below:

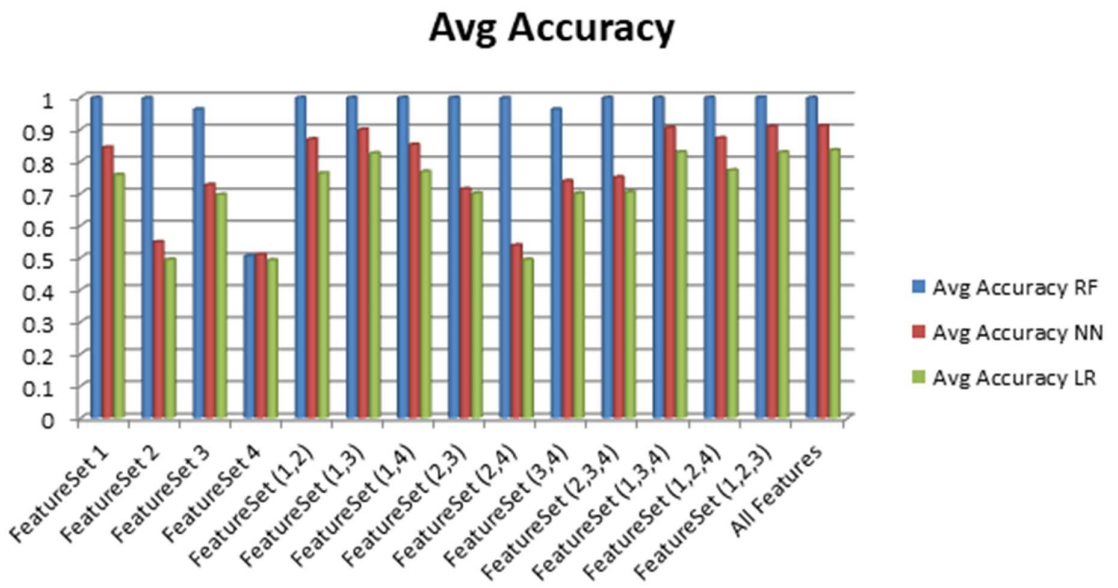

Figure 4: Chart for average accuracy

Form result mentioned in table and chart, it is found that all machine learning models performs well on combination of feature set rather than individual features but the best result for accuracy in all feature set is given by random forest model. It is also found that we should apply combination of feature set to found better results.

5. Conclusion

The botnet is one of the most dangerous attacks in the network amongst the various existing malware. The botnet is analysed depending on its architectural design which is developed by the botmaster. Botmaster uses various types of topologies and tools to make botnets strong and complex. So that it becomes harder to detect the botnet. Botnet attacks are constantly more sophisticated, and this is expected to get even worse with the massive increase of connected objects and virtualised infrastructures. This paper explores how flow-based traffic analysis and supervised machine learning can be used to provide that. We developed a botnet detection system that relies on flow-level network traffic analysis and supervised MLAs for capturing patterns of malicious botnet traffic. The developed method achieves more accuracy around 99.8%. The future work shall be devoted in analysis and identification of new and effective features with respect to periodic flows. There is a scope in development of more improvised methods that also include other parameters. Beside accuracy other parameters like precision and recall should also be evaluated.

References

[1] Kaur, N. and Singh, M., 2016, August. Botnet and botnet detection techniques in cyber realm. In 2016 International Conference on Inventive Computation Technologies (ICICT) (Vol. 3, pp. 1-7). IEEE 2016.

[2] Liu, Jing, Yang Xiao, Kaveh Ghaboosi, Hongmei Deng, and Jingyuan Zhang. "Botnet: classification, attacks, detection, tracing, and preventive measures." EURASIP journal on wireless communications and networking 2009, no. 1 (2009): 692654.

[3] Zeidanloo, Hossein Rouhani, Azizah BT Manaf, Payam Vahdani, Farzaneh Tabatabaei, and Mazdak Zamani. "Botnet detection based on traffic monitoring." In 2010 International Conference on Networking and Information Technology, pp. 97-101. IEEE, 2010.

[4] H. Lin, J. Hu, W. Xiaoding, M. F. Alhamid and M. J. Piran, “Towards secure data fusion in industrial IoT using transfer learning,” *IEEE Transactions on Industrial Informatics*, pp. 1–9, 2020, (Early Access).

[5] A. Rehman, S. U. Rehman, M. Khan, M. Alazab and G. T. R., “CANintelliIDS: Detecting in-vehicle intrusion attacks on a controller area network using CNN and attention-based GRU,” *IEEE Transactions on Network Science and Engineering*, pp. 1–11, 2021, (Early Access).

[6] H. F. Atlam and G. B. Wills, "IoT security, privacy, safety and ethics," *In Digital Twin Technologies and Smart Cities*, 1st ed., Switzerland, Springer Nature Switzerland AG 2020, chp. 8, pp. 123–149, 2020.

[7] S. Hameed, F. Idris Khan and B. Hameed, "Understanding security requirements and challenges in internet of things (IoTA review," *Journal of Computer Networks and Communications*, vol. *2019*, pp. 1–14, 2019.

[8] S. Anwar, J. Mohamad Zain, M. F. Zolkipli, Z. Inayat, S. Khan et al., "From intrusion detection to an intrusion response system: Fundamentals, requirements, and future directions," *Algorithms*, vol. *10*, no. *2*, pp. 1–24, 2017.

[9] M. Letafati, A. Kuhestani, K. K. Wong and M. J. Piran, "A lightweight secure and resilient transmission scheme for the internet of things in the presence of a hostile jammer," *IEEE Internet of Things Journal*, vol. *8*, no. *6*, pp. 4373–4388, 2021.

[10] S. Suganth and D. Usha, "A survey of intrusion detection system in IoT devices.," *International Journal of Advanced Research*, vol. *6*, pp. 23–30, 2018.

[11] G. Srivastava, G. Thippa Reddy, N. Deepa, B. Prabadevi and P. K. Reddy M, "An ensemble model for intrusion detection in the internet of softwarized things," in Adjunct Proc. of the 2021 Int. Conf. on Distributed Computing and Networking, New York, NY, USA, pp. 25–30, 2021.

[12] R. M. Swarna Priya, P. K. R. Maddikunta, M. Parimala, S. Koppu, T. R. Gadekallu et al., "An effective feature engineering for DNN using hybrid PCA-gWO for intrusion detection in IoMT architecture," *Computer Communications*, vol. *160*, pp. 139–149, 2020.

[13] Feily, Maryam, Alireza Shahrestani, and Sureswaran Ramadass. "A survey of botnet and botnet detection." In 2009 Third International Conference on Emerging Security Information, Systems and Technologies, pp. 268-273. IEEE, 2009.

[14] W. N. H. Ibrahim *et al.*(2021), "Multilayer Framework for Botnet Detection Using Machine Learning Algorithms," in *IEEE Access*, vol. 9, pp. 48753-48768, 2021, doi: 10.1109/ACCESS.2021.3060778.

[15] F. Fernández-Peña and A. Zurita-Amores (2019), "Botnets Detection in DNS logs using machine learning," *IEEE 14th Iberian Conference on Information Systems and Technologies (CISTI)*, 2019, pp. 1-5, doi: 10.23919/CISTI.2019.8760760.

[16] Dange, S., Chatterjee, M. (2020). IoT Botnet: The Largest Threat to the IoT Network. In: Jain, L., Tsihrintzis, G., Balas, V., Sharma, D. (eds) Data Communication and Networks. Advances in Intelligent Systems and Computing, vol 1049. Springer, Singapore. https://doi.org/10.1007/978-981-15-0132-6_10

[17] Ashraf, Javed & Moustafa, Nour & Khurshid, Hasnat & Debie, Essam & Haider, Waqas & Wahab, Abdul. (2020). A Review of Intrusion Detection Systems Using Machine and Deep Learning in Internet of Things: Challenges, Solutions and Future Directions. Electronics. 9. 10.3390/electronics9071177.

[18] A. Muhammad, M. Asad and A. R. Javed (2020), "Robust Early Stage Botnet Detection using Machine Learning," *IEEE International Conference on Cyber Warfare and Security (ICCWS)*, 2020, pp. 1-6, doi: 10.1109/ICCWS48432.2020.9292395.

[19] Khalid Alissa, Tahir Alyas, Kashif Zafar, Qaiser Abbas, Nadia Tabassum, Shadman Sakib, "Botnet Attack Detection in IoT Using Machine Learning", Computational Intelligence and Neuroscience, vol. 2022, Article ID 4515642, 14 pages, 2022. https://doi.org/10.1155/2022/4515642

[20] De Caldas Filho FL, Soares SCM, Oroski E, de Oliveira Albuquerque R, da Mata RZA, de Mendonça FLL, de Sousa Júnior RT. Botnet Detection and Mitigation Model for IoT Networks Using Federated Learning. *Sensors*. 2023; 23(14):6305. https://doi.org/10.3390/s23146305

[21] G. Gu, P. Porras, V. Yegneswaran, and M. Fong, "BotHunter: Detecting malware infection through ids-driven dialog correlation," in Proceedings of the USENIX Security Symposium. USENIX Association, 2007.

[22] G. Gu, J. Zhang, and W. Lee, "BotSniffer: Detecting botnet command and control channels in network traffic," in Proceedings of the Network and Distributed System Security Symposium (NDSS), 2008.

[23] S. García, M. Grill, J. Stiborek, and A. Zunino, "An empirical comparison of botnet detection methods," Computers & Security, vol. 45, pp.100–123, 2014.