Scalable Anomaly Detection in Cybersecurity: A Data Mining Approach Using the CTU-13 Dataset

Halime Sıla ÖZYURT
Faculty of Engineering and Natural Sciences, Biruni University
Computer Engineering
İstanbul, Türkiye
210408006@st.biruni.edu.tr

Damla TEZAL
Faculty of Engineering and Natural Sciences, Biruni University
Computer Engineering
İstanbul, Türkiye
220408031@st.biruni.edu.tr

Buse Emine SÖNMEZ
Faculty of Engineering and Natural Sciences, Biruni University
Computer Engineering
İstanbul, Türkiye
210408026@st.biruni.edu.tr

***Abstract*—The advent of rapid digital transformation, coupled with an increasing dependence on interconnected devices, has underscored the necessity of robust cybersecurity solutions to address the ever-evolving threats posed. This research project utilizes the CTU-13 dataset, a benchmark collection designed to reflect real-world scenarios involving botnets, to investigate sophisticated methodologies for botnet and anomaly detection. By combining approaches from data mining and machine learning, this study employs both supervised and unsupervised learning algorithms to accurately identify botnet traffic. The integration of graph mining techniques for the clustering of communication patterns, along with deep learning models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), significantly augments the detection of nefarious activities within intricate network settings. Experimental evaluations have validated the efficacy of this hybrid approach, achieving an F1 score of 0.985 and a detection accuracy of 99.99%. Through this study, the pivotal importance of the CTU-13 dataset in the development of adaptive and scalable mechanisms for detecting attacks is highlighted. Future endeavors will concentrate on enhancing computational efficiency and broadening the applicability of the framework to a variety of real-world datasets.**

Keywords—CTU-13 dataset, botnet detection, anomaly detection, data mining, deep learning, graph mining, cybersecurity, machine learning.

INTRODUCTION

The advent of digital transformation and interconnected devices has transformed the way we interact with technology. However, this technological advancement has led to an equally important challenge: cybersecurity. As critical infrastructures, businesses and personal devices increasingly rely on interconnected networks, the risk of cyberattacks, data breaches and system vulnerabilities has increased exponentially. Cybersecurity researchers and professionals are constantly looking for innovative approaches to combat these challenges, with data mining and machine learning emerging as undeniable tools in this effort. The development of effective cybersecurity solutions is highly dependent on datasets that accurately represent real-world attack scenarios and network behavior. These datasets are essential for training and testing intrusion detection systems (IDS), anomaly detection algorithms and other security mechanisms.

By providing the necessary context and variety to model complex attack vectors and normal traffic patterns, they enable researchers to detect, explain and mitigate evolving cyber threats.

The CTU-13 dataset, which was collected at the CTU University in the Czech Republic in 2011, contains 13 scenarios of botnet traffic mixed with other types of traffic and background activity. Of many existing cybersecurity datasets, the CTU-13 dataset holds a significant position due to its focus on botnet activity in realistic network environments [4].

In particular, it provides a benchmark for evaluating the performance of attack detection systems in identifying botnet traffic [1].

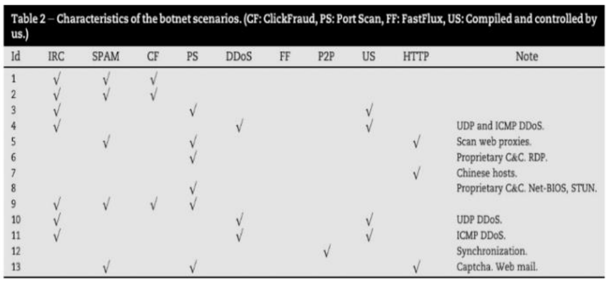

Table 2 – Characteristics of the botnet scenarios. (CF: ClickFraud, PS: Port Scan, FF: FastFlux, US: Compiled and controlled by us.)

Id	IRC	SPAM	CF	PS	DDoS	FF	P2P	US	HTTP	Note
1	√	√	√							
2	√	√	√							
3	√			√				√		
4	√				√			√		UDP and ICMP DDoS.
5		√		√					√	Scan web proxies.
6				√						Proprietary C&C. RDP.
7									√	Chinese hosts.
8				√						Proprietary C&C. Net-BIOS, STUN.
9	√	√	√	√						
10	√				√			√		UDP DDoS.
11	√				√			√		ICMP DDoS.
12							√			Synchronization.
13		√		√					√	Captcha. Web mail.

Table 1.1: Characteristics of botnet scenarios [1]

The importance of CTU-13 lies not only in its comprehensive coverage of botnet behavior, but also in filling the gap between theoretical research and practical application. By providing a detailed representation of network traffic, it allows researchers to explore the application of machine learning and data mining techniques to detect advanced cyberattacks. These techniques, ranging from supervised classification to unsupervised anomaly detection, have shown great promise in improving the accuracy and efficiency of attack detection systems. [2,3].

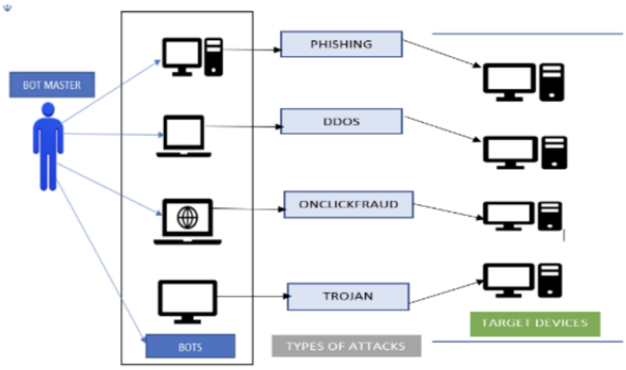

Table 1.2: Botnet Architecture [4]

This report focuses on familiarizing the CTU-13 dataset with its structure and key features and exploring its potential applications in cybersecurity. It describes how the dataset contributes to developing highly adaptive attack detection mechanisms, emphasizing its interest in data mining and machine learning. By exploring the nuances of the dataset, this study aims to communicate its importance in advancing cybersecurity research and reducing the growing cyberattack challenge.

LITERATURE REVIEW

The increase in sophisticated cyber threats, such as botnets and network anomalies, has made traditional detection methods increasingly deficit. Traditional methods like rule-based systems, signature-based detection, and machine learning models, including Random Forest and Support Vector Machines (SVM), struggle to approach the complexities of modern cyberattacks, specifically zero-day threats and localized botnets. These methods, while foundational, deal with obstacles in scalability, adaptability, and false-positive rates, submitting them less effective in dynamic and large-scale network environments. As a consequence, researchers have turned to advanced methodologies, such as graph mining and deep learning, to defeat these limitations. These technologies influence modern computational techniques to enhance the detection of complex and evolving threats, offering a more powerful and efficient alternative for cybersecurity systems.

On the flip side, deep learning techniques have revolutionized the way we detect anomalies in cybersecurity. These approaches, which include Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), excel at dealing with unstructured information like system logs (syslogs) and network traffic (dataflows). The LogHub dataset has played a crucial role in this area, especially for analyzing system log anomalies. Within this dataset, CNNs have achieved an F1 score of 0.999 in cases where they had labeled data, and semi-supervised approaches scored 0.938. Moreover, models based on RNNs, such as DeepLog and LogAnomaly, have proven their effectiveness in analyzing sequential data for anomaly detection by taking advantage of the order of events [6].

In terms of handling data movement across networks, advanced machine learning approaches like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have consistently outperformed conventional methods. By utilizing datasets like NIDS and CTU-13, these algorithms have demonstrated exceptional performance indicators, including an impressive 0.985 F1 score for identifying anomalies in data flow on the NIDS dataset. Hybrid models, such as Convolutional Long Short-Term Memory networks (C-LSTM), integrate both convolutional and recurrent layers to improve the detection of anomalies in sequences. These models are particularly adept at recognizing both the spatial and temporal aspects, rendering them ideal for navigating intricate network scenarios [6] [7].

Research into techniques like Log2Vec and IP2Vec has also been conducted to improve the initial processing of syslog and dataflow data. Log2Vec, drawing inspiration from word2vec, transforms log templates into numerical vectors by considering semantic connections, thereby enhancing the comprehension of context for the purpose of identifying anomalies. Nonetheless, its significant computational expense and variable results underscore the necessity for refinement. In a similar vein, IP2Vec categorizes IPs according to their behavioral likenesses, which boosts anomaly detection but elevates the need for more computational resources. [3].

Even with these progressions, obstacles persist. Datasets such as CTU-13 and LogHub, though commonly utilized, do not completely mirror the intricacies found in actual-world situations. Moreover, the computational demands of deep learning models, especially their reliance on extensive, high-quality labeled datasets, restrict their ability to scale in environments with limited resources. A potential solution lies in hybrid strategies that combine graph mining for preliminary grouping with deep learning for precise categorization. By simplifying the data through clustering, these techniques enable deep learning models to concentrate on areas of higher risk, enhancing both efficiency and the use of resources [5] [6] [7] [8].

"Real-Time Identification of Cyber Threats: An Overview of Supervised Learning Approaches with the CTU-13 dataset" explores the use of supervised learning algorithms for the immediate identification of cyber threats through the CTU-13 dataset. This dataset includes a wide variety of network traffic events, offering a strong basis for creating intrusion detection models that can differentiate between harmless and harmful activities. The study utilizes established supervised learning techniques to improve the accuracy and dependability of intrusion detection systems, with the goal of correctly recognizing and classifying examples of both harmless and harmful activities within network data. [3].

Interconnection of Graph mining and deep learning is changed framework of botnet and anomaly detection, addressing the limitations of conventional methods. Graph mining is great for preprocessing and scalability, while deep learning provides better accuracy and flexibility for data analysis. Future research works on combining these methods, for improving efficiency and testing them to ensure they are scalable and reliable. Altogether, this creates a strong system for dealing with cyber threats. By using these improvements, cybersecurity can better handle the growing challenges of advanced attacks, protecting important systems and networks more accurately and reliably.

References

[1] Stratosphere IPS. (n.d.). CTU-13 dataset. Retrieved November 17, 2024. [Online]. Avaliable: https://www.stratosphereips.org/datasets-ctu13

[2] Impact Cyber Trust. (n.d.). CTU-13 botnet dataset. Retrieved November 17, 2024. [Online] Avaliable: https://impactcybertrust.org/dataset_view?idDataset=945

[3] A. Sharma and H. Babbar, "Detecting cyber threats in real-time: A supervised learning perspective on the CTU-13 dataset," *5th International Conference for Emerging Technology (INCET)*, Karnataka, India, May 24-26, 2024..

[4] S. K. Srinarayani, Dr. B. Padmavathi, and Mrs. D. Kavitha, "Detection of botnet traffic using deep learning approach," *Proceedings of the International Conference on Sustainable Computing and Data Communication Systems (ICSCDS-2023)*, Chennai, India, 2023.

[5] D. J. Borah and A. Sarma, “Detection of Peer-to-Peer Botnets Using Graph Mining,” *International Journal of Computer Networks & Communications (IJCNC)*, vol. 15, no. 2, Mar. 2023.

[6] A. A. Ahmed, W. A. Jabbar, A. S. Sadiq, and H. Patel, “Deep learning-based classification model for botnet attack detection,” *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, no. 10,Feb. 2020.

[7] K. Macková, D. Benk, and M. Šrotýř, “Enhancing Cybersecurity Through Comparative Analysis of Deep Learning Models for Anomaly Detection,” *Proceedings of the 10th International Conference on Information Systems Security and Privacy (ICISSP 2024)*, Prague, Czech Republic, pp. 682-690, 2024

[8] K. Sinha, A. Viswanathan, and J. Bunn, “Tracking Temporal Evolution of Network Activity for Botnet Detection,” Aug. 2019.