PHASE 1 REPORT: CTU-13 BOTNET DATASET EXPLORATION

Name Surname
Nilsu BÜLBÜL (230304055)
Nisa AKSOY (230304047)

Dept. Name Of Organization
Computer Engineering of FBU

Name Of Organization
Data Mining for Cybersecurity
Group-12

This paper presents an exploratory analysis of the CTU-13 botnet dataset, a widely used benchmark in cybersecurity research. The objective of this Phase 1 study is to examine the structure, features, and labeling scheme of the dataset, as well as to review academic research that applies machine learning and anomaly detection methods to CTU-13. The dataset contains 13 scenarios of real network traffic, including botnet, normal, and background activities, making it suitable for intrusion detection and behavioral analysis. The literature survey highlights common techniques, such as Random Forest, LSTM-based models, and anomaly detection algorithms, used in prior work. This report provides the foundation for Phase 2, where data preprocessing, exploratory data analysis (EDA), and machine learning classification will be implemented.

Keywords: CTU-13 dataset, botnet detection, cybersecurity, machine learning, anomaly detection, network traffic analysis.

A. . Introduction

Botnet attacks represent a major threat in cybersecurity due to their ability to control large numbers of infected devices. Detecting such attacks requires high-quality datasets that reflect real-world network behavior. The CTU-13 dataset, developed by the Stratosphere IPS research group at the Czech Technical University, is one of the most widely used datasets for botnet traffic analysis. In this report, we explore the structure, purpose, and features of the CTU-13 dataset. Additionally, we review several academic studies that use this dataset for machine learning and anomaly detection. This Phase 1 report prepares the foundation for Phase 2, where data preprocessing, EDA, and machine learning classification will be conducted.

B. .Dataset Overview

The CTU-13 dataset contains 13 scenarios of real network traffic captured during controlled botnet executions. The traffic includes three main types: normal, background, and botnet traffic. Each scenario represents a unique botnet family or communication method such as IRC, HTTP, or P2P. The dataset is commonly used to evaluate machine learning-based intrusion detection systems.

B.1 Source and Purpose

The CTU-13 dataset was collected in 2011 by the Stratosphere IPS Laboratory at the Czech Technical University. Its purpose is to provide a realistic benchmark dataset for cybersecurity research, especially for detecting botnet behavior in network traffic. The dataset mixes real botnet activity with legitimate traffic, allowing researchers to test detection models under realistic conditions.

B.2 Scenarios

The dataset consists of 13 different scenarios, each produced by executing a unique botnet sample. Scenarios differ in protocol type, number of infected machines, traffic volume, and botnet behavior. Each scenario contains:
- Botnet traffic
- Normal traffic
- Background traffic

This diversity makes CTU-13 useful for both supervised and unsupervised machine learning research.

B.3 Features

The dataset provides bidirectional NetFlow files that include many informative features. Common features include:
- Source and Destination IP
- Source and Destination Ports
- Protocol type
- Flow duration
- Packet counts (forward and backward)
- Byte counts (forward and backward)

These features are suitable for classification and anomaly detection.

B.4 Labels

Each NetFlow record contains a label describing the traffic type:
- Botnet: Traffic generated by infected hosts
- C&C: Command-and-control communication
- Normal: Legitimate traffic
- Background: Unrelated traffic

These labels allow researchers to train and evaluate machine learning models.

B.5 Strengths and Limitations

Strengths:
- Real botnet behavior captured in actual network environments
- Labeled traffic suitable for supervised learning
- Wide variety of botnet behaviors and protocols

Limitations:
- Data collected in 2011; some botnet behavior may be outdated
- Some scenarios highly imbalanced (few botnet flows)
- Mixed full traffic pcap files not publicly available for privacy reasons

C. Literature Survey

This section summarizes academic studies that used the CTU-13 dataset for machine learning-based botnet detection, deep learning approaches, and anomaly detection.

C.1 Machine Learning Based Studies

A study published in 2023 performed a comparative evaluation of algorithms such as Random Forest, SVM, and Naive Bayes on the CTU-13 dataset. The results showed that Random Forest achieved the best performance due to its robustness to noise and its ability to model nonlinear relationships. Feature selection methods were shown to improve performance by reducing redundant attributes.

C.2 Deep Learning Studies

Another study explored the use of Long Short-Term Memory (LSTM) networks to analyze temporal patterns in CTU-13 traffic. The researchers found that LSTM models successfully captured sequential patterns in botnet communication and outperformed traditional ML models in scenarios involving C&C activity. Convolutional Neural Networks (CNNs) were also tested and showed good performance when representing flows as grid-based feature maps.

C.3 Anomaly Detection Studies

Several papers applied unsupervised learning techniques to detect botnet anomalies within CTU-13. Methods such as One-Class SVM, Isolation Forest, and clustering achieved promising results. These studies highlight the challenge of class imbalance in CTU-13 and suggest anomaly detection as a solution for identifying rare botnet patterns.

C.4 Feature Research Gaps

Although CTU-13 is a widely used dataset, there are still open research challenges:

- Developing models capable of real-time botnet detection
- Improving generalization to modern botnet families
- Handling extreme class imbalance
- Combining supervised and unsupervised detection methods

D. Conclusion

This Phase 1 report provided a detailed overview of the CTU-13 dataset along with a review of key academic studies. The dataset remains a strong benchmark for evaluating intrusion detection and botnet classification methods. In Phase 2, we will apply preprocessing, exploratory data analysis (EDA), and machine learning models to perform botnet detection on selected CTU-13 scenarios.

REFERENCES

[1] https://www.stratosphereips.org/datasets-ctu13

[2] https://www.researchgate.net/publication/377653307_Evaluating_ML_models_on_CTU-13_and_IOT-23_Datasets.