Proposal: Analyzing DDoS Attack Patterns for Improved Detection

DATA 450 Capstone

Bijay Adhikari

February 7, 2025

1 Introduction

Distributed Denial of Service (DDoS) attacks are a significant threat to the modern internet. These attacks aim to disrupt the normal traffic of a server or network infrastructure by overwhelming it with a flood of network traffic coming from malicious or compromised systems. These attacks are very frequent and sophisticated at the same time. In 2023, there were more than 13 million DDoS attacks reported globally (NetScout 2024). These attacks can cause long service outages and lead to serious financial losses for businesses in all industries. This project analyzes a public DDoS dataset to identify prevalent attack vectors, differentiate malicious traffic characteristics, and explore attack patterns related to time and protocol. The ultimate goal is to improve our understanding of DDoS attack dynamics and contribute to the development of effective yet simpler cybersecurity strategies for such attacks.

2 Dataset

The dataset I plan to use for this project is the "DDoS 2019" (Sharafaldin et al. 2019) dataset provided by the Canadian Institute for Cybersecurity (CIC) at the University of New Brunswick.

The dataset can be downloaded directly from the CIC website. As per the CIC, the dataset was created in a realistic network environment. The data was captured in CSV format and includes network traffic features extracted from packet captures using the CICFlowMeter tool (Sharafaldin et al. 2019). The dataset contains various types of DDoS attacks as well as benign traffic. The following variables are intended to be used in the analysis:

- Flow Duration: Duration of the network flow (nanoseconds)

- Total Fwd Packets: Total number of packets sent in the forward direction
- Total Backward Packets: Total number of packets sent in the backward direction
- Flow Bytes/s: Number of bytes per second within a flow
- Flow Packets/s: Number of packets per second within a flow
- Protocol: Protocol used (e.g., TCP, UDP, ICMP)
- Packet Length Mean: Average length of packets in a flow
- Label: Classification of the flow (Benign or DDoS type)

3 Data Acquisition and Processing

The data can be obtained by downloading the dataset directly from the CIC's website. The first step in processing will involve exploring the dataset to understand its structure, data types, and to identify any missing values. For the data cleaning, missing values will be checked and handled appropriately. This may involve imputation, such as using the mean or median for numerical features or the mode for categorical features, or the removal of rows with excessive missing data. Once missing values are addressed, the next step will be to ensure that all columns have the correct data types, with numerical features properly formatted and categorical features correctly encoded. The dataset will also be examined for infinite or NaN values. If any additional categorical features beyond 'Protocol' and 'Label' are found relevant, they will be numerically encoded, using techniques like one-hot encoding if necessary.

4 Research Questions and Methodology

1. Which DDoS vectors (UDP flood, HTTP flood, etc.) appear most frequently in the dataset? To answer this, I will filter the dataset for DDoS attacks, and count the occurrences of each DDoS attack vector from the 'Label' column. I will then normalize these counts to percentages and visualize the results using a bar chart, with each bar representing a DDoS vector and the height representing its frequency.

Estimated Time: 2 hours

2. How do packet size, flow duration, and protocol usage differ between DDoS and benign flows? To answer this, I will select features like 'Fwd Packet Length Mean', 'Flow Duration', and 'Protocol'. I will create subsets for benign and DDoS flows. For packet size and flow duration, I will calculate descriptive statistics, visualize distributions using box and violin plots, and perform statistical tests. For 'Protocol', I will calculate and visualize frequency distributions for both flow types using side-by-side bar charts.

Estimated Time: 5 hours

3. Are attacks more likely at certain times of day or under specific network conditions? To answer this, for the time of day, I will extract the ‘hour of day’, group data by hour and calculate the DDoS proportion per hour. I will visualize it with a line chart. For network conditions, if a proxy like ‘Flow Bytes/s’ is available, I will correlate the hourly network load with the hourly DDoS proportion.

Estimated Time: 5 hours

4. Which features (e.g., packet rates, flow bytes/s) might be the strongest indicators of an ongoing DDoS? To answer this, I will select relevant features and train a Random Forest or Logistic Regression classifier to predict DDoS attacks. I will then extract feature importance scores and also calculate the correlation of each feature with the ‘is_DDoS’ label. Lastly, I will rank features by both importance and correlation to identify the strongest indicators.

Estimated Time: 5 hours

5. Does TCP or UDP dominate in these attacks, and does one protocol create significantly different traffic signatures? To answer this, for protocol dominance, I will filter for TCP and UDP flows separately, calculate the DDoS proportion within each protocol group, and visualize it using a bar chart. For traffic signatures, I will compare the distributions of key features for TCP DDoS and UDP DDoS flows using box and violin plots.

Estimated Time: 3 hours

5 Work plan

Week 4 (2/10 - 2/16):

- Data cleaning and merging (4 hours)
- Exploratory Data Analysis (EDA) (3 hours)

Week 5 (2/17 - 2/23):

- Address missing values and normalize data (3 hours)
- Answer Question 1(4 hours)

Week 6 (2/24 - 3/2):

- Answer Question 2 (5 hours)
- Begin Question 3 (5 hours)

Week 7 (3/3 - 3/9):

- Complete Question 3 (3 hours)

- Start building the Random Forest model for Question 4 (4 hours)
- Presentation prep and practice (4 hours)

Week 8 (3/10 - 3/16): *Presentations given on Wed-Thu 3/12-3/13.*

- Poster prep (4 hours)
- Presentation peer review (1.5 hours)

Week 9 (3/24 - 3/30): *Poster Draft 1 due Monday morning 3/24 at 9am. Poster Draft 2 due Sunday night 3/30.*

- Peer feedback (2 hours)
- Poster revisions (1.5 hours)

Week 10 (3/31 - 4/6): *Final Poster due Sunday 4/6.*

- Peer feedback (1.5 hours)
- Poster revisions (2 hours)

Week 11 (4/7 - 4/13):

- Validate models with cross-validation (2 hours).

Week 12 (4/14 - 4/20):

- Write blog post methods/results sections (4 hours).

Week 13 (4/21 - 4/27): *Blog post draft 1 due Sunday night 4/28.*

- Finalize blog post (4 hours)

Week 14 (4/28 - 5/4):

- Peer feedback (3 hours)
- Blog post revisions (4 hours)

Week 15 (5/5 - 5/8): *Final blog post due Tues 5/7. Blog post read-throughs during final exam slot, Thursday May 8th, 8:00-11:20am.*

- Blog post revisions (2 hours)
- Peer feedback (2 hours)

References

NetScout. 2024. “2H 2023 DDoS Threat Intelligence Report.” NETSCOUT Systems.

Sharafaldin, I., A. H. Lashkari, A. A. Ghorbani, and M. Zincirkiran. 2019. “A Detailed Investigation of the Datasets Used in Evaluating Network Intrusion Detection Systems.” *Online*.