Network Traffic Analysis Report: A Comprehensive Study

Your Name

April 6, 2025

Abstract

This report provides an in-depth analysis of a network traffic dataset comprising 2,071,657 flows collected from June 11 to June 17, 2010, across six CSV files combined into a single DataFrame. I used the combined data just because it is easy to confirm the attacks on span of days, like seeing the activity in free hours like at night. Using Python tools such as Pandas, NumPy, Matplotlib, and probabilistic data structures (HyperLogLog, Count-Min Sketch, Bloom Filter), the study examines traffic patterns, estimates unique IPs, identifies heavy hitters, and detects anomalies and suspicious behaviors. Key findings include TCP/IP dominance (79.4%), significant traffic from IPs like 192.168.5.122, and evidence of potential security threats such as stealthy port scans, slow DDoS attacks, and IP hopping. The report balances exact and approximate methods to offer insights into network performance and security, culminating in a risk assessment to guide mitigation efforts.

1 Introduction

Network traffic analysis is critical in today's digital landscape to monitor performance, ensure security, and detect malicious activities. This report analyzes a dataset of 2,071,657 network flows recorded over a week in June 2010, sourced from six files (`network_analysis_data1.csv` to `network_analysis_data6.csv`). The objectives are to:

1. Characterize traffic patterns, including protocols, active IPs, and packet sizes.
2. Employ probabilistic data structures to estimate unique IPs and identify frequently contacted destinations efficiently.
3. Detect anomalies (e.g., traffic spikes, unusual packet sizes) and suspicious patterns (e.g., port scans, DDoS, IP hopping).

The analysis leverages both exact computations and approximate algorithms to handle the dataset's scale, providing a foundation for network management and threat detection, with a final risk assessment to evaluate identified threats.

2 Methodology

The analysis was conducted in a Jupyter Notebook using Python 3.12. The methodology is divided into three tasks, each addressing specific aspects of the dataset.

2.1 Data Preprocessing

Six CSV files were loaded and concatenated into a single Pandas DataFrame (`df`) with 2,071,657 rows. Key preprocessing steps included:

- Converting `startDateTime` and `stopDateTime` to datetime objects using `pd.to_datetime`.
- Deriving columns like `avg_packet_size` (`totalSourceBytes / totalSourcePackets`) and `duration` (`stopDateTime - startDateTime`).
- Handling division-by-zero errors by replacing with NaN where applicable.

2.2 Task 1: Traffic Characterization

Basic statistics were computed:

- Total flows: `len(df)`.
- Top protocols: `df[’protocolName’].value_counts().head(5)`.
- Most active IPs: `df[’source’].value_counts().head(10)` and `df[’destination’].value_cou`
- Average packet size: `df[’totalSourceBytes’].mean() / df[’totalSourcePackets’].mean()`
- Traffic volume over time: Grouped by hourly windows (`df[’startDateTime’].dt.floor(’h’)`).

Anomalies were identified using a threshold of mean + 2 standard deviations.

2.3 Task 2: Probabilistic Data Structures

Three probabilistic methods were applied:

- **HyperLogLog (HLL)**: Estimated unique IPs with 16-bit registers, compared against exact counts.
- **Count-Min Sketch (CMS)**: Identified top 10 destination IPs with a 5x1000 table, validated against exact counts.
- **Bloom Filter**: Tested IP membership with a 250,000-bit filter and 5 hash functions, measuring false positives.

Performance metrics (time, memory, error rates) were recorded.

2.4 Task 3: Anomaly and Pattern Detection

Anomalies and suspicious behaviors were detected using statistical and heuristic approaches:

- **Packet Size Anomalies**: Used 99th percentile threshold.
- **Flow Duration Anomalies**: Applied mean + 2σ and Median Absolute Deviation (MAD).
- **Stealthy Port Scans**: Aggregated by source-destination pairs, flagged long durations.

- **Slow DDoS**: Identified sustained moderate traffic with high source counts.
- **IP Hopping**: Clustered sources by shared destinations and temporal overlap.
- **Encrypted Traffic**: Calculated payload entropy, flagged high-entropy flows on non-standard ports.
- **C&C Patterns**: Detected low-byte, high-flow communications.

3 Results

3.1 Task 1: Network Traffic Characterization

The dataset revealed extensive network activity:

- **Total Flows**: 2,071,657, indicating a high-traffic network over the week.
- **Top 5 Protocols**: TCP/IP dominates, reflecting typical internet traffic.

Protocol	Count	Percentage
tcp_ip	1,644,056	79.36%
udp_ip	419,246	20.23%
icmp_ip	8,211	0.40%
igmp	77	0.004%
ip	66	0.003%

Table 1: Top 5 Protocols by Flow Count

- **Top 10 Source IPs**:
- **Top 10 Destination IPs**:
- **Average Packet Size**: 124.04 bytes, with a variance of 1,172.25 bytes^2, suggesting moderate variability.
- **Most Common Pair**: 192.168.5.122 → 198.164.30.2 (232,409 flows), a critical communication channel.

Traffic Spikes: Three significant anomalies were detected:

- June 14, 22:00: 793,893,427 bytes
- June 15, 02:00: 784,152,020 bytes
- June 15, 16:00: 388,361,383 bytes

These exceed the mean (34,919,444.25 bytes) + 2σ (96,589,857.57 bytes), indicating potential incidents.

Source IP	Flows
192.168.5.122	268,267
192.168.2.107	208,379
192.168.4.118	135,374
192.168.1.101	116,292
192.168.4.121	105,454
192.168.1.105	101,359
192.168.2.109	99,183
192.168.3.116	97,241
192.168.2.110	90,658
192.168.3.115	88,915

Table 2: Top 10 Most Active Source IPs

Destination IP	Flows
198.164.30.2	232,409
192.168.5.122	199,437
203.73.24.75	193,200
125.6.164.51	106,826
67.220.214.50	49,298
202.210.143.140	36,189
82.98.86.183	25,214
95.211.98.12	25,095
209.112.44.10	21,824
62.140.213.243	20,509

Table 3: Top 10 Most Active Destination IPs

3.2 Task 2: Probabilistic Data Structures

- **HyperLogLog**:
  - Exact unique IPs: 34,801
  - Estimated: 34,817
  - Error: 0.05% (well within 10%)
  - Time: HLL: 17.5351s; Exact: 0.4915s
  - Memory: HLL: 80 bytes; Exact: 2,158,365 bytes

  HLL offers memory efficiency at the cost of computation time.

- **Count-Min Sketch**: Average error: 0.52%; CMS time: 15.1453s; Memory: 140,168 bytes; Exact: 0.3133s; Memory: 3,795,418 bytes.

- **Bloom Filter**:
  - Size: 250,000 bits
  - Exact IPs: 33,589

Destination IP	Exact Count	CMS Count	Error (%)
198.164.30.2	232,409	232,788	0.16
192.168.5.122	199,437	199,648	0.11
203.73.24.75	193,200	193,302	0.05
125.6.164.51	106,826	107,527	0.66
67.220.214.50	49,298	49,355	0.12
202.210.143.140	36,189	36,353	0.45
82.98.86.183	25,214	25,362	0.59
95.211.98.12	25,095	25,219	0.49
209.112.44.10	21,824	22,041	0.99
62.140.213.243	20,509	20,842	1.62

Table 4: Top 10 Heavy Hitters: Exact vs. CMS

- False positives: 30/1,212 (2.48%)
- Bits set: 122,513

3.3 Task 3: Anomaly and Suspicious Pattern Detection

- **Protocol Distribution Anamoly**:
  - Mean: 849706476.67 flows; Std: 2005153490.22 flows
  - Threshold (MAD): 852136539.67 flows
  - Protocol Anomalies:
    tcp/ip 4940765626

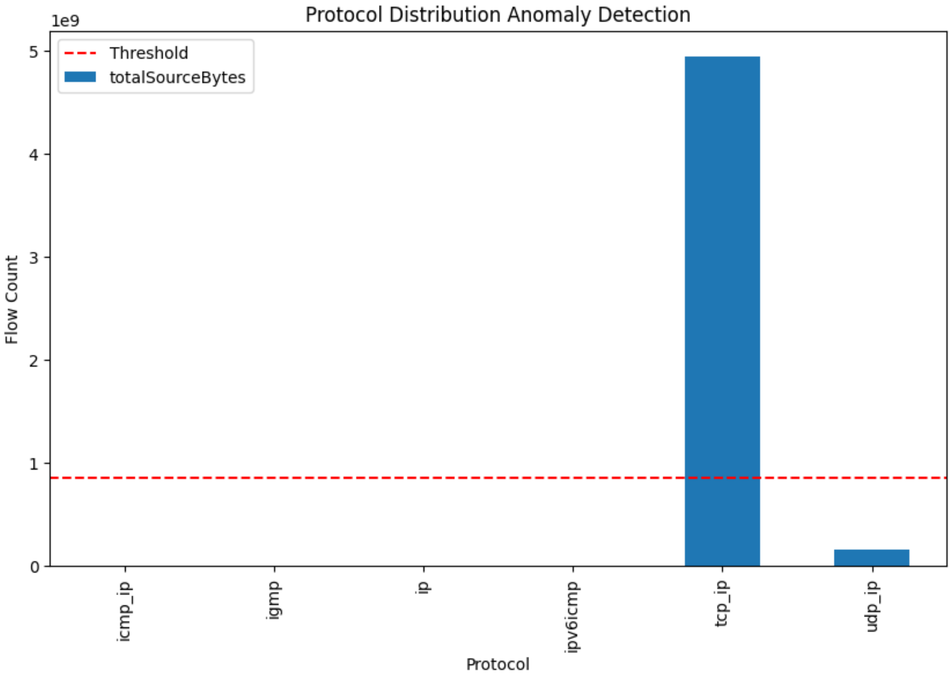

Figure 1: Graph of anamolies detection based on Protocol Distribution.

- **Packet Size Anomalies**:
  - Mean: 85.06 bytes; Std: 34.24 bytes
  - Threshold (99th percentile): 247.00 bytes
  - Anomalies: 6,512 flows
  - Sample: 192.168.2.113 → 192.168.2.255 (251.5 bytes), 192.168.5.123 → 192.168.5.255 (259.0 bytes)

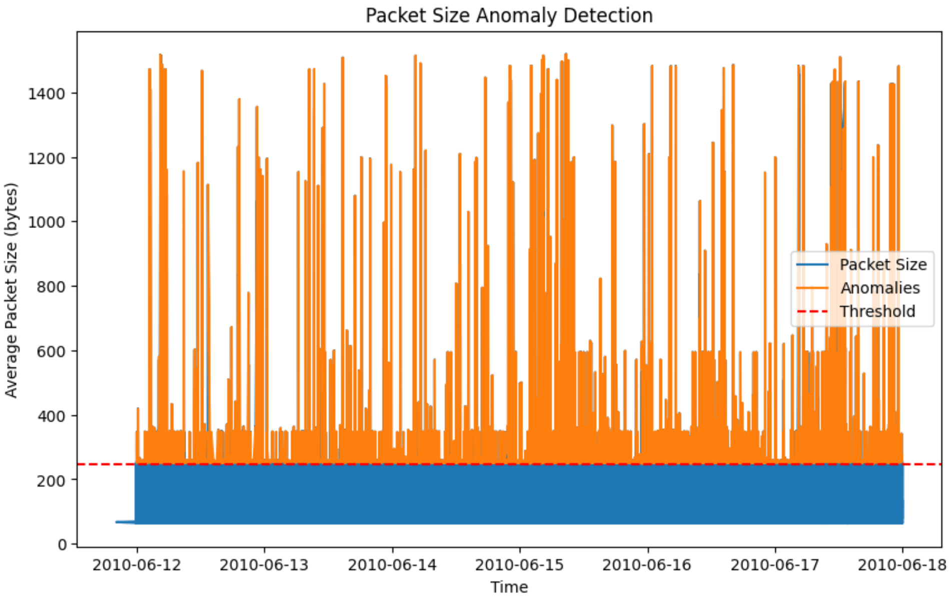

Figure 2: Graph of anamolies detection based on packet sizes.

- **Flow Count Anomalies**:
  - Hourly Mean: 34,919,444.25 bytes; Std: 96,589,857.57 bytes
  - Threshold: 131,509,301.81 bytes
  - Anomalies: June 12, 04:00 (133,409,213 bytes), June 14, 22:00 (793,893,427 bytes)
- **Stealthy Port Scans**: 51 candidates:
  - Threshold: 1,288.53s
  - Examples: 192.168.1.101 → 192.168.1.255 (2,416,500s), 192.168.2.109 → 192.168.2.255 (2,352,780s)
- **Slow DDoS**: 144 candidates:
  - Example: 192.168.4.121 (452,829 bytes, 2,078,820s, 6 sources)
- **IP Hopping**: 2 groups:
  - Group 0: 9 IPs (e.g., 192.168.5.122, 192.168.1.101)
  - Group 1: 21 IPs (e.g., 192.168.3.114, 192.168.2.107)

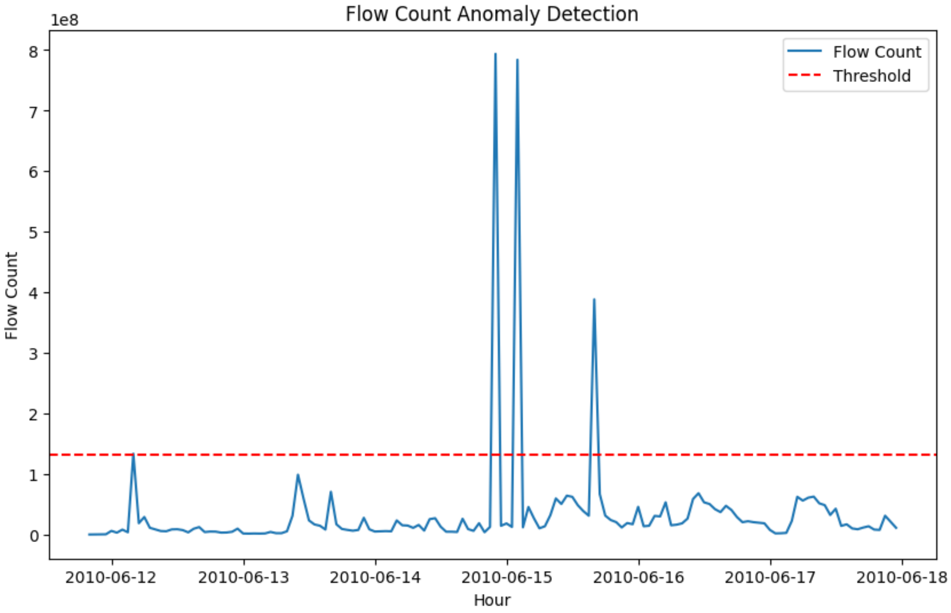

Figure 3: Graph of anamolies detection based on flow counts.

- **High Entropy Flows**: 1,013 flows:
  - Example: 192.168.2.112 → 131.202.243.84 (src: 7.54, dst: 5.01)
- **Suspicious Encrypted Traffic**: 384 flows:
  - Example: 192.168.2.107 → 95.79.144.25 (port 19785, dst: 7.55)
- **C&C Patterns**: 3,985 pairs:
  - Example: 109.93.202.76 → 192.168.2.107 (304.71 bytes/flow, 4.12 packets/flow)

4 Discussion

The dominance of TCP/IP (79.4%) aligns with its role as the backbone of internet communication, while UDP/IP (20.23%) supports real-time applications. IPs like 192.168.5.122 and 198.164.30.2 are central to the network, appearing as both sources and destinations, suggesting they may be servers or critical nodes. The average packet size (124.04 bytes) indicates typical small-packet traffic, but the high variance (1,172.25 bytes^2) points to occasional large transfers.

Probabilistic methods proved effective: HLL's 0.05% error demonstrates its accuracy for cardinality estimation, though its 17.5s runtime versus 0.5s for exact counts highlights a trade-off. CMS's 0.52% error for heavy hitters is impressive, but its memory efficiency (140KB vs. 3.8MB) comes at a 15s computational cost. The Bloom Filter's 2.48% false positive rate is acceptable for blocklist checks, balancing speed and accuracy.

Anomaly detection revealed actionable insights. Traffic spikes (e.g., 793M bytes on June 14) could indicate DDoS attacks or data exfiltration, while packet size anomalies (e.g., 259 bytes) suggest unusual payloads. Stealthy port scans (e.g., 2.4M seconds) and slow DDoS candidates (e.g., 2M seconds duration) imply reconnaissance or persistent

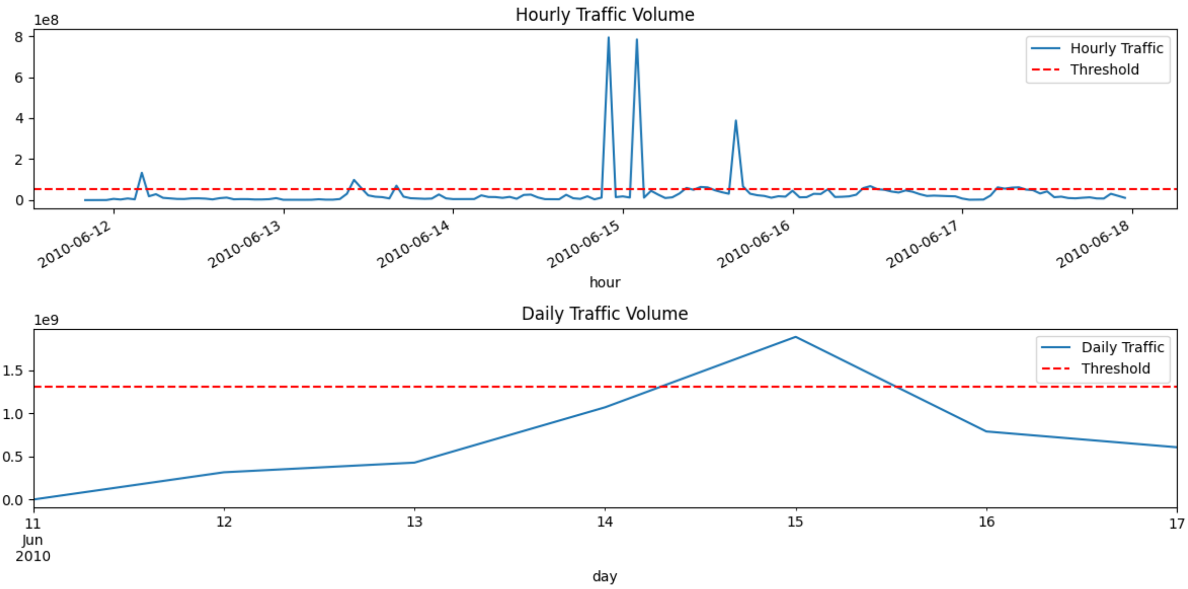

Figure 4: Graph of traffic flow in hours vs Days

attacks. IP hopping groups and high-entropy flows (e.g., entropy ¿ 7.5) suggest evasion tactics or encrypted malware, especially on non-standard ports.

Limitations include potential OCR errors in the dataset, lack of explicit port fields in some analyses, and reliance on static thresholds. Future work could incorporate dynamic machine learning models, additional contextual data (e.g., geolocation), and real-time monitoring capabilities.

5 Risk Assessment Report

This section evaluates the risks identified in the network traffic analysis, assessing their likelihood, impact, and recommended mitigation strategies. Risks are derived from anomalies and suspicious patterns detected in Task 3.

5.1 Risk Identification and Assessment

1. **Traffic Volume Spikes**

   - *Description*: Three significant spikes (e.g., 793,893,427 bytes on June 14, 22:00) exceed the threshold of 131,509,301.81 bytes.
   - *Likelihood*: High – Multiple occurrences within a week suggest recurring events.
   - *Impact*: Severe – Could overwhelm network resources, disrupt services, or indicate data exfiltration.
   - *Risk Level*: High

2. **Packet Size Anomalies**

   - *Description*: 6,512 flows with packet sizes ¿ 247.00 bytes (e.g., 192.168.5.123 → 192.168.5.255, 259.0 bytes).

- *Likelihood*: Moderate – Represents 0.31% of flows but indicates targeted activity.
- *Impact*: Moderate – May suggest unusual data transfers or malware payloads.
- *Risk Level*: Moderate

3. **Stealthy Port Scans**
    - *Description*: 51 source-destination pairs with durations ¿ 1,288.53s (e.g., 192.168.1.101 → 192.168.1.255, 2,416,500s).
    - *Likelihood*: Moderate – Infrequent but deliberate reconnaissance attempts.
    - *Impact*: High – Precursor to targeted attacks, potentially compromising network security.
    - *Risk Level*: High

4. **Slow DDoS Attacks**
    - *Description*: 144 destinations with sustained moderate traffic (e.g., 192.168.4.121, 2,078,820s duration).
    - *Likelihood*: Moderate – Persistent activity across multiple targets.
    - *Impact*: High – Degrades service availability over time, difficult to detect.
    - *Risk Level*: High

5. **IP Hopping Behavior**
    - *Description*: 2 groups with shared destinations (e.g., Group 1: 21 IPs including 192.168.3.114).
    - *Likelihood*: Low – Complex coordination required, but detected instances are significant.
    - *Impact*: High – Indicates evasion tactics, potentially masking malicious activity.
    - *Risk Level*: Moderate to High

6. **Suspicious Encrypted Traffic**
    - *Description*: 384 flows with high entropy (¿7.5) on non-standard ports (e.g., 192.168.2.107 → 95.79.144.25, port 19785).
    - *Likelihood*: Moderate – 0.02% of flows, but concentrated on unusual ports.
    - *Impact*: High – Possible encrypted malware or tunneling, bypassing standard security.
    - *Risk Level*: High

7. **Command and Control (C&C) Communication**
    - *Description*: 3,985 pairs with low bytes/flow (e.g., 109.93.202.76 → 192.168.2.107, 304.71 bytes/flow).
    - *Likelihood*: High – Common pattern in malware communication.
    - *Impact*: Severe – Could enable botnet control or data theft.
    - *Risk Level*: High

5.2 Risk Summary

Risk	Likelihood	Impact	Risk Level
Traffic Volume Spikes	High	Severe	High
Packet Size Anomalies	Moderate	Moderate	Moderate
Stealthy Port Scans	Moderate	High	High
Slow DDoS Attacks	Moderate	High	High
IP Hopping Behavior	Low	High	Moderate to High
Suspicious Encrypted Traffic	Moderate	High	High
C&C Communication	High	Severe	High

Table 5: Summary of Identified Risks

5.3 Mitigation Recommendations

1. **Traffic Volume Spikes**:
   - Deploy rate-limiting and traffic filtering at key nodes (e.g., 192.168.5.122).
   - Investigate logs for June 14–15 to identify sources and payloads.
2. **Packet Size Anomalies**:
   - Monitor flows exceeding 247 bytes for content analysis.
   - Implement packet inspection tools to detect anomalies in real-time.
3. **Stealthy Port Scans**:
   - Block IPs exhibiting prolonged scan behavior (e.g., 192.168.1.101).
   - Enhance firewall rules to limit port access.
4. **Slow DDoS Attacks**:
   - Use anomaly detection systems to flag sustained moderate traffic.
   - Isolate affected destinations (e.g., 192.168.4.121) for forensic analysis.
5. **IP Hopping Behavior**:
   - Track IP groups (e.g., Group 1) using behavioral analytics.
   - Correlate with external threat intelligence for known hopping patterns.
6. **Suspicious Encrypted Traffic**:
   - Inspect non-standard ports (e.g., 19785, 5555) with deep packet inspection.
   - Quarantine high-entropy flows for malware analysis.
7. **C&C Communication**:
   - Block IPs with C&C patterns (e.g., 109.93.202.76) at the perimeter.
   - Deploy endpoint detection to identify infected hosts (e.g., 192.168.2.107).

6 Conclusion

This study provides a detailed portrait of a week-long network traffic dataset, revealing both operational patterns and security concerns. TCP/IP's prevalence, key IPs like 192.168.5.122, and probabilistic method efficacy underscore the network's structure and analytical feasibility. Detected anomalies and suspicious patterns—spikes, scans, and encrypted flows—necessitate immediate investigation to mitigate risks, as detailed in the risk assessment. This report demonstrates the power of combining statistical and probabilistic techniques for scalable, insightful network analysis, with actionable recommendations to enhance security.