Classification of Internet Traffic

Alok Shriram

Need for Classification

- Classification required
  - To isolate traffic of interest
  - To treat special types of traffic in a different manner
- Some types of classification already seen in AI learning systems.
- Some types of classification seen in Data mining.

Three Techniques

- A Framework for Classifying Denial of Service Attacks (Single or Multiple Source Attacks)
- Identification of Repeated Attacks Using Network Traffic Forensics.
- Class of Service Mapping for QoS.

Identification of Repeated Attacks Using Network Traffic Forensics

- To Identify repeated attacks
- Forensic evidence used to investigate and establish facts
- Depending on Intent attackers punishment is decided

Objective

- Build an attack fingerprinting system
- Make this system of creating fingerprints automatic
  - Fingerprint is any characteristic feature of an attack which can uniquely identify it.
- Automatic matching system
- Identify repeated attacks

Methodology in a Nutshell

- Given an attack scenario
  - Figure out if attack has occurred previously.
- For this we filter attack
- Create attack fingerprint
- Compare attack to previously fingerprinted attack
Creating Attack Fingerprint

- Convert packet trace into time series
- Consider interval of time $p$
  - Packet arrivals $[t, t + p)$
- For $T$ second trace $T/p$ samples
- Max frequency $1/2p$ Hz
- Use $p=1$ msec and attack segment length =2 s

Creating Attack Fingerprint (1)

- Thus we have time series $x(t)$.
- Compute autocorrelation function (ACF) of time series
- Compute ACF for different values of $L$ to get $r_k(L)$
- Compute FFT of $r_k(L)$
  - Periodicity shows up as dominant frequency.

Creating the fingerprint (3)

- $F_a$ consists of all segment fingerprints $X_k$
- Use $F_a$ to compute digest
  - $M_a$ = mean of $X_k$
  - $C_a$ = covariance of $X_k$
- $N_a/|X_k| \geq 10$
  - Thus $N_a=20$
Comparing Fingerprints(1)

- Use a comparator to match similarity
- Bayes ML classifier
  - Assumptions
    - Spectral profiles normal w.r.t dominant frequency
    - Each scenario equally likely
    - Attacks are independent

Comparing Fingerprints(2)

- With each attack we just need some information to compare each segment against signature
- Quantify separation between current attack and signatures

Analyzing the results

- Low$_{CA}$ 5% quartile indicate the at least 5% match very accurately
- 95%-5% small range of this indicates precision.

Experimental Results (1)
Experiments and Results (2)

A Framework for Classifying Denial of Service Attacks

- Denial Of Service Attacks are of two types
  - Single Source
  - Multiple Source
- Identifying the number of sources helps in mitigation strategies

Objective

- Develop framework to classify attacks as single or multiple source
  - Use Ramp up behavior
  - Port numbers
  - Spectral Characteristics of attack traffic
- Spectral content cannot be spoofed
- Could be used in DOS detection and response systems

Two Types of Attacks

- Software Attacks
- Flooding Attacks
  - Single Source
  - Multiple Source
  - Reflector Attacks

Classifying Attacks

- Three Methods that are used for classification
  - Header Content
  - Ramp-up Behavior
  - Spectral Characteristics

Header Content

- Use fragment ID field and TTL field
  - Single hosts monotonically increasing
  - Multiple Hosts
    - Many ID sequences
    - Two sequence considered unique if they have an IDgap >16
    - ID gap is there to tolerate moderate packet reordering.
**Ramp-up Behavior**

- Single sources don’t exhibit a ramp-up behaviour
- Multiple source with large number of processes
  - Exhibit ramp up behavior
  - Clock and RTT skews cause gradual buildup
  - By observing this we can guess the number of sources.

**Spectral Analysis**

- Stuff about spectra analysis here from previous slides.

**Experiments: Packet Header Analysis**

<table>
<thead>
<tr>
<th>Attack Class</th>
<th>Attack</th>
<th>Range (packet/s)</th>
<th>Range (bits)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-source</td>
<td>37</td>
<td>100-1500</td>
<td>2500-5000</td>
</tr>
<tr>
<td>Multi-source</td>
<td>10</td>
<td>1000-8000</td>
<td>1700-10000</td>
</tr>
<tr>
<td>Redirected</td>
<td>20</td>
<td>1000-2000</td>
<td>3000-5000</td>
</tr>
<tr>
<td>Untrusted</td>
<td>11</td>
<td>4000-6500</td>
<td>12000-20000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Protocol</th>
<th>Packet Type</th>
<th>Attack Class</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCP</td>
<td>ACK</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>SYN/ACK</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>M</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>U</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ICMP</td>
<td></td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>F</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UDP</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Number of packets in each class based on header analysis.

<table>
<thead>
<tr>
<th>Protocol</th>
<th>Packet Type</th>
<th>Attack Class</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCP</td>
<td>ACK</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>SYN/ACK</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>M</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>U</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ICMP</td>
<td></td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>F</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UDP</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Detailed analysis of packet headers. 1 indicates single-source, 2 indicates multi-source, 3 indicates distributed sources, and 7 indicates unknown attacks. The number in parenthesis indicates attacks terminating within our ISP while the first number indicates total attacks.

**Experiments: Arrival Rate Analysis**

- (a) Attack intensity in packets/s and Ki/s
- (b) Attack intensity in packets for each class

**Experiments: Ramp Up Behavior Analysis**

- (a) Multiple source addresses observed in attack
- (b) Subset spoofed source addresses

Figure 5: Due to lack of synchronization among the zombies, multi-source attacks exhibit initial ramp-up behavior.
Experiments: Spectral Content Analysis

- Single Source Dominant high frequencies
- Multi Source attacks Dominant low Frequencies

Experiment: Explanation

How do two sources combine to form lower frequency??

Class of Service Mapping for QoS

- Support different applications
- With different quality demands
- Concept has been around for some time
  - What ails QoS?
    - The ability to identify types of traffic

Objective

- Develop a signature based classification framework
- Class of Service to Traffic mapping problem
- How to choose statistics that accurately represent traffic behavior.

Traffic Classification (In the dark ages)

- Based on Port Numbers
- These techniques had several limitations
  - More than one application using the same port
  - P2P does not use any standardized ports.
  - Some applications tunnel through other application ports
  - Different ports used to circumvent control.
Implementing CoS Mapping

- Three Stage process
  - Statistics Collection
  - Classification
  - Rule Creation

Statistics Collection

- Place monitors and collect network stats
- Need to collect aggregate stats
- Form a vector of statistics
- Ideally statistics should be updatable recursively or in an online manner.

Instance of recursive Classification

1. **average:**
   \[ \bar{x}_{j+1} = \frac{1}{j+1}x_{j+1} + \frac{j}{j+1}\bar{x}_j, \]

2. **variance:**
   \[ \text{var} (X_{j+1}) = \frac{1}{j}X_{j+1}^2 + \frac{j-1}{j} \text{var} (X_j) + \frac{j}{j-1}X_j^2 - \frac{j+1}{j}\bar{x}_{j+1}^2. \]

Classification

- Now we have a collection of statistics indexed by aggregate
- Use classification algorithm to classify traffic
- This classification can have a direct quality mapping

What type of traffic can there be?

- Interactive -> Real time interaction.
- Streaming -> Multimedia with RT constraints.
- Bulk Data Transfers-> Large volumes of data over the internet.
- Transactional-> Small volumes of traffic.

What statistics can we collect

- Packet Level features
  - Mean Packet Size
  - RMS size
- Flow Summaries
  - Mean flow duration
  - Mean data volume
What statistics can we collect

- **Connection Level**
  - Track Connection level Characteristics
  - Symmetry of connection
  - Advertised window size
- **Intra-flow**
  - IAT between packets
- **Multi Flow**
  - Features across different flows.

Classification methods

- Two methods of classification
  - Linear Discriminant Analysis (LDA)
  - Nearest Neighbor (NN)

Given \( k \) classes \( m \) features and \( n \) training data points
- Can we classify traffic into characteristics types?

Simple Classification Results

![Simple Classification Results](image1.png)

Streaming vs. Data

![Streaming vs. Data](image2.png)

Temporal Difference

![Temporal Difference](image3.png)

What does this have to do with NIDS?

- If we can classify traffic as the DOS type traffic
- Provide QoS of zero to it.
  - Basically means deny service to that traffic
The END