Towards Reliable Traffic Classification Using Visual Motifs

Wilson Lian\textsuperscript{1} John McHugh\textsuperscript{1,2} Fabian Monrose\textsuperscript{1}

\textsuperscript{1}University of North Carolina at Chapel Hill
\textsuperscript{2}RedJack, LLC

FloCon 2010
Overview

- Background
- Visual Motifs
- Traffic Classification
- Evaluation
Motivation
Motivation

Towards Reliable Traffic Classification Using Visual Motifs
Towards Reliable Traffic Classification Using Visual Motifs

Wilson Lian, John McHugh, Fabian Monrose
Assumptions

- Reliable transport via TCP
- Stream Cipher
  - No access to payload
  - Length preservation
- Negligible packet loss & retransmission
Related Work

- Using Visual Motifs to Classify Encrypted Traffic, Wright et al. 2006
- Intelligent Classification and Visualization of Network Scans Muelder et al. 2008.
Timeline Heatmaps

Image credit: Wright et al. 2006
Timeline Heatmaps

Lineplot for http Connection

Image credit: Wright et al. 2006
Timeline Heatmaps

Image credit: Wright et al. 2006
Timeline Heatmaps

Image credit: Wright et al. 2006
Timeline Heatmaps

Image credit: Wright et al. 2006
Timeline Heatmaps

Heatmap for http

Image credit: Wright et al. 2006
Timeline Heatmaps

(a) HTTP

(b) SMTP

(c) AIM

(d) SSH

Image credit: Wright et al. 2006
Unigram Heatmaps

(a) HTTP
(b) SMTP
(c) AIM
(d) SSH

Image credit: Wright et al. 2006
Bigram Heatmaps
Heatmap Construction

Client -> Server

- SYN 48 bytes
- SYN-ACK 48 bytes
- ACK 40 bytes
- HTTP Request 891 bytes
- 1500 bytes
- 40 bytes
- 40 bytes
- 1500 bytes
- 40 bytes
- 40 bytes
- 1500 bytes
- 40 bytes

(48, -48)
(-48, 40)
(40, 891)
(891, -40)
(-40, -270)
(-270, -1500)
(-1500, 40)
(40, -1500)
(-1500, 40)
## Heatmap Construction

<table>
<thead>
<tr>
<th>(-48, 40)</th>
<th>(40, 891)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(-1500, 40)</td>
<td>(40, -1500)</td>
</tr>
<tr>
<td>(-1500, 40)</td>
<td>(40, -1500)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(-40, -270)</th>
<th>(48, -48)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(-270, -1500)</td>
<td>(891, -40)</td>
</tr>
<tr>
<td>(-270, -1500)</td>
<td>(40, -1500)</td>
</tr>
</tbody>
</table>
**Heatmap Construction**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>3/9 = 33.3%</td>
<td>1/9 = 11.1%</td>
</tr>
<tr>
<td>2/9 = 22.2%</td>
<td>3/9 = 33.3%</td>
</tr>
</tbody>
</table>
Heatmap Construction

Background
Visual Motifs
Traffic Classification
Evaluation

Wilson Lian, John McHugh, Fabian Monrose
Towards Reliable Traffic Classification Using Visual Motifs
Bigram Heatmaps
### Modeling Protocol Behavior

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>3/9 = 33.3%</td>
<td>1/9 = 11.1%</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2/9 = 22.2%</td>
<td>3/9 = 33%</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>
Modeling Protocol Behavior

![Bar chart showing bin probabilities]

- Bin 0: 0.333
- Bin 1: 0.111
- Bin 2: 0.222
- Bin 3: 0.333
- Bin 4: 0.333

Wilson Lian, John McHugh, Fabian Monrose
Towards Reliable Traffic Classification Using Visual Motifs
Comparing Protocol Models
Comparing Protocol Models

\[ A_{total} = \sum_{k=1}^{n} A_k \]

\[ B_{total} = \sum_{k=1}^{n} B_k \]

\[ \text{Score}_{A \leftrightarrow B} = \sum_{i=1}^{n} \left| \frac{A_i}{A_{total}} - \frac{B_i}{B_{total}} \right| \]

\[ = \frac{1}{A_{total} \cdot B_{total}} \sum_{i=1}^{n} |A_i \cdot B_{total} - B_i \cdot A_{total}| \]
Comparing Protocol Models

Score = .233 + .589 + .072 + .283 = 1.177

<table>
<thead>
<tr>
<th>Difference</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.333</td>
</tr>
<tr>
<td>2</td>
<td>.111</td>
</tr>
<tr>
<td>3</td>
<td>.222</td>
</tr>
<tr>
<td>4</td>
<td>.333</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>.100</td>
<td>.050</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Difference</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>.333−.100</td>
<td>.233</td>
</tr>
<tr>
<td>.111−.700</td>
<td>.589</td>
</tr>
<tr>
<td>.222−.150</td>
<td>.072</td>
</tr>
<tr>
<td>.333−.050</td>
<td>.283</td>
</tr>
</tbody>
</table>
Comparing Protocol Models
Comparing Protocol Models

Score = 0.067 + 0.039 + 0.072 + 0.033 = 0.211

\( |0.333 - 0.400| = 0.067 \)
\( |0.111 - 0.150| = 0.039 \)
\( |0.333 - 0.300| = 0.033 \)
\( |0.222 - 0.150| = 0.072 \)
Classifying Samples: Easy as 1-2-3

1. Create training models for desired protocols
2. Build distribution for sample network trace
3. Find training model with lowest difference score

\[
\text{Score}_{A\leftrightarrow B} = \sum_{i=1}^{n} \left| \frac{A_i}{A_{\text{total}}} - \frac{B_i}{B_{\text{total}}} \right|
\]

\[
= \frac{1}{A_{\text{total}} \cdot B_{\text{total}}} \sum_{i=1}^{n} \left| A_i \cdot B_{\text{total}} - B_i \cdot A_{\text{total}} \right|
\]
Evaluation

- How much traffic must be collected for:
  - Training
  - Testing
- Precision?
  \[
  \text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}
  \]
- Recall?
  \[
  \text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}
  \]
Data

- CRAWDAD Dataset
- Weekdays: January 19, 2004 – February 6, 2004
- Ports with sufficient traffic
  - $\geq 1$M packets
  - 0.3% of ports $\rightarrow 95.21\%$ of packets
- Keep top 10 ports by number of sessions observed
- No ground truth

<table>
<thead>
<tr>
<th>Total Packets</th>
<th>1.3 Billion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traffic Volume</td>
<td>707 GB</td>
</tr>
<tr>
<td>Observed Ports</td>
<td>64,214</td>
</tr>
<tr>
<td>Sessions</td>
<td>5.2 Million</td>
</tr>
<tr>
<td>Port 80 Sessions</td>
<td>1.7 Million</td>
</tr>
</tbody>
</table>
Trial :=
1. Randomly sample some percentage of available data for each port and train classifier
2. Randomly sample some number of the remaining data points for each port and create testing samples
3. Classify testing samples

50 Trials
Training Size Selection

Average Recall for Varying Training Size

6.8% improvement

Wilson Lian, John McHugh, Fabian Monrose
Towards Reliable Traffic Classification Using Visual Motifs
Towards Reliable Traffic Classification Using Visual Motifs

Training Size Selection

Average Precision for Varying Training Size

- 6.6% improvement
- Average Precision: 0.953
- Training Sample Size (%): 10%
Testing Size Selection

Average Recall for Varying Testing Size

- Testing Sample Size (Data Points)
- Average Recall

0.88
0.89
0.9
0.91
0.92
0.93
0.94
0.95
0.96
0.97
0.98
0.99
1

5.6% improvement

Wilson Lian, John McHugh, Fabian Monrose
Towards Reliable Traffic Classification Using Visual Motifs
Testing Size Selection

Average Precision for Varying Testing Size

Average Precision

Testing Sample Size (Data Points)

4.7% improvement

Wilson Lian, John McHugh, Fabian Monrose
Towards Reliable Traffic Classification Using Visual Motifs
**Results**

50 Trials

15% Training Set Size

50,000 Data Points Testing Set Size

96.5% Precision

96.0% Recall
Goal: Eliminate close calls

Require 1st place candidate to lead 2nd place by certain amount to make decision

Standard deviation of scores
Methodology v2.0

- Randomly sample some percentage of available data for each port and **train** classifier
- Randomly sample some number of the remaining data points for each port and create **testing** samples
- Attempt to classify testing samples
  - If all testing samples reach threshold, done.
  - If any testing sample fails, rebuild testing samples and try again.
Classification Confidence

50 Trials

5% Training Set Size

35,000 Data Points Testing Set Size

1.0 Lead Threshold

96.9% Precision

96.6% Recall
Classification Confidence

Testing Session Sampled For 50 Trials
35,000 Data Point Testing Samples

Wilson Lian, John McHugh, Fabian Monrose
Towards Reliable Traffic Classification Using Visual Motifs
Ground Truth Testing

MIT Lincoln Labs DARPA Data
50 trials, 5% training sample size, 35,000 data point testing
sample size, 1.25 lead threshold

- Precision: 98.3%
- Recall: 98.0%
Results

50 Trials
5% Training Set Size
35,000 Data Points Testing Size
1.25 Lead Threshold

Ground Truth

Confusion Matrix for 50 Trials-0.05pctTraining-35kTesting-Disjoint-1.25 Lead Threshold
Evasion

One might attempt to thwart our technique by padding all packets to MTU. Reduces problem to 4-quadrant problem. Can still make decisions based on relative prevalence of each quadrant.
Current/Future Work

- Packet loss/re-transmission may cause unpredictable results
- On-line classification
- Training and testing from separate datasets
- UDP
- Subcategorization
Conclusion

- Modeling protocol behavior using only packet size, direction, and order
- Resistant to encryption and padding
- Average precision and recall $> 97\%$
- Quick and reliable traffic inspection
- Useful for pre-screening traffic for deeper analysis
Questions?

Thanks for listening.

Q & A

wwlian@gmail.com