Traffic Classification based on Visualization

Zhibin Yu
Realtime Image Processing & Telecommunication Lab.
Kyungpook National University
South Korea

December 8th, 2011
Overview

Figure A Is that possible to use face recognition technology to classify network traffic?
Introduction

Input traffic flows

Image enhancement

Pattern recognition (PCA)

Normalization of packet size and interval

Display 2-D images

Results

Figure B Flow chart
Related Work

- Port-based Approaches
  - Fast but unreliable
- Signature-based approaches
  - Accurate but inflexible
- Statistical-based approaches
  - Weak in small flows
- Machine Learning Approaches
  - Accurate but time costly
- Traffic Classification Metric

\[ F = \frac{2TP}{2TP + FN + FP} \]

- Where TP is true positive, FN is false negative and FP is false positive.
Proposed Method (1/5)

- Feature selection
  - Packet size
  - Packet inter-arrival time

Figure C: Cumulative Distribution of Packet Size and Packet Inter-arrival time using our experiment dataset
Proposed Method (2/5)

- **Image Normalization**
  - Definition

\[
X = 512 \times \sqrt{\frac{\text{Packet}_{\text{size}}}{\text{MTU}}} \\
Y = 512 \times \left(\frac{\text{Packet}_{\text{Interval}}}{\text{Max}_{\text{Interval}}}\right)^a
\]

- MTU=1500
- Max_{Interval}=3600s
- \(a=0.1\)
Proposed Method (3/5)

- **Image Normalization**
  - Initialized Images

Figure D: Four local images generated by FTP-data and OICQ flows.
(a) FTP-data1. (b) FTP-data2. (c) OICQ1. (d) OICQ2.
Proposed Method (4/5)

- **Image enhancement**
  - Mountain clustering and visualization

\[
M_a = \left( \sum_{i=1}^{N} \exp\left(-\frac{\|a-a_i\|^2}{2\sigma^2}\right) \right)^b
\]

\[
B_a = 255 \times \left( \frac{M_a}{M_{\text{max}}} \right)
\]

- N is the total number of packets in this flow
- \(M_a\) is the mountain height of point a calculated by equation 1
- b is the parameter which controls the difference between peak and plain
- \(B_a\) is the brightness value of point a
Proposed Method (5/5)

- Image enhancement
  - Mountain clustering

Figure E: Image enhancement after mountain clustering.
(a) Original image. (b) Mountain clustering value. (c) Image after enhancement when $b=0.3 \sigma=2$.
Evaluation (1/5)

- Evaluation
  - Data Description

<table>
<thead>
<tr>
<th>Traffic name</th>
<th>Average flow size (Kbytes)</th>
<th>Average interval (Seconds)</th>
<th>Average packet size (Bytes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HTTP</td>
<td>37.9</td>
<td>0.759</td>
<td>646</td>
</tr>
<tr>
<td>FTP</td>
<td>7.64</td>
<td>11.621</td>
<td>84</td>
</tr>
<tr>
<td>FTP-data</td>
<td>21386.7</td>
<td>0.0666</td>
<td>576</td>
</tr>
<tr>
<td>OICQ (Chatting)</td>
<td>18.9</td>
<td>3.531</td>
<td>200</td>
</tr>
<tr>
<td>POP3</td>
<td>35.3</td>
<td>0.246</td>
<td>317</td>
</tr>
<tr>
<td>SMTP</td>
<td>19.9</td>
<td>0.0835</td>
<td>464</td>
</tr>
<tr>
<td>Web-download</td>
<td>21734.4</td>
<td>0.0425</td>
<td>1142</td>
</tr>
<tr>
<td>PPStream (P2PTV)</td>
<td>1911.6</td>
<td>0.284</td>
<td>371</td>
</tr>
</tbody>
</table>

Table 1: Data description
Evaluation (2/5)

- Performance
  - Comparison with different parameters

Figure F: (a) Initialized image generated by a PPS flow (11.4MB).
(b) $b=0.1 \ \sigma=4$. (c) $b=0.2 \ \sigma=4$. (d) $b=0.5 \ \sigma=4$. (e) $b=0.2 \ \sigma=1$ (f) $b=0.2 \ \sigma=2$ (g) $b=0.2 \ \sigma=4$
Evaluation (3/5)

- Performance
  - Comparison with different flow size

Figure G: Performance between different flow size
(a) Initialized images generated by an elephant FTP-data flow (53,401KB). (b) Initialized images generated by a mice FTP-data flow (4.43KB). (c) Image (a) after enhancement with parameter $b=0.3 \sigma=2$. (d) Image (b) after enhancement with parameter $b=0.3 \sigma=2$. 

16/12/11

Realtime Image Processing & Telecommunication Lab.
• Performance
  – Performance on encryption detection
  – 386 ftp-data flows based on SFTP protocol using SSH2 are tested

Figure H: Performance of encryption traffic classification
(a) An image generated from an SSH flow (13.4MB) with parameter b=0.3 \( \sigma = 2 \).
(b) An image generated from an FTP-data flow (13.0MB) with parameter b=0.3 \( \sigma = 2 \).
(c) Recognition result by PCA.
Evaluation (5/5)

- **Evaluation**
  - **Recognition result**

<table>
<thead>
<tr>
<th>Traffic type</th>
<th>F-measure (%)</th>
<th><strong>b=0.2</strong></th>
<th><strong>b=0.3</strong></th>
<th><strong>b=0.4</strong></th>
<th><strong>b=0.5</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>FTP</td>
<td>Initialized</td>
<td>σ=1</td>
<td>σ=2</td>
<td>σ=4</td>
<td>σ=1</td>
</tr>
<tr>
<td></td>
<td>image</td>
<td>77.78</td>
<td>91.67</td>
<td>97.22</td>
<td>97.12</td>
</tr>
<tr>
<td>FTP-data</td>
<td>b=0.2</td>
<td>59.76</td>
<td>90.28</td>
<td>81.94</td>
<td>78.32</td>
</tr>
<tr>
<td>HTTP</td>
<td>b=0.3</td>
<td>86.39</td>
<td>73.06</td>
<td>75.86</td>
<td>60.56</td>
</tr>
<tr>
<td>OICQ (Chatting)</td>
<td>b=0.4</td>
<td>83.39</td>
<td>75.02</td>
<td>94.44</td>
<td>86.11</td>
</tr>
<tr>
<td>POP3</td>
<td>b=0.5</td>
<td>76.12</td>
<td>97.21</td>
<td>100</td>
<td>98.33</td>
</tr>
<tr>
<td>SMTP</td>
<td>b=0.6</td>
<td>94.44</td>
<td>100</td>
<td>99.27</td>
<td>100</td>
</tr>
<tr>
<td>PPStream,(P2PTV))</td>
<td>b=0.7</td>
<td>80.56</td>
<td>91.67</td>
<td>97.22</td>
<td>94.44</td>
</tr>
<tr>
<td>Web-download</td>
<td>b=0.8</td>
<td>87.23</td>
<td>92.21</td>
<td>90.57</td>
<td>89.44</td>
</tr>
<tr>
<td>Average</td>
<td>b=0.9</td>
<td>80.72</td>
<td>88.76</td>
<td>92.07</td>
<td>87.95</td>
</tr>
</tbody>
</table>

Figure 8: Anomalies identified both methods, sorted by traffic.
Conclusion

- We proposed a simple method to classify network traffic based on visualization and pattern recognition.
- We get more than 93% F-measure through classification.
- Our proposed method is able to detect encrypted flows.
- This algorithm can reduce the gap between elephant flow and mice flow.
Thank you!