State of the Art in Traffic Classification:
A Research Review

min zhang
wolfgang john
kc claffy
nevill browlee
Outline

- Motivation
- Research review and taxonomy
- Survey analysis: P2P
- Discussion and conclusion
Motivation

- Today’s Internet
  - evolving in scope and complexity
  - applications adapt rapidly to detection attempts
  - emerging obfuscation techniques

- Many classification approaches in literature
  - using whatever traffic samples available
  - no systematic integration of results
Motivation contd.

- Filling this gap, our research review
  - creates a structured taxonomy of traffic classification papers and their datasets
  - helps to answer popular questions
  - reveals open issues and challenges
Research review and taxonomy

- 64 papers published between 1994 and 2008

- **Definition:** *traffic classification*
  
  **Methods** of classifying traffic **data sets** based on **features** passively observed in the traffic, according to specific **classification goals**.

http://www.caida.org/research/traffic-analysis/classification-overview
Research review and taxonomy contd.

- **Data sets**: more than 80 data sets used for 64 papers!
  
  Categorized by: Time of collection, link type, capture environments, geographic location, payload length, etc

- **Classification goals**: coarse or finer-grained
Research review and taxonomy contd.

Features

Figure 1: Trends of applications and features
Research review and taxonomy contd.

- Methods
  - exact matching: port number, payload, etc
  - heuristic methods, e.g. on connection patterns
  - machine learning methods: supervised and unsupervised

http://www.caida.org/research/traffic-analysis/classification-overview
Survey analysis: P2P

- How much P2P?
  1.2% to 93% across the 18 (out of 64) papers
Survey analysis: P2P contd.

- How much P2P? (cont’)

<table>
<thead>
<tr>
<th>Geo Location</th>
<th>Year</th>
<th>Range of P2P Volume</th>
<th>Paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Europe</td>
<td>2005</td>
<td>60-80%</td>
<td>[15]</td>
</tr>
<tr>
<td></td>
<td>2006</td>
<td>79-93%</td>
<td>[7, 8]</td>
</tr>
<tr>
<td>North America</td>
<td>2003</td>
<td>8%, 10.7%</td>
<td>[9]</td>
</tr>
<tr>
<td></td>
<td>2004</td>
<td>14%, 9.9%</td>
<td>[9]</td>
</tr>
<tr>
<td></td>
<td>2003-04</td>
<td>9.2-70%</td>
<td>[10, 6, 12]</td>
</tr>
<tr>
<td></td>
<td>2006</td>
<td>21-35%</td>
<td>[3, 5, 4]</td>
</tr>
<tr>
<td>Asia</td>
<td>2002</td>
<td>21.5%</td>
<td>[14]</td>
</tr>
<tr>
<td></td>
<td>2005</td>
<td>1.34% (port-based)</td>
<td>[2]</td>
</tr>
<tr>
<td></td>
<td>2008</td>
<td>1.29% (port-based)</td>
<td>[2]</td>
</tr>
</tbody>
</table>
Discussion and Conclusions

- Shortcomings of current traffic classification efforts:
  - 80 data sets by 64 papers → lack of shared, current data sets as reference data
  - no clear definition of P2P or file-sharing → lack of standardized measures and classification goals

- Poor comparability of results!!!
Discussion and Conclusions contd.

- So how much of modern Internet traffic is P2P?
  
  "there is a wide range of P2P traffic on Internet links; see your specific link of interest and classification technique you trust for more details."

- This review can answer further questions:
  - TCP/UDP ratio?
  - Amount of encrypted traffic?
  - Tunneled traffic?
  - ...

- Thanks
- 謝謝
- Bedankt
- Merci
- Danke
- Ευχαριστώ
- Grazie
- ありがとう
- 감사합니다
- Dzieki
- Gracias
- شكرا

http://www.caida.org/research/traffic-analysis/classification-overview/