Distilling the Internet’s Application Mix from Packet-Sampled Traffic

Passive and Active Measurement Conference 2015
New York City

Philipp Richter
TU Berlin

Nikolaos Chatzis
TU Berlin

Georgios Smaragdakis
TU Berlin / MIT

Anja Feldmann
TU Berlin

Walter Williger
NIKSUN, Inc.

prichter@inet.tu-berlin.de
The Internet’s Application Mix

• ”What is the application mix in today’s Internet”?
  • Optimizing network performance
  • Provisioning network resources
  • Identifying new trends in Internet usage

• Numerous academic and commercial studies
• Typically focus on a single or a few locations
• We study the application mix seen on tens of thousands of peering links at a Large European IXP
Agenda

- Dataset & Challenges
- Related Work & Applicability
- Classification Approach
- Results
A Large European IXP

- Daily traffic (2013): ~14 PB
- sFlow export
- Random Sampling 1/16K Packets
- Snaplen 128 Bytes
- Weekly Snapshosts dating back to 2011

most recent snapshot (2013-09, 496 networks, 1 week)

<table>
<thead>
<tr>
<th>packets sampled</th>
<th>bytes sampled</th>
<th>IPv4 / IPv6</th>
<th>TCP / UDP</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.3B</td>
<td>5.9TB</td>
<td>99.37% / 0.63%</td>
<td>83.7% / 16.3%</td>
</tr>
</tbody>
</table>
Dataset Characteristics: Sampling

• Typically one packet per sampled flow
• Can be any packet (e.g., just a TCP ACK)
• One packet = unidirectional visibility
• a “random set” of packets

5-tuple aggregation:
packets seen per sampled flow
**Dataset Characteristics: Sampling**

86% of sampled TCP flows: one packet

- Typically one packet per sampled flow
- Can be any packet (e.g., just a TCP ACK)
- One packet = unidirectional visibility
- a “random set” of packets

5-tuple aggregation: packets seen per sampled flow
Classification Approaches

(I) Payload-based Approach

• Match application signatures (i.e., handshakes)
• Produces accurate results
• Challenge: Most sampled packets are “in the middle” of a flow and contain only binary data.

(II) Port-based Approach

• Match port-numbers to well-known applications
• Problems: Applications hiding behind well-known ports, applications using random ports (P2P)
• Applicable as-is
Classification Approaches (cont.)

(III) Flow feature-based Approach

- Match per-flow properties (i.e., #packets, #avg. packet size etc.)
- *Not applicable, no per-flow statistics available*

(IV) Host behavior-based Approach

- Social interaction between hosts (e.g., BLINC)
- Network-wide interaction of hosts (e.g., TDGs)
- *Partially applicable*

taxonomy based on Kim et al.
Classification Approaches

(III) Flow feature-based Approach
- Match per-flow properties (i.e., #packets, #avg. packet size etc.)
- Not applicable, no per-flow statistics available

(IV) Host behavior-based Approach
- Social interaction between hosts (e.g., BLINC)
- Network-wide interaction of hosts (e.g., TDGs)
- Partially applicable

We combine several approaches

taxonomy based on Kim et al.
Our Classification Approach

1. Pre-Classification Phase
   derive state which will be leveraged later

2. Classification Phase
   actual classification of packets
Pre-Classification Phase

**state - server-side:**
- Server-Endpoint (IP,port)
  e.g., ~2.7M HTTP endpoints
- For SSL-based applications:
  SSL signature on well-known port
  e.g., ~210K HTTPS endpoints (IP,port)

**state - client-side:**
- BitTorrent peer IPs (~38.9M)
- Web Client IPs (~37.7M)

We extract Connection Endpoints ("state")
(1) Payload Signatures

1. For each packet:
   - payload signature
   - classify to respective protocol

Completeness (bytes):
- 11.7%

---

| SRC   | DST      | Message...
|-------|----------|-------------
| 1.2.3.4:80 | 5.6.7.8:34325 | HTTP/1.1 200 OK ...

Signature match?
- -> HTTP
(2) Server Endpoint Matching

For each packet:

1. Classify the payload signature
2. Match to a known server endpoint

Classify to respective protocol

Completeness (bytes):
11.7% 78.2%

<table>
<thead>
<tr>
<th>SRC</th>
<th>DST</th>
<th>...</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.2.3.4:80</td>
<td>5.6.7.8:34325</td>
<td>... A0FD03480CF ...</td>
<td>...</td>
</tr>
</tbody>
</table>

1.2.3.4:80 HTTP pre-classified

Endpoint match?
-> HTTP
(3) Fallback: Port-Based Classification

For each packet:
1. Payload signature
2. Known server endpoint
3. Known port number

Classify to respective protocol

Completeness (bytes):
- SRC 1.2.3.4:1935
- DST 5.6.7.8:34325
- ... A0FD03480CF ...

Port match?
-> RTMP
(4) Catching the TCP BitTorrent Traffic

for each packet

1. payload signature
2. known server endpoint
3. known port number
4. between BT speakers

classify to respective protocol

"BT/P2P likely"

completeness (bytes)

11.7% 78.2% 82.7% 92.9%

SRC DST
1.2.3.4:42364 5.6.7.8:34325
...

A0FD03480CF ...

SRC and DST BitTorrent Peers?
-> BT/P2P likely
(5) Tie Breaker: Other P2P likely

for each packet

1. payload signature
2. known server endpoint
3. known port number
4. between BT speakers
5. between BT or Web clients

classify to respective protocol

“BT/P2P likely”
“P2P likely”

completeness (bytes)

11.7% 78.2% 82.7% 92.9% 94.2%

SRC
1.2.3.4:42364

DST
5.6.7.8:34325

… A0FD03480CF …

SRC and DST BitTorrent Peers
OR Web Clients?
-> P2P likely
The Application Mix: Aggregate

- HTTP(S) dominates ~67%
- other applications (e.g., RTMP, mail, news) ~6%
- BitTorrent/BT/P2P likely ~22%
- unclassified ~5%
The Application Mix: Over Time

- Diurnal patterns, e.g., P2P dominates in off-hours
- Historical view shows increasing dominance of HTTP(S) and significant HTTPS increase in 2013.
## The Application Mix: Same Everywhere?

<table>
<thead>
<tr>
<th>Study</th>
<th>Vantage Point(s)</th>
<th>Method</th>
<th>HTTP(S)</th>
<th>other</th>
<th>BT/P2P</th>
<th>unclassified</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labovitz, 2010</td>
<td>5 large ISPs</td>
<td>payload</td>
<td>52.1</td>
<td>24</td>
<td>18.3</td>
<td>5.5</td>
</tr>
<tr>
<td>Labovitz, 2010</td>
<td>110 Networks</td>
<td>port</td>
<td>52</td>
<td>10</td>
<td>1</td>
<td>37</td>
</tr>
<tr>
<td>Maier, 2009</td>
<td>Access ISP</td>
<td>payload</td>
<td>57.6</td>
<td>23.5</td>
<td>13.5</td>
<td>10</td>
</tr>
<tr>
<td>Gerber, 2011</td>
<td>Backbone ISP</td>
<td>payload</td>
<td>60</td>
<td>28</td>
<td>12</td>
<td>N/A</td>
</tr>
<tr>
<td>Czyz, 2014</td>
<td>260 Networks</td>
<td>port</td>
<td>69.2</td>
<td>4</td>
<td>&lt;7</td>
<td>20</td>
</tr>
<tr>
<td>Sandvine, 2014</td>
<td>VA, North America</td>
<td>payload</td>
<td>~70</td>
<td>N/A</td>
<td>6</td>
<td>N/A</td>
</tr>
<tr>
<td>Sandvine, 2014</td>
<td>VA, Europe</td>
<td>payload</td>
<td>~65</td>
<td>N/A</td>
<td>15</td>
<td>N/A</td>
</tr>
<tr>
<td>Sandvine, 2014</td>
<td>VA, Asia-Pacific</td>
<td>payload</td>
<td>~60</td>
<td>N/A</td>
<td>30</td>
<td>N/A</td>
</tr>
<tr>
<td>Sandvine, 2014</td>
<td>VA, Latin America</td>
<td>payload</td>
<td>~65</td>
<td>N/A</td>
<td>9.4</td>
<td>N/A</td>
</tr>
<tr>
<td>this study</td>
<td>European IXP</td>
<td>various</td>
<td>66.9</td>
<td>5.9</td>
<td>21.4</td>
<td>5.8</td>
</tr>
</tbody>
</table>

around 55-70% HTTP(S), another 10%-20% Peer-to-Peer across different vantage points.

Is that what I can expect on any link I measure? Should I design my network for these applications?
The Application Mix: Per Network Type

- Content/CDN almost 100% HTTP
- HTTPS increase driven by only a few networks
- P2P not only between Eyeballs! Hoster/IaaS too!

Dissecting per network shows a different appmix!
The Application Mix: Per Link

- Aggregate mix by no means representative of single link
- Many links just have one dominant protocol
- The business type of the ASes gives hints on app mix
The Application Mix: Per Link

- Aggregate mix by no means representative of single link
- Many links just have one dominant protocol
- The business type of the ASes gives hints on app mix

content <> eyeball: HTTP
The Application Mix: Per Link

- Aggregate mix by no means representative of single link
- Many links just have one dominant protocol
- The business type of the ASes gives hints on app mix

**eyeball <> eyeball: P2P**
The Application Mix: Per Link

- Aggregate mix by no means representative of single link
- Many links just have one dominant protocol
- The business type of the ASes gives hints on app mix

**hoster/IaaS: diverse application mix**
Summary

• By using a *stateful* approach, we can largely overcome the limitations of random packet sampling
• We can classify up to 95% of the bytes exchanged

• Our results:
  • Application mix similar to commonly reported
  • Dissecting per Network Type reveals different appmix
  • Business types of involved networks give hints