Rendezvous-based Traffic Classification, Measurement, and Analysis

ISC/CAIDA Data Collaboration Workshop
October 22, 2012

David Plonka &
Paul Barford
{plonka,pb}@cs.wisc.edu
Outline

• Rendezvous-based Traffic Analysis
  – What is it? Why use it?

• Implementation: TreeTop
  – a DNS rendezvous-based analysis tool
    [Plonka & Barford, IMC 2009, SATIN 2011, work in progress]
  – flow export with rendezvous annotations

• Sample Applications:
  – Aggregate traffic measurement by service
  – Passive performance measurement of services on IPv6 versus IPv4
Rendezvous-based Traffic Analysis?

- Traffic classification and analysis has focussed on target traffic features (IP headers, DPI, etc.)

- However, Internet hosts learn IP addresses by some *rendezvous* mechanism, e.g.:
  - By static configuration (IP addrs in config files)
  - The Domain Name System (DNS)
  - Application-specific mechanisms (URLs, p2p)

- Inform traffic analysis by considering, “How does this host know this IP address?” rather than simply, “With what IP address did this host interact?”
Why Focus on Rendezvous?

*Rendezvous*: how hosts “present themselves”

- For standard protocols, rendezvous information is not private and is of low-volume
  - Separate and separable from private payloads
  - Can be monitored in situations where target traffic is *high-volume, sampled, or encrypted*

- Rendezvous info can indicate when other analysis or classification techniques are effective and when they're not
  - e.g., bolstered *port-based classification*
    
    [Kim, et al., 2008] [Plonka & Barford, 2011]
Traffic Observation Points

Internet host

DNS server

client
Traffic Observation Points

SPI, high-volume

DPI, low-volume
Rendezvous-annotated Flow Export

*TreeTop* uses two annotation approaches for flow source and destination addresses:

- **Direct**: *TreeTop* discovers that the given client end-host knows a remote IP address by a domain name from a prior DNS A or AAAA query.

- **Consensus**: we *infer*, by *shared* consensus of other client end-hosts, that the hosts could have used the DNS to similarly resolve the peer's name. *Name sampling* is performed to clarify otherwise ambiguous names.
TreeTop: radix tries and domain trees
TreeTop enhanced with nmsg support

We select **nmsg** because it provides:

- an extensible mechanism for encapsulating rendezvous and IP traffic trace (flow) data
- a means of transmitting streams to distributed encapsulation and online analysis elements
- a serialized file format for offline analyses
- a scripting interface to build prototype components and perform ad hoc analyses
Rendezvous-annotated Flow Export

Legend:
- process/executable
- data object/file
- process writes to file
- process reads from file

Recursive DNS server

flow export from router(s)

Rendezvous-annotated Flow Export diagram with process and data flow.
[2011-06-08 21:52:26.000000000] [7:1 WISC nfdump]
ts: 1307569945
te: 1307569945
td: 0.064000
sa: 203.0.113.71
da: 192.0.2.32
sp: 80
dp: 55983
pr: 6
ibyt: 396630
snamed: CLIENT_DNS_NAMED
sn: static.ak.facebook.com
ip_version: IPV4
Rendezvous-annotated Flow Export (2)

[2011-06-08 20:14:11.000000000] [7:1 WISC nfdump]
ts: 1307564050
ten: 1307564050
td: 0.064000
sa: 2001:0db8::face:b00c:0:3
da: 2001:0db8::2:1
sp: 443
dp: 53646
pr: 6
ibyt: 34297
snamed: INFERRED_DNS_NAMED
sm_sample: de-de.facebook.com.
ip_version: IPV6
Rendezvous-annotated Flow Export (3)

[219] [2011-06-08 00:11:10.000000000] [7:1 WISC nfdump]

ts: 1307491869
te: 1307491869
td: 0.128000
sa: 2001:0db8::2:1
da: 2001:0db8:fff4::79
sp: 56451
dp: 80
pr: 6
ibyt: 849
dnamed: INFERRED_DNS_NAMED
dn: *
dn_sample: rss.slashdot.org.
ip_version: IPV6
Aggregate Traffic: named & unnamed
Aggregate Traffic by Domain Name
# World IPv6 Day Performance Study: Trace Data Characteristics

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trace duration</td>
<td>24 hours</td>
</tr>
<tr>
<td>DNS query responses</td>
<td>~14.2M</td>
</tr>
<tr>
<td>DNS IPv4 client addresses</td>
<td>2028</td>
</tr>
<tr>
<td>DNS IPv6 client addresses</td>
<td>23</td>
</tr>
<tr>
<td>DNS AAAA queries</td>
<td>~114.3K</td>
</tr>
<tr>
<td>DNS AAAA NOERROR responses</td>
<td>~6.2K</td>
</tr>
<tr>
<td>Flows - IPv4</td>
<td>~58.8M</td>
</tr>
<tr>
<td>Flows - IPv6</td>
<td>~2.4M</td>
</tr>
</tbody>
</table>
## World IPv6 Day:
### Popular IPv6 FQDNs

<table>
<thead>
<tr>
<th>Rank</th>
<th>FQDN</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><a href="http://www.google.com">www.google.com</a>.</td>
</tr>
<tr>
<td>2</td>
<td><a href="http://www.google-analytics.com">www.google-analytics.com</a>.</td>
</tr>
<tr>
<td>4</td>
<td>ssl.gstatic.com</td>
</tr>
<tr>
<td>5</td>
<td>safebrowsing.clients.google.com.</td>
</tr>
<tr>
<td>6</td>
<td>mail.google.com.</td>
</tr>
<tr>
<td>7</td>
<td>safebrowsing-cache.google.com.</td>
</tr>
<tr>
<td>8</td>
<td>clients1.google.com.</td>
</tr>
<tr>
<td>9</td>
<td><a href="http://www.youtube.com">www.youtube.com</a>.</td>
</tr>
<tr>
<td>10</td>
<td>view.atdmt.com.</td>
</tr>
<tr>
<td>12</td>
<td>news.google.com.</td>
</tr>
<tr>
<td>13</td>
<td>maps.google.com.</td>
</tr>
<tr>
<td>14</td>
<td>ssl.google-analytics.com.</td>
</tr>
<tr>
<td>15</td>
<td>addons.mozilla.org.</td>
</tr>
<tr>
<td>16</td>
<td>docs.google.com.</td>
</tr>
<tr>
<td>17</td>
<td>chat-encrypted.mail.google.com.</td>
</tr>
<tr>
<td>18</td>
<td>translate.google.com.</td>
</tr>
<tr>
<td>19</td>
<td>mail-attachment.googleusercontent.com.</td>
</tr>
<tr>
<td>20</td>
<td>sites.google.com.</td>
</tr>
</tbody>
</table>
Facebook Active Client IP Addresses

![Graph showing the distribution of active client IP addresses over different hours of the day for IPv4 and IPv6.]
Gmail Active Client IP Addresses
Facebook WWW Flow Bit Rates
Gmail WWW Flow Bit Rates

![Graph showing bit rates over time for IPv4 and IPv6.]
Facebook WWW Flow Bit Rates (detail)
Gmail WWW Flow Bit Rates (detail)
Sharing Opportunities

• Use of dnsdb as basis for consensus labeling?

• Streams of anonymized recursive DNS query/responses?

• Tap other rendezvous mechanisms?

• Aggregate measurements, e.g. flow volumes, by DNS rendezvous?
Rendezvous-based Traffic Classification, Measurement, and Analysis

FIN

David Plonka
&
Paul Barford
{plonka,pb}@cs.wisc.edu