METHODOLOGICAL AND TECHNICAL CHALLENGES OF LARGE SCALE DNS ANALYSIS

Jingxiu Su
Researches

• Education:
  ➢ Feb.2016—Jan.2017 Polytech, University of Savoie
    Joint Ph.D in Systems for advertisement on mobiles
  ➢ Sept.2013—Present Institute of Computing Technology, Chinese Academy of Science
    Master+Ph.D in Computer System Architecture
    Bachelor in Computer science and technology

• Project List:
  ➢ Mar.2013—Jun.2013 Developing a prototype of Service-Oriented Future Internet architecture
  ➢ Nov.2015—Oct.2016 Analysis of Online Advertisement Behaviors on DNS trace

• Paper List:
  ➢ Index-Trie: Efficient Archival and Retrieval of Network Traffic
    ➢ Computer Networks journal paper
  ➢ Methodological and technical challenges of large scale DNS analysis
    ➢ Will Submit to IEEE Infocom 2018 conference
Outline

Dataset

Identify of Web Trackers

User Sessions Extraction

Two main Issues and Solutions

Results of tracking behavior
Dataset

- DNS cache log data
- 2 days in July 2015
- 27 regions in China
- Record Format:
  
  User IP | Requested Domain | Timestamp | Resolved IP List | Type

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>223.73.80.194</td>
<td>wifi.dianxin.net</td>
<td>20150716040012</td>
<td>117.185.16.47</td>
<td>0</td>
</tr>
<tr>
<td>223.73.80.194</td>
<td>wifi.dianxin.net</td>
<td>20150716040012</td>
<td>117.185.16.47</td>
<td>0</td>
</tr>
<tr>
<td>223.73.80.194</td>
<td>mtalk.google.com</td>
<td>20150716040013</td>
<td>120.196.0.5</td>
<td>0</td>
</tr>
<tr>
<td>223.73.80.194</td>
<td>mtalk.google.com</td>
<td>20150716040013</td>
<td>120.196.0.5</td>
<td>0</td>
</tr>
<tr>
<td>223.73.80.194</td>
<td>a.root-servers.net</td>
<td>20150716040014</td>
<td>198.41.0.4</td>
<td>0</td>
</tr>
</tbody>
</table>

- Distinct values:

<table>
<thead>
<tr>
<th>Num Records</th>
<th>Num IP</th>
<th>Num destinations</th>
</tr>
</thead>
<tbody>
<tr>
<td>149,619,580,908</td>
<td>18,507,392</td>
<td>711,660,375</td>
</tr>
</tbody>
</table>
Web tracking
Identify of trackers

- **Label Trackers**
  - Filter lists: Easylist, Simple malvertising, Simple Ads…
- **Top 40 advertisers**

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>tudou.com</td>
<td>11</td>
<td>mobileapptracking.com</td>
<td>21</td>
</tr>
<tr>
<td>2</td>
<td>applovin.com</td>
<td>12</td>
<td>adsmogo.com</td>
<td>22</td>
</tr>
<tr>
<td>3</td>
<td>mopub.com</td>
<td>13</td>
<td>admob.com</td>
<td>23</td>
</tr>
<tr>
<td>4</td>
<td>scorecardresearch.com</td>
<td>14</td>
<td>51.la</td>
<td>24</td>
</tr>
<tr>
<td>5</td>
<td>nexage.com</td>
<td>15</td>
<td>doubleclick.net</td>
<td>25</td>
</tr>
<tr>
<td>6</td>
<td>playtomic.com</td>
<td>16</td>
<td>appsflyer.com</td>
<td>26</td>
</tr>
<tr>
<td>7</td>
<td>questionmarket.com</td>
<td>17</td>
<td>baidu.com</td>
<td>27</td>
</tr>
<tr>
<td>8</td>
<td>tapjoy.com</td>
<td>18</td>
<td>inmobi.com</td>
<td>28</td>
</tr>
<tr>
<td>9</td>
<td>2mdn.net</td>
<td>19</td>
<td>51yes.com</td>
<td>29</td>
</tr>
<tr>
<td>10</td>
<td>flurry.com</td>
<td>20</td>
<td>googlesyndication.com</td>
<td>30</td>
</tr>
</tbody>
</table>
Extract User Sessions

• User Session:
  • A target site
  • A sequence of trackers

• Threshold = 5 seconds
Issue 1: NAT Impact

- Issue:
  Several users behind one single IP
Issue 1: NAT Detection

• Observation:
For each source IP: the Average, Variance, and Maximum number of DNS requests per second
Issue 1: NAT Detection

• Solution:
  step1: Divide to 2 subsets by IP list:
  Mobile users, ADSL users
  step2: Using GMM model to classify

<table>
<thead>
<tr>
<th>category</th>
<th>Number</th>
<th>Proportion</th>
<th>Average</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mobile users</td>
<td>1</td>
<td>87%</td>
<td>0.012</td>
<td>2.45</td>
</tr>
<tr>
<td>Mobile users</td>
<td>2</td>
<td>13%</td>
<td>5.03</td>
<td>29.21</td>
</tr>
<tr>
<td>ADSL users</td>
<td>1</td>
<td>75%</td>
<td>0.046</td>
<td>3.83</td>
</tr>
<tr>
<td>ADSL users</td>
<td>2</td>
<td>25%</td>
<td>0.78</td>
<td>18.60</td>
</tr>
</tbody>
</table>

• Assignment:
  class 1: Single user IP
  class 2: NATed IP
Issue 2: DNS cache impact

• Domain Name System lookup procedure

• DNS cache affect
Issue 2: DNS cache impact

• Rescaling solution:
  Step1: Divide DNS requests into user sessions
  Step2: Merge different user sessions relative to the same target site

Maximum Session Size: 40 -> 200
Geography of Traffic share

Traffic share of target sites:
CN: 73%
US: 24%
CH: 1.3%
HK: 1.6%
Others: 0.05%

Traffic share of trackers:
US: 87%
GB: 7.2%
CN: 3.2%
CH: 1.3%
Traffic share of Top 20 sites

• Demo
http://bl.ocks.org/SJXdct/1bbc4cb97c9a5dd4a22d5df7fd97a1b4
Tracking traffic between top 100 sites and 50 trackers
Traffic exchange between top sites and trackers

10 trackers (50% of global traffic)
13 sites (70% of global traffic)
Thanks