Perspectives on Measuring and Analyzing Online Ads

Paul Barford

Spring 2014
Motivation
Outline

• Overview
  – Background, objectives and challenges

• The publisher perspective
  – m.Labs & PPV nets

• The user perspective
  – Ad Uprising & adscape
Objectives

• Investigate behavior and characteristics of the ad eco-system
  – Develop techniques for measuring aspects of the online ad eco-system
  – Compile diverse data repositories

• Develop new mechanisms that improve performance, yield, security and privacy
  – Many opportunities!

• Commercial impact
  – But it’s a cluttered space!
Publisher revenue deletions

Invalid traffic includes both clicks and impressions that Google suspects to not be the result of genuine user interest

• Standard means for valid traffic - AdWords
• Google simply notifies publishers that invalid traffic led to $XX deduction from you account
  – Large internal group that monitors traffic quality
• m.Labs – Web user & traffic quality analytics
  – Identify invalid impressions and clicks in real time

pb@cs.wisc.edu
Traffic generation

• Type “purchase web traffic” in Google
  – MANY traffic generation offerings

• Simple threats: script-based page retrieval
  – Ubiquitous - $12/10K impressions

• More complex threats: botnets
  – Geotargeting, clicks, and other characteristics
  – As much as $100/10K impressions

• Pay-per-view networks
  – Websites that load 3rd party pages in an obfuscated fashion when accessed by users
Honeypot websites

- Series of websites developed to be targets for traffic generation
  - Look and feel of a “real” site
- Instrumentation
  - Gather as much data per access as possible
Purchased traffic

BuildTraffic arrival process

AeTraffic arrival process
Deep dive into PPV nets

• When a user accesses a site running a PPV tag a pop-under window is generated
  – Typically requires a user action

• Pop-under calls PPV network server
  – Delivers details on user and site

• PPV network will deliver URL’s of sites buying traffic
  – Often to 0 height frames
  – Frequent reloads

Scope and impact of PPV nets

• Many PPV sites publish their volume
  – Average of 17.16M unique visitors and 6.29B page views per provider per day are claimed
• We searched Jan-June ‘12 Common Crawl DB for PPV tags from 10 providers
  – Over 4M PPV tags found on over 11K domains
• We used MuStats to estimate daily page views on identified pages
  – Over 168M daily page views
• Over $15M/month in wasted ad spend from 10 PPV networks alone!
Reset: the user perspective

• Research question: what display ads are being delivered to users?

• Targeting problem: select an ad to deliver to a user accessing a particular web page
  – Objectives: build awareness, click through
  – Context, geography, placement, behavior, etc.
  – Targeting mechanisms are intrinsic to online ad eco-system

• AdUprising – Identifying the Internet Adscape
  – What is being shown where and to whom?
Challenges in Adscape identification

• Scope – millions of publisher sites
• Complexities of ad campaigns
  – Demo, geo, site lists, caps, etc.
• Complexities of publisher ad placements
  – Premium, exchanges, backfill, etc.
• Complexities of ad delivery and targeting mechanisms
  – This is what we’re seeking to understand
Building a display ad crawler

- We seek to understand ad delivery by harvesting ads from a broad set of sites

- Ad crawler requirements
  - Distinguish and collect ads from other images
  - Collect related data
  - Accommodate gigantic scale and highly dynamic nature of ads in a gentle fashion
  - Personalization

- We developed a scalable, profile-based ad crawler based on Firefox/firefly
Building user profiles

- We assume *profiles* can be built based on browsing careful selection of sites
  - Assume single interest users
- Site selection is base on Alexa categories
- Profiles are created by browsing top 100 websites from an Alexa category
  - Profiles are basis for ad collection
- Profile maintenance is a challenge when gathering ads from different sites
  - We find that profiles change significantly
Collecting ads

- **Objective**: gather as many distinct ads as possible
- **Simplifications**
  - Single collection site (no geo diversity)
  - No consideration of time varying effects
  - Limited crawl to fixed period and fixed sites
- **134 sites crawled using 60 different profiles**
  - 462 [website, profile] pairs
- **Over 175K unique ads gathered from ~3.5K entities**
Ad targeting

Ratio of ads targeted by profile

Ratio of ads targeted by age group
Ads impressions by category

Time (accurate to hour)

- Arts (TV&Video)
- Computers
- Shopping
- Business
- Sports
- News

pb@cs.wisc.edu
Profiles vs. ad categories

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Arts</td>
<td>0.098</td>
<td>0.239</td>
<td>0.199</td>
<td>0.037</td>
<td>0.031</td>
<td>0.016</td>
<td>0.048</td>
<td>0.026</td>
<td>0.022</td>
<td>0.038</td>
<td>0.000</td>
<td>0.187</td>
<td>0.029</td>
<td>0.030</td>
</tr>
<tr>
<td>Business</td>
<td>0.099</td>
<td>0.299</td>
<td>0.145</td>
<td>0.015</td>
<td>0.018</td>
<td>0.011</td>
<td>0.024</td>
<td>0.044</td>
<td>0.015</td>
<td>0.083</td>
<td>0.000</td>
<td>0.173</td>
<td>0.038</td>
<td>0.036</td>
</tr>
<tr>
<td>Computers</td>
<td>0.125</td>
<td>0.099</td>
<td>0.415</td>
<td>0.021</td>
<td>0.012</td>
<td>0.008</td>
<td>0.031</td>
<td>0.007</td>
<td>0.012</td>
<td>0.030</td>
<td>0.000</td>
<td>0.193</td>
<td>0.025</td>
<td>0.022</td>
</tr>
<tr>
<td>Health</td>
<td>0.070</td>
<td>0.164</td>
<td>0.181</td>
<td>0.020</td>
<td>0.083</td>
<td>0.014</td>
<td>0.017</td>
<td>0.033</td>
<td>0.027</td>
<td>0.101</td>
<td>0.001</td>
<td>0.199</td>
<td>0.065</td>
<td>0.025</td>
</tr>
<tr>
<td>Home</td>
<td>0.102</td>
<td>0.230</td>
<td>0.287</td>
<td>0.002</td>
<td>0.011</td>
<td>0.030</td>
<td>0.003</td>
<td>0.017</td>
<td>0.034</td>
<td>0.100</td>
<td>0.010</td>
<td>0.123</td>
<td>0.032</td>
<td>0.019</td>
</tr>
<tr>
<td>News</td>
<td>0.070</td>
<td>0.313</td>
<td>0.197</td>
<td>0.014</td>
<td>0.038</td>
<td>0.008</td>
<td>0.003</td>
<td>0.050</td>
<td>0.037</td>
<td>0.060</td>
<td>0.002</td>
<td>0.116</td>
<td>0.037</td>
<td>0.056</td>
</tr>
<tr>
<td>Recreation</td>
<td>0.135</td>
<td>0.204</td>
<td>0.199</td>
<td>0.006</td>
<td>0.014</td>
<td>0.012</td>
<td>0.005</td>
<td>0.037</td>
<td>0.086</td>
<td>0.039</td>
<td>0.001</td>
<td>0.154</td>
<td>0.039</td>
<td>0.070</td>
</tr>
<tr>
<td>Reference</td>
<td>0.058</td>
<td>0.171</td>
<td>0.235</td>
<td>0.009</td>
<td>0.021</td>
<td>0.024</td>
<td>0.007</td>
<td>0.007</td>
<td>0.030</td>
<td>0.163</td>
<td>0.002</td>
<td>0.168</td>
<td>0.102</td>
<td>0.003</td>
</tr>
<tr>
<td>Shopping</td>
<td>0.106</td>
<td>0.132</td>
<td>0.231</td>
<td>0.004</td>
<td>0.012</td>
<td>0.005</td>
<td>0.004</td>
<td>0.012</td>
<td>0.013</td>
<td>0.036</td>
<td>0.000</td>
<td>0.409</td>
<td>0.021</td>
<td>0.015</td>
</tr>
<tr>
<td>Sports</td>
<td>0.087</td>
<td>0.207</td>
<td>0.249</td>
<td>0.001</td>
<td>0.025</td>
<td>0.012</td>
<td>0.012</td>
<td>0.017</td>
<td>0.032</td>
<td>0.108</td>
<td>0.001</td>
<td>0.158</td>
<td>0.029</td>
<td>0.063</td>
</tr>
</tbody>
</table>

Alexa profiles in rows, ad categories in columns, percentage of ads shown in cells

Summary and status

• Fraud is a gigantic problem in online advertising
  – From simple scripts to sophisticated bots to PPV networks
• m.Labs has developed filters to identify fraud
• m.Labs open experimental platform
  – Data repository and API for ad fraud detection

• The Internet Adscape is huge and diverse
• Initial study to characterize the Adscape
• Ad Uprising: ongoing data collection and analysis
Thank you!

• Igor Canadi
• Darja Krushevskaja
• Qiang Ma
• Muthu
• Kevin Springborn
• Charles Thomas