Insights from a 1-million-site Measurement of Online Tracking

and how our data can help your research!

Steven Englehardt  
@s_englehardt

Dillon Reisman  
@dillonthehuman

Arvind Narayanan  
@random_walker

PRINCETON UNIVERSITY
Visiting 2 websites results in 84 third parties contacted
Open Web Privacy Measurement (OpenWPM)

https://github.com/citp/OpenWPM
The Princeton Web Census

Monthly
1 Million Site Crawl

Collecting:

- Javascript Calls
- All javascript files
- HTTP Requests and Responses
- Storage (cookies, Flash, etc)
Insights from the Princeton Web Census

New metric to rank third parties

Trackers impede HTTPS adoption

Consolidation of third-parties

45 out of top 50 3rd parties cookie sync

Online Tracking: A 1-million-site Measurement and Analysis (CCS 2016)
Impact of OpenWPM and the Princeton Web Census
### Open-sourcing early can help spur adoption

<table>
<thead>
<tr>
<th>Study using OpenWPM</th>
<th>Conference</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Web Never Forgets: Persistent Tracking Mechanisms in the Wild</td>
<td>CCS</td>
<td>2014</td>
</tr>
<tr>
<td>Cognitive disconnect: Understanding Facebook Connect login permissions</td>
<td>OSN</td>
<td>2014</td>
</tr>
<tr>
<td><strong>Cookies that give you away: The surveillance implications of web tracking</strong></td>
<td>WWW</td>
<td>2015</td>
</tr>
<tr>
<td>Upgrading HTTPS in midair: HSTS and key pinning in practice</td>
<td>NDSS</td>
<td>2015</td>
</tr>
<tr>
<td>Web Privacy Census</td>
<td>Tech Science</td>
<td>2015</td>
</tr>
<tr>
<td>Variations in Tracking in Relation to Geographic Location</td>
<td>W2SP</td>
<td>2015</td>
</tr>
<tr>
<td>No Honor Among Thieves: A Large-Scale Analysis of Malicious Web Shells</td>
<td>WWW</td>
<td>2016</td>
</tr>
<tr>
<td><strong>Online Tracking: A 1-million-site Measurement and Analysis</strong></td>
<td>CCS</td>
<td>2016</td>
</tr>
<tr>
<td>Dial One for Scam: Analyzing and Detecting Technical Support Scams</td>
<td>NDSS</td>
<td>2017</td>
</tr>
</tbody>
</table>
Open-sourcing early can help spur adoption

<table>
<thead>
<tr>
<th>Study using OpenWPM</th>
<th>Conference</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Web Never Forgets: Persistent Tracking Mechanisms in the Wild</td>
<td>CCS</td>
<td>2014</td>
</tr>
<tr>
<td>Cognitive disconnect: Understanding Facebook Connect login permissions</td>
<td>OSN</td>
<td>2014</td>
</tr>
<tr>
<td>Cookies that give you away: The surveillance implications of web tracking</td>
<td>WWW</td>
<td>2015</td>
</tr>
<tr>
<td>Upgrading HTTPS in midair: HSTS and key pinning in practice</td>
<td>NDSS</td>
<td>2015</td>
</tr>
<tr>
<td>Web Privacy Census</td>
<td>Tech Science</td>
<td>2015</td>
</tr>
<tr>
<td>Variations in Tracking in Relation to Geographic Location</td>
<td>W2SP</td>
<td>2015</td>
</tr>
<tr>
<td>No Honor Among Thieves: A Large-Scale Analysis of Malicious Web Shells</td>
<td>WWW</td>
<td>2016</td>
</tr>
<tr>
<td>Dial One for Scam: Analyzing and Detecting Technical Support Scams</td>
<td>NDSS</td>
<td>2017</td>
</tr>
</tbody>
</table>
Measurement work can influence standards

https://w3c.github.io/fingerprinting-guidance
How can measurement influence adoption of new tracking techniques?

Canvas

WebRTC

Audio

Battery

https://webtransparency.cs.princeton.edu/webcensus/
Canvas fingerprinting returns in the absence of measurement

May 2014: 5% of sites

Aug 2014: ~0.1% of sites

Jan 2016: 2.6% of sites

Percentage of the Alexa top 100k sites
AudioContext fingerprinting the Tor Browser

271 samples from the Tor Browsers
- 7 distinct fingerprints (2 fingerprints account for 80% of samples)
- Overlap with fingerprints from Firefox shows these largely reveal OS of device
Browsers remove BatteryStatus API citing privacy
Browsers remove BatteryStatus API citing privacy
Blocking tools miss less popular trackers & fingerprinters

Ghostery
AdBlock Plus
Third-party cookie blocking

https://webtransparency.cs.princeton.edu/webcensus/
Future directions of the Web Census

- Complete instrumentation of JS
- Static + dynamic analysis of trackers
- Detection of trackers with ML
- Blocking tools for users
Our data is available!

## Data

The data is available as bzipped PostgreSQL dumps. The schema file used in all of the datasets is available [here](https://webtransparency.cs.princeton.edu/webcensus/index.html#data).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Million Site Stateless</td>
<td>Parallel Stateless Crawl</td>
</tr>
<tr>
<td>100k Site Stateful</td>
<td>Parallel Stateful Crawl -- 10,000 site seed profile</td>
</tr>
<tr>
<td>10k Site ID Detection (1)</td>
<td>Sequential Stateful Crawl -- Flash enabled -- Synced with ID Detection (2)</td>
</tr>
<tr>
<td>10k Site ID Detection (2)</td>
<td>Sequential Stateful Crawl -- Flash enabled -- Synced with ID Detection (1)</td>
</tr>
<tr>
<td>55k Site Stateless with cookie blocking</td>
<td>Parallel Stateless Crawl -- Firefox set to block all third-party cookies</td>
</tr>
<tr>
<td>55k Site Stateless with Ghostery</td>
<td>Parallel Stateless Crawl -- Ghostery extension installed and set to block all possible trackers</td>
</tr>
<tr>
<td>55k Site Stateless with HTTPS Everywhere</td>
<td>Parallel Stateless Crawl -- HTTPS Everywhere installed</td>
</tr>
</tbody>
</table>
Making data exploration easier

**Problem:** Querying our data — and making sense of the output — comes with a steep learning curve

```sql
query = "SELECT DISTINCT res.url, v1.name, v1.value FROM " \\
    "http_responses_view as res " \\
    "LEFT JOIN http_response_cookies_view as v1 " \\
    "ON v1.response_id = res.id " \\
    "WHERE res.top_url = %s AND v1.name != '' " \\
    " union " \\
    "SELECT DISTINCT req.url, v2.name, v2.value FROM " \\
    "http_requests_view as req " \\
    "LEFT JOIN http_request_cookies_view as v2 " \\
    "ON v2.request_id = req.id " \\
    "WHERE req.top_url = %s AND v2.name != ''"
```
Making data exploration easier

**Problem:** Querying our data — and making sense of the output — comes with a steep learning curve

```python
query = "SELECT DISTINCT res.url, v1.name, v1.value FROM " \\
"http_responses_view as res " \\
"LEFT JOIN http_response_cookies_view as v1 " \\
"ON v1.response_id = res.id " \\
"WHERE res.top_url = %s AND v1.name != ' ' " \\
"union " \\
"SELECT DISTINCT req.url, v2.name, v2.value FROM " \\
"http_requests_view as req " \\
"LEFT JOIN http_request_cookies_view as v2 " \\
"ON v2.request_id = req.id " \\
"WHERE req.top_url = %s AND v2.name != ' ' "
```

**Solution:** We are making our data analysis libraries available via Jupyter Notebook!
Census.py

- We’ll give you access to our Notebook server, complete with tools that will provide an abstraction layer over our web census data.
- Example API:
  - `get_third_party_responses_by_domain(domain)`
  - `get_cookie_syncs_on_domain(domain)`
  - `is_tracker(url, first_party=False, is_js=True, is_img=False)`
  - `get_trackers(domain)`

Get access: https://groups.google.com/forum/#!forum/web-census-explorers
Example: Getting third party requests by domain

```python
In [14]:
results = census.get_third_party_responses_by_domain(con, 'http://nytimes.com')

third_party_trackers = {results[x][url_ps] for x in results if results[x][is_tracker]}

print "Number of third_party trackers on domain: " + str(len(third_party_trackers))

for url in results:
    print url
    print \tIs a script? ' + str(results[url][is_js])
    print \tIs a tracker? ' + str(results[url][is_tracker])
    print \tPS+1: ' + results[url][url_ps]
```

Get access: https://groups.google.com/forum/#!forum/web-census-explorers
Example: Getting third party requests by domain

```python
In [14]: results = census.get_third_party_responses_by_domain(con, 'http://nytimes.com')

third_party_trackers = {results[x]['url_ps'] for x in results if results[x]['is_tracker']}

print "Number of third_party trackers on domain: " + str(len(third_party_trackers))

for url in results:
    print url
    print \tIs a script? ' + str(results[url]['is_js'])
    print \tIs a tracker? ' + str(results[url]['is_tracker'])
    print \tPS+1: ' + results[url]['url_ps']
```

Number of third_party trackers on domain: 27
https://static01.nyt.com/images/2016/08/18/business/19WHEELS-ss-slide-0LQP/19WHEELS-ss-slide-0LQP-thumbStandard.jpg
Is a script? False
Is a tracker? False
PS+1: nyt.com

https://static01.nyt.com/images/2016/03/insider/03insider-savage01/03insider-savage01-thumbStandard-v2.jpg
Is a script? False
Is a tracker? False
PS+1: nyt.com

https://tpc.googlesyndication.com/simgad/15453271384116304559
Is a script? False
Is a tracker? True
PS+1: googlesyndication.com

http://googleleads.g.doubleclick.net/pagead/gen_204?id=wfocus&gclid=CMqE0Km81c4CFVVeDAod0MoOTQ&fg=1
Is a script? False
Is a tracker? True
```

Get access: https://groups.google.com/forum/#!forum/web-census-explorers
Thanks for listening!

Full Paper:
senglehardt.com/papers/ccs16_online_tracking.pdf

Princeton Web Census Data and Analysis:
webtransparency.cs.princeton.edu/webcensus/

Collaborate:
webtap.princeton.edu/research/

Email: ste@cs.princeton.edu    Twitter: @s_englehardt    Web: senglehardt.com
dreisman@princeton.edu