The Web Never Forgets

Persistent Tracking mechanisms in the Wild

by Gunes Acar, Christian Eubank, Steven Englehardt, Marc Juarez, Arvind Narayanan, Claudia Diaz
Contents

- Introduction
  - What is Web Tracking?
  - Motivations
  - Security Consequences
- Web Tracking Techniques & Situations in 2016
  - Canvas API
  - Evercookies
  - Cookie Syncing
- Mitigation Discussion
Information Available in a Web Session

High Entropy
- Session time
- Plugins
- Credentials for authentication
- Cookies
- Hardwares that used

Low Entropy
- Browser agent (Mozilla ver., for example)
- Area code, IP address
- Device type, screen size
- OS/architecture, fonts, software installed

Bolded words mean they can be used for web tracking.

Refer to https://browserleaks.com
Motivations

As a website owner
- Personal preferences
- Browser support, avoid potential incompatibility
- Fight against fraud reviews

As an attacker (3rd party)
- Ad service, real time bidding
- Tracking high profile individuals/groups
Security Consequences

- Sensitive data leaked
- Legal issues
Web Tracking Techniques in 2016

- Canvas Fingerprinting
  - Invisible images that give subtle differences in the rendering of the same text.
- Evercookies
  - Local storage (Flash, IndexedDB, LocalStorage, CacheStorage, etc.).
- Cookie Syncing
  - Workaround to the Same-Origin Policy that allows different trackers to share information
  - SOP recap
    - Cookies are stored for domainA/pathA cannot be accessed by domainA/pathB or domainB/*.
Canvas Fingerprinting

Why can Canvas APIs be abused for fingerprinting?

Figure 1: Canvas fingerprinting basic flow of operations
Canvas Fingerprinting

- **Data Collection**
  - Instrument Firefox to collect scripts.
  - Attribute Canvas to script files.

- **False Positive Removal**
  - Scripts that both read and write the Canvas.
  - Images whose size is greater than \(16 \times 16\) pixels.
  - Images that are not compressed.

- **Result**
  - More than 5.5% of Alexa Top 100k actively ran fingerprinting scripts.
  - The less popular a website is, the more Canvas fingerprinting it employs.
  - 95% of the scripts belongs to one provider, addthis.com.

<table>
<thead>
<tr>
<th>Rank interval</th>
<th>% of sites with canvas fingerprinting scripts</th>
</tr>
</thead>
<tbody>
<tr>
<td>([1, 1K])</td>
<td>1.80</td>
</tr>
<tr>
<td>([1K, 10K])</td>
<td>4.93</td>
</tr>
<tr>
<td>([10K, 100K])</td>
<td>5.73</td>
</tr>
</tbody>
</table>

Table 2: Percentage of sites that include canvas fingerprinting scripts on the homepage, found in top 100K Alexa sites divided in intervals of variable length. Websites in the 1 to 1K rank interval are 2.5 times less likely to embed a canvas fingerprinting script than a site within 1K-10K interval.
What are the Potential Reasons?

The less popular a website is, the more Canvas fingerprinting it employs.

- More popular ones have more conscious about their users
- More popular ones have better ways to track user other than fingerprinting

<table>
<thead>
<tr>
<th>Rank interval</th>
<th>% of sites with canvas fingerprinting scripts</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1, 1K)</td>
<td>1.80</td>
</tr>
<tr>
<td>[1K, 10K)</td>
<td>4.93</td>
</tr>
<tr>
<td>[10K, 100K]</td>
<td>5.73</td>
</tr>
</tbody>
</table>

Table 2: Percentage of sites that include canvas fingerprinting scripts on the homepage, found in top 100K Alexa sites divided in intervals of variable length. Websites in the 1 to 1K rank interval are 2.5 times less likely to embed a canvas fingerprinting script than a site within 1K-10K interval.

Figure 3: Frequency of canvas fingerprinting scripts on the home pages of Top Alexa 100K sites.
What are the Potential Reasons?

Why use a perfect pangram?

- All letters can be tested
- Minimize its impact
- More information collected
- Include syntax checking
- Cloaking by target website

<table>
<thead>
<tr>
<th>Fingerprinting script</th>
<th>Number of including sites</th>
<th>Text drawn into the canvas</th>
</tr>
</thead>
<tbody>
<tr>
<td>ct1.addthis.com/static/r07/core130.js</td>
<td>5282</td>
<td>Cwm fjordbank glyphs vext quiz, @</td>
</tr>
<tr>
<td>iligatas.com/script/fingerprint.min.js</td>
<td>115</td>
<td><a href="http://valve.github.io">http://valve.github.io</a></td>
</tr>
<tr>
<td>src.kitcode.net/fp2.js</td>
<td>68</td>
<td><a href="http://valve.github.io">http://valve.github.io</a></td>
</tr>
<tr>
<td>admicro1.venmedia.vn/fingerprint/fipp.js</td>
<td>31</td>
<td><a href="http://admicro.vn/">http://admicro.vn/</a></td>
</tr>
<tr>
<td>amazonaws.com/af-bdas/bquery.js</td>
<td>26</td>
<td>Centillion</td>
</tr>
<tr>
<td>*shorte.st/js/packed/smeadvert-intermediate-ad.js</td>
<td>14</td>
<td><a href="http://valve.github.io">http://valve.github.io</a></td>
</tr>
<tr>
<td>stat.ringer.ca/js/fingerprint.min.js</td>
<td>4</td>
<td><a href="http://valve.github.io">http://valve.github.io</a></td>
</tr>
<tr>
<td>cya2.net/js/STAT/89946.js</td>
<td>3</td>
<td>ABCDEFGHIJKLMNOPQRSTUVWXYZ abcdefghijklmnopqrstuvwxyz0123456789+/</td>
</tr>
<tr>
<td>pof.com</td>
<td>2</td>
<td><a href="http://www.plentyofish.com">http://www.plentyofish.com</a></td>
</tr>
<tr>
<td>*.rackedn.com/mongoose.fp.js</td>
<td>2</td>
<td><a href="http://api.gonorthloads.com">http://api.gonorthloads.com</a></td>
</tr>
<tr>
<td>9 others*</td>
<td>9</td>
<td>(Various)</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>5559</strong></td>
<td><strong>5548 unique¹</strong></td>
</tr>
</tbody>
</table>

Table 1: Canvas fingerprinting domains found on Top Alexa 100K sites.

*: Some URLs are truncated or omitted for brevity. See Appendix for the complete list of URLs.
1: Some sites include canvas fingerprinting scripts from more than one domain.
Evercookies

- Detecting the User IDs
  - Use two different instances to crawl one website.
  - Conduct a differential analysis on the cookies.
- Flash is Gone! [1]
  - Flash Cookies Respawning HTTP Cookies
  - HTTP Cookies Respawning Flash Cookies
- IndexedDB for Web Tracking [2]
  - Used to store customized configuration, for example, user’s theme preference.
  - Don’t expire unless deleted explicitly.


Figure 2: Respawning HTTP cookies by Flash evercookies: (a) the webpage stores an HTTP and a Flash cookie (LSO), (b) the user removes the HTTP cookie, (c) the webpage respawns the HTTP cookie by copying the value from the Flash cookie.
Evercookies

How can sites keep using evercookies and evade detection?

- Obfuscate the ID
- Dynamic IDs
  - Simple idea: id = random number || id

Figure 2: Respawning HTTP cookies by Flash evercookies: (a) the webpage stores an HTTP and a Flash cookie (LSO), (b) the user removes the HTTP cookie, (c) the webpage respawn the HTTP cookie by copying the value from the Flash cookie.
Cookie Syncing

- User ID Cookie owned by one domain A is sent to another domain B.
- Domain B is able to associate the User ID in domain A with its own User ID.
Cookie Syncing

- Detecting Cookie Synchronization
  - If an ID appears in a requested URL, the requested domain learns the ID.
  - If an ID appears in the referrer URL, the requested domain and location domain (if it exists) learn the ID.
  - If an ID appears in the location URL, the requested domain learns the ID.

- Crawler Setup
  - Allowing all cookies.
  - Allowing all cookies, but “Do Not Track”. (trivial impact, omitted)
  - Blocking 3rd-party cookies.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Third party cookie policy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Allow</td>
</tr>
<tr>
<td># IDs</td>
<td>1308</td>
</tr>
<tr>
<td># ID cookies</td>
<td>1482</td>
</tr>
<tr>
<td># IDs in sync</td>
<td>435</td>
</tr>
<tr>
<td># ID cookies in sync</td>
<td>596</td>
</tr>
<tr>
<td># (First*) Parties in sync</td>
<td>(407)</td>
</tr>
<tr>
<td># IDs known per party</td>
<td>1/2.0/1/33</td>
</tr>
<tr>
<td># Parties knowing an ID</td>
<td>2/3.4/2/43</td>
</tr>
</tbody>
</table>

Table 4: Comparison of high-level cookie syncing statistics when allowing and disallowing third-party cookies (top 3,000 Alexa domains). The format of the bottom two rows is minimum/mean/median/maximum. *Here we define a first-party as a site which was visited in the first-party context at any point in the crawl.
Cookie Syncing

What can a tracker learn about users?
Cookie Syncing

Comparison

- It can recover more of a user’s history w/ 3rd-party cookie.
- By merging databases based on synced IDs, more sites are able to recover user’s browsing history.
  - But how prevalent backend database merges is unknown.

Figure 5: Proportions of user history known when allowing and blocking third party cookies under the two different merging schemes. Note that since the x-axis is sorted by the proportion of a user’s history that a domain can recover, the domains may appear in different orders for the different models.
How to Prevent History Discovering by Cookie Syncing?
Why is Opt-out Not Effective?
Mitigation

Effectiveness

Friendliness

Customised browser

Law forced opt-out policy

Terminal browser

Clear cookie

Noscript