Detecting and Defending Against Third-Party Tracking on the Web

Franziska Roesner, Tadayoshi Kohno, David Wetherall

April 26, 2012
NSDI
Third-Party Web Tracking

Bigger browsing profiles = *increased value* for trackers
= *reduced privacy* for users

(Hypothetical tracking relationships only.)
Tracking is Complicated

• Much discussion of tracking, but limited understanding of how it actually works.

• Our goals:
  – Understand the tracking ecosystem.
    • How is tracking actually done in the wild?
    • What kinds of browsing profiles do trackers compile?
    • How effective are defenses available to users?
  – Address gaps with new defense (ShareMeNot).
Outline

• How Tracking Works
  – Tracking Mechanisms
  – Tracking Taxonomy

• Measurements

• Defenses
Mechanisms Required By Trackers

• Ability to store user identity in the browser
  – Browser cookies
  – HTML5 LocalStorage and Flash cookies (LSOs)
  – Not considering more exotic storage mechanisms or approximate fingerprinting

• Ability to communicate visited page and user identity back to tracker
  – Identity: Cookies attached to requests
  – Visited page: HTTP referrers
  – Both: scripts that embed information in URLs
Tracking: The Simple Version

• **Within-Site:** First-party cookies are used to track repeat visits to a site.

• **Cross-Site:** Third-party cookies are used by trackers included in other sites to create profiles.

9:30am: user 789 visited site1.com

9:31am: user 789 visited site2.com
# Our Tracking Taxonomy

<table>
<thead>
<tr>
<th>Name</th>
<th>Scope</th>
<th>User Visits Directly?</th>
<th>Overview</th>
</tr>
</thead>
<tbody>
<tr>
<td>N/A</td>
<td>Within-Site</td>
<td>Yes</td>
<td>Site does its own on-site analytics.</td>
</tr>
<tr>
<td><strong>Evolution: Embedding analytics libraries</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Analytics</td>
<td>Within-Site</td>
<td>No</td>
<td>Site uses third-party analytics engine (e.g., Google Analytics).</td>
</tr>
<tr>
<td>Vanilla</td>
<td>Cross-Site</td>
<td>No</td>
<td>Site embeds third-party tracker that uses third-party storage (e.g., Doubleclick).</td>
</tr>
<tr>
<td><strong>Evolution: Third-party cookie blocking</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Forced</td>
<td>Cross-Site</td>
<td>Yes (forced)</td>
<td>Site embeds third-party tracker that forced the user to visit directly (e.g., via popup).</td>
</tr>
<tr>
<td>Referred</td>
<td>Cross-Site</td>
<td>No</td>
<td>Tracker relies on another cross-site tracker to leak unique identifier values.</td>
</tr>
<tr>
<td>Personal</td>
<td>Cross-Site</td>
<td>Yes</td>
<td>Site embeds third-party tracker that the user otherwise visits directly (e.g., Facebook).</td>
</tr>
</tbody>
</table>
Quirks of Third-Party Cookie Blocking

- Option blocks the **setting** of third-party cookies: all browsers
- Option blocks the **sending** of third-party cookies: **only Firefox**

- Result: Once a third-party cookie is somehow set, **it can be used** (in most browsers).
Forced Tracking

High-level point:
On most browsers, if a tracker can ever set a cookie, third-party cookie blocking is rendered ineffective.
## Our Tracking Taxonomy

<table>
<thead>
<tr>
<th>Type (Name)</th>
<th>Scope</th>
<th>User Visits Directly?</th>
<th>Overview</th>
</tr>
</thead>
<tbody>
<tr>
<td>N/A</td>
<td>Within-Site</td>
<td>Yes</td>
<td>Site does its own on-site analytics.</td>
</tr>
<tr>
<td><strong>Evolution: Embedding analytics libraries</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Analytics</td>
<td>Within-Site</td>
<td>No</td>
<td>Site uses third-party analytics engine (e.g., Google Analytics).</td>
</tr>
<tr>
<td><strong>Evolution: Third-party cookie blocking</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vanilla</td>
<td>Cross-Site</td>
<td>No</td>
<td>Site embeds third-party tracker that uses third-party storage (e.g., Doubleclick).</td>
</tr>
<tr>
<td>Forced</td>
<td>Cross-Site</td>
<td>Yes</td>
<td>Site embeds third-party tracker that forced the user to visit directly (e.g., via popup).</td>
</tr>
<tr>
<td><strong>Evolution: Complex ad networks</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Referred</td>
<td>Cross-Site</td>
<td>No</td>
<td>Tracker relies on another cross-site tracker to leak unique identifier values.</td>
</tr>
<tr>
<td>Personal</td>
<td>Cross-Site</td>
<td>Yes</td>
<td>Site embeds third-party tracker that the user otherwise visits directly (e.g., Facebook).</td>
</tr>
</tbody>
</table>
Referred Tracking

High-level point:
One tracker with client-side state can enable tracking by **partners without client-side state**.

2:34pm: site1.com: user 522
2:35pm: site2.com: user 522
# Our Tracking Taxonomy

<table>
<thead>
<tr>
<th>Type (Name)</th>
<th>Scope</th>
<th>User Visits Directly?</th>
<th>Overview</th>
</tr>
</thead>
<tbody>
<tr>
<td>N/A</td>
<td>Within-Site</td>
<td>Yes</td>
<td>Site does its own on-site analytics.</td>
</tr>
<tr>
<td><strong>Evolution: Embedding analytics libraries</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Analytics</td>
<td>Within-Site</td>
<td>No</td>
<td>Site uses third-party analytics engine (e.g., Google Analytics).</td>
</tr>
<tr>
<td>Vanilla</td>
<td>Cross-Site</td>
<td>No</td>
<td>Site embeds third-party tracker that uses third-party storage (e.g., Doubleclick).</td>
</tr>
<tr>
<td><strong>Evolution: Third-party cookie blocking</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Forced</td>
<td>Cross-Site</td>
<td>Yes (forced)</td>
<td>Site embeds third-party tracker that forced the user to visit directly (e.g., via popup).</td>
</tr>
<tr>
<td><strong>Evolution: Complex ad networks</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Referred</td>
<td>Cross-Site</td>
<td>No</td>
<td>Tracker relies on another cross-site tracker to leak unique identifier values.</td>
</tr>
<tr>
<td>Personal</td>
<td>Cross-Site</td>
<td>Yes</td>
<td>Site embeds third-party tracker that the user otherwise visits directly (e.g., Facebook).</td>
</tr>
<tr>
<td><strong>Evolution: Social networks</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Personal Tracking

- **Just loading** these buttons (not clicking on them) enables tracking.
- Users **visit these sites directly**.
- This tracking is often **not anonymous** (linked to accounts).
# Our Tracking Taxonomy

<table>
<thead>
<tr>
<th>Type (Name)</th>
<th>Scope</th>
<th>User Visits Directly?</th>
<th>Overview</th>
</tr>
</thead>
<tbody>
<tr>
<td>N/A</td>
<td>Within-Site</td>
<td>Yes</td>
<td>Site does its own on-site analytics.</td>
</tr>
<tr>
<td><strong>Evolution: Embedding analytics libraries</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Analytics</td>
<td>Within-Site</td>
<td>No</td>
<td>Site uses third-party analytics engine (e.g., Google Analytics).</td>
</tr>
<tr>
<td>Vanilla</td>
<td>Cross-Site</td>
<td>No</td>
<td>Site embeds third-party tracker that uses third-party storage (e.g., Doubleclick).</td>
</tr>
<tr>
<td><strong>Evolution: Third-party cookie blocking</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Forced</td>
<td>Cross-Site</td>
<td>Yes</td>
<td>Site embeds third-party tracker that forced the user to visit directly (e.g., via popup).</td>
</tr>
<tr>
<td>Referred</td>
<td>Cross-Site</td>
<td>No</td>
<td>Tracker relies on another cross-site tracker to leak unique identifier values.</td>
</tr>
<tr>
<td>Personal</td>
<td>Cross-Site</td>
<td>Yes</td>
<td>Site embeds third-party tracker that the user otherwise visits directly (e.g., Facebook).</td>
</tr>
<tr>
<td><strong>Evolution: Complex ad networks</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Evolution: Social networks</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Outline

• How Tracking Works
  – Tracking mechanisms
  – Tracking taxonomy

• Measurements

• Defenses
Measurement Tool: TrackingTracker

• Firefox add-on
• **Based on taxonomy** of client-side mechanisms
• Crawls the web, **automatically categorizing** trackers
• Monitors:
  – Third-party requests
  – Cookies, HTML5 LocalStorage, Flash LSOs
    (considers state that changes across clean measurement runs)
  – Identifier leaks
Measurement Study

• 3 data sets
  – Alexa Top 500
    • 5 pages per domain: main page and up to 4 links
  – Alexa Non-Top 500
    • Sites ranked #501, #601, #701, etc.
    • 5 pages per domain: main page and up to 4 links
  – AOL search logs
    • 300 unique queries for 35 random users
Tracking Prevalence (Top 500)

- 524 unique trackers on 500 domains

- 457 domains (91%) embed at least one tracker.
  (97% of those include at least one cross-site tracker.)

- 50% of domains embed between 4 and 5 trackers.

- One domain includes 43 trackers.
Top 20 Trackers on Top 500 Domains

- **Within-Site**
- **Cross-Site (Personal)**
- **Cross-Site (Anonymous)**

- google-analytics.com: 297
- doubleclick.net: 189
- facebook.com: 154
- google.com: 149
- quantserve.com: 109
- twitter.com: 105
- atdmt.com: 93
- imrworldwide.com: 81
- revsci.net: 60
- advertising.com: 45
- addthis.com: 44
- adnxs.com: 40
- serving-sys.com: 34
- youtube.com: 33
- addthiscdn.com: 32
- bluekai.com: 30
- mediaplex.com: 29
- 91bluemedia.com: 27
- 26e3.com: 26
Each line represents one user.

Doubleclick: Avg 39% (Max 66%)
Facebook: Avg 23% (Max 45%)
Google: Avg 21% (Max 61%)
LocalStorage and Flash Cookies

- Surprisingly little use of these mechanisms!
- Of 524 trackers on Alexa Top 500:
  - Only 5 set unique identifiers in LocalStorage
  - 35 set unique identifiers in Flash cookies
- Respawnning:
  - LS → Cookie: 1 case; Cookie → LS: 3 cases
  - Flash → Cookie: 6 cases; Cookie → Flash: 7 cases
Outline

• How Tracking Works
  – Tracking mechanisms
  – Tracking taxonomy

• Measurements

• Defenses
Defenses to Reduce Tracking

• We explore several in the paper:
  – Third-party cookie blocking
  – Do Not Track header
  – Popup blocking
  – Clearing client-side state
  – Disabling JavaScript
  – Private browsing mode
Personal Tracking Revisited

• Most popular, based on measurements:
  – Facebook, Google, Twitter, AddThis, YouTube, LinkedIn, Digg, Stumbleupon

• Third-party cookie blocking is ineffective.

• Existing browser extension solutions remove the buttons (undesirable to some users).

• Can we reduce tracking but allow use?
ShareMeNot

http://sharemenot.cs.washington.edu

• A browser extension that protects against tracking from third-party social media buttons while still allowing them to be used.

• Firefox version: removes cookies from relevant requests until user clicks button.
  – Similar: Priv3 Firefox add-on

• Chrome version: replace buttons with local stand-in button until user click.
## Effectiveness of ShareMeNot (Top 500)

<table>
<thead>
<tr>
<th>Tracker</th>
<th>Without ShareMeNot</th>
<th>With ShareMeNot</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facebook</td>
<td>154</td>
<td>9</td>
</tr>
<tr>
<td>Google</td>
<td>149</td>
<td>15</td>
</tr>
<tr>
<td>Twitter</td>
<td>93</td>
<td>0</td>
</tr>
<tr>
<td>AddThis</td>
<td>34</td>
<td>0</td>
</tr>
<tr>
<td>YouTube</td>
<td>30</td>
<td>0</td>
</tr>
<tr>
<td>LinkedIn</td>
<td>22</td>
<td>0</td>
</tr>
<tr>
<td>Digg</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>Stumbleupon</td>
<td>6</td>
<td>0</td>
</tr>
</tbody>
</table>
Summary

• Introduced taxonomy of tracking behavior for any client-side identifiers.
  – Analytics, Vanilla, Forced, Referred, Personal
• Studied tracking in the wild with browser measurements.
  – Revealed rich tracking ecosystem.
  – Results can assist informed broader discussions.
• Developed ShareMeNot, a new privacy-enhancing defense for personal tracking.

http://sharemenot.cs.washington.edu/