Reverse-engineering Online Tracking

From niche research field to easy-to-use tool

Steven Englehardt

webtap.princeton.edu
DELETE COOKIES?!
Evercookies

Respawn cookies using alternative locations

- Flash cookies, HTML5 localStorage, ETags, etc.
If you’re going to track me, please use cookies

Ed Felten
July 7th, 2009
freedom-to-tinker.com

https://freedom-to-tinker.com/blog/felten/if-youre-going-track-me-please-use-cookies/
Panopticlick

How Unique — and Trackable — Is Your Browser?

Is your browser configuration rare or unique? If so, websites may be able to track you, even if you limit or disable cookies.

Panopticlick tests your browser to see how unique it is based on the information it will share with sites it visits. Click below and you will be given a uniqueness score, letting you see how easily identifiable you might be as you surf the web.

Only anonymous data will be collected by this site.
Canvas Fingerprinting

流程概述:
1. FillText()
2. ToDataURL()
   - data:image/png;base64,iVBOR
     w0KGgoAAAANSUhEUgAAA
     SwAAACWCAAYAAABkW7XS
     AAAeq0leXgV1d0...
3. Hash()
2009

If you’re going to track me, please use cookies
If you’re going to track me, please use **cookies**

If you’re going to track me, please use **browser storage**
Cross-Device Tracking

NOV 16, 2015
CONSTITUTION CENTER
400 7th St SW, Washington, DC 20024 | Directions & Nearby

If you’re going to track me, please *use cookies*

2009

2010

If you’re going to track me, please *use browser storage*
If you’re going to track me, please use

2009
cookies

2010

use browser storage

2015
limit it to one device
If you’re going to track me, please limit it to one device.
Measurement can help!
Web measurement hurdles

1. Engineering Debt
<table>
<thead>
<tr>
<th>Paper</th>
<th>Targets</th>
<th>Infrastructure</th>
<th>Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leakage of PII via OSN ('09)</td>
<td>PII leaks</td>
<td>M*</td>
<td>LHH</td>
</tr>
<tr>
<td>Privacy diffusion on the web ('09)</td>
<td>Tracking: cookies</td>
<td>F, PS</td>
<td>Proxy</td>
</tr>
<tr>
<td>Challenges in measuring ('10)</td>
<td>Personalization: ads</td>
<td>Proxy</td>
<td>730 queries</td>
</tr>
<tr>
<td>Flash cookies and privacy ('10)</td>
<td>Tracking: cookies, LSOs</td>
<td>M*</td>
<td>Proxy</td>
</tr>
<tr>
<td>Privacy leakage in mOSN ('10)</td>
<td>PII leaks</td>
<td>M*</td>
<td>Proxy</td>
</tr>
<tr>
<td>Flash cookies and privacy II ('11)</td>
<td>Tracking: cookies, LSOs</td>
<td>M*</td>
<td>Proxy</td>
</tr>
<tr>
<td>Privacy leakage vs. protection measures ('11)</td>
<td>PII leaks</td>
<td>M*</td>
<td>Proxy</td>
</tr>
<tr>
<td>Respawn HTTP Cookies ('11)</td>
<td>Tracking: cookies, LSOs</td>
<td>UA*</td>
<td>FourthParty</td>
</tr>
<tr>
<td>Self-help tools ('11)</td>
<td>Tracking: cookies</td>
<td>UA*</td>
<td>FourthParty</td>
</tr>
<tr>
<td>Where everybody knows your username ('11)</td>
<td>PII leaks</td>
<td>M*</td>
<td>FourthParty</td>
</tr>
<tr>
<td>Detecting and defending ('12)</td>
<td>Tracking: cookies</td>
<td>FF, TT</td>
<td>TrackingTracker</td>
</tr>
<tr>
<td>Detecting price and search discrimination ('12)</td>
<td>Price discrimination</td>
<td>SA, CH, IE, JS</td>
<td>Proxy</td>
</tr>
<tr>
<td>Mac users steered to pricier hotels ('12)</td>
<td>Personalization: ads</td>
<td>F, SL</td>
<td></td>
</tr>
<tr>
<td>Measuring the effectiveness of privacy tools ('12)</td>
<td>Personalization: ads</td>
<td>F, SL</td>
<td></td>
</tr>
<tr>
<td>Websites vary prices ('12)</td>
<td>Personalization: prices, deals</td>
<td>Proxy</td>
<td></td>
</tr>
<tr>
<td>What they do with what they know ('12)</td>
<td>Personalization: ads</td>
<td>Proxy</td>
<td></td>
</tr>
<tr>
<td>AdReveal ('13)</td>
<td>Tracking: fingerprinting</td>
<td>F, CH</td>
<td>Custom plugin</td>
</tr>
<tr>
<td>Cookieless monster ('13)</td>
<td>Tracking: fingerprinting, Ads</td>
<td>M, UA</td>
<td>2184 names</td>
</tr>
<tr>
<td>Crowd-assisted search ('13)</td>
<td>Tracking: fingerprinting</td>
<td>CR, SL, CJ, PJ</td>
<td>Proxy</td>
</tr>
<tr>
<td>Discrimination in online ad delivery ('13)</td>
<td>Personalization: search</td>
<td>Browser Code</td>
<td></td>
</tr>
<tr>
<td>FPDetective ('13)</td>
<td>Personalization: search</td>
<td>Browser Code</td>
<td></td>
</tr>
<tr>
<td>Know your personalization ('13)</td>
<td>Tracking: cookies, JS</td>
<td>F, SL</td>
<td>Custom plugin</td>
</tr>
<tr>
<td>Measuring personalization of web search ('13)</td>
<td>General tracking</td>
<td>F, PS, SL</td>
<td>5K sites</td>
</tr>
<tr>
<td>Selling off privacy at auction ('13)</td>
<td>Personalization: ads</td>
<td>F, SL</td>
<td>FourthParty</td>
</tr>
<tr>
<td>Shining the floodlights ('13)</td>
<td>Personalization: search</td>
<td>F, SL</td>
<td>FourthParty</td>
</tr>
<tr>
<td>Statistical approach ('13)</td>
<td>Tracking: cookies</td>
<td>F, SL</td>
<td>FourthParty</td>
</tr>
<tr>
<td>Adscape ('14)</td>
<td>Price discrimination and steering</td>
<td>F, SL</td>
<td>FourthParty</td>
</tr>
<tr>
<td>Bobble ('14)</td>
<td>Price discrimination of airline tickets</td>
<td>F, SL</td>
<td>FourthParty</td>
</tr>
<tr>
<td>Information flow experiments ('14)</td>
<td>Price discrimination</td>
<td>F, SL</td>
<td>FourthParty</td>
</tr>
<tr>
<td>Third-party OSN applications ('14)</td>
<td>Price discrimination</td>
<td>F, SL</td>
<td>FourthParty</td>
</tr>
<tr>
<td>Price discrimination and steering ('14)</td>
<td>Price disc, steering</td>
<td>PJ</td>
<td></td>
</tr>
<tr>
<td>Price discrimination of airline tickets ('14)</td>
<td>Price discrimination</td>
<td>PJ</td>
<td></td>
</tr>
</tbody>
</table>

*FF = Firefox, CH = Chrome, CR = Chromium, IE = Internet Explorer, SA = Safari, SL = Selenium, JS = JavaScript, PJ = PhantomJS, PS = PageStats, PY = Python, TT = TrackingTracker, CJ = CasperJS, UA = Unknown automation, M = manual, LHH = Live HTTP Headers, Asterisk = inferred
Many Studies, Many Platforms

● Automation:
  ○ 7 used Selenium (Full browser)
  ○ 4 used PhantomJS/CapsperJS (Headless webkit)

● Instrumentation
  ○ 5 used FourthParty
  ○ 9 used a Proxy
Many Studies, Many Platforms

- **Automation:**
  - 7 used Selenium (Full browser)
  - 4 used PhantomJS/CapsperJS (Headless webkit)

- **Instrumentation**
  - 5 used FourthParty
  - 9 used a Proxy

FourthParty is the only shared code
Web measurement hurdles

1. Engineering Debt

2. Lasting Impact
Canvas Fingerprinting in May 2014

- Acar, et.al (2014)
- 5% of Top 100k
Canvas Fingerprinting in May 2014

- Acar, et.al (2014)
- 5% of Top 100k
Canvas Fingerprinting in October 2015

Over 100 first-party domains on the Top 100k
Canvas Fingerprinting in October 2015

Over 100 first-party domains on the Top 100k

Overcoming these hurdles:

1. A Common Platform

2. A Web Privacy Census
OpenWPM
OpenWPM
OpenWPM
OpenWPM

Browser Instance

 Firefox

 mitmproxy

 Web

 FourthParty

 Se
OpenWPM

- Supports browsing with persistent state
  - Browser keeps profile through crashes and freezes.

- Real Browser
  - Extensions
  - Privacy Features
  - WebRTC, Audio, Video, WebGL

- Stable
A Web Privacy Census

Monthly
1 Million Site Crawl
A Web Privacy Census

Monthly
1 Million Site Crawl

Collecting:
- Javascript Calls
- All javascript files
- HTTP Requests and Responses
- Storage (cookies, Flash, etc)
<table>
<thead>
<tr>
<th>Type</th>
<th>Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stateful</td>
<td>● ID Cookies</td>
</tr>
<tr>
<td></td>
<td>● Respawning</td>
</tr>
<tr>
<td></td>
<td>● Cookie syncing</td>
</tr>
<tr>
<td>Stateless</td>
<td>● Ghostery</td>
</tr>
<tr>
<td></td>
<td>● AdBlock Plus</td>
</tr>
<tr>
<td></td>
<td>● HTTPS Everywhere</td>
</tr>
</tbody>
</table>
A Web Privacy Census

1. Measure how effective tools are
2. Quickly deploy new measurements
3. Release data and analysis monthly
Detecting WebRTC Local IP Sniffing
1. I saw a tweet that nytimes.com is IP sniffing

Mike O'Neill
@incloud

WebRTC being used now by embedded 3rd party on nytimes.com to report visitors' local IP addresses.
2. I added code to JS Instrumentation for next crawl

// Access to webRTC
instrumentPrototype(window.mozRTCPeerConnection.prototype,
"mozRTCPeerConnection");
3. I wrote some analysis code

- Grab all urls that execute
  - mozRTCPeerConnection.onicecandidate
  - mozRTCPeerConnection.createDataChannel
  - mozRTCPeerConnection.createOffer

- Check JS Files to confirm
4. Results (October 2015)

- 121 first-party sites
  - 29 in the top 10k
- 24 unique scripts
- Only 1 of which is blocked by EasyList/EasyPrivacy
With regular measurement we can:

1. Inform the public
2. Build block lists
3. Change the incentives
2020

If you’re going to track me, ___________________
2020

If you’re going to track me, ___ I’ll know! _____________
Help us make the web more private!

- **Contribute?**
  - github.com/citp/OpenWPM

- **Collaborate?**
  - webtap.princeton.edu