Web Tracking and Fingerprinting

Vitaly Shmatikov
It’s the Internet! Of course they know you’re a dog. They also know your favorite brand of pet food and the name of the cute poodle at the park that you have a crush on!
Tracking via Cookies

◆ **Cookie**: value set by Web server, automatically sent by the browser on subsequent requests to same(ish) origin

◆ Link two sessions at same site

◆ Link sessions between different sites (third-party cookies)

◆ Can be combined with user-identifying information
Third-Party Tracking

**Third-party cookies:**
- Disabled by default (Safari)
- Can be disabled by user (many browsers)
- Cannot be disabled (Android)

... but there are many other tracking technologies
Behavioral Targeting

- Ad network
- Advertisers
- Publishers
- The New York Times
- okcupid
- Google
## Partial List of Ad Networks

<table>
<thead>
<tr>
<th>24/7 Real Media</th>
<th>33Across</th>
<th>Acerno</th>
<th>Acxiom Relevance-X</th>
<th>AdAdvisor</th>
<th>AdBrite</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adify</td>
<td>AdInterax (Yahoo!)</td>
<td>AdJuggler</td>
<td>AdShuffle</td>
<td>ADTECH (AOL)</td>
<td>Advertising.com (AOL)</td>
</tr>
<tr>
<td>Aggregate Knowledge</td>
<td>Akamai</td>
<td>AlmondNet</td>
<td>Atlas (Microsoft)</td>
<td>AudienceScience</td>
<td>Bizo</td>
</tr>
<tr>
<td>Blue Kai</td>
<td>BlueLithium (Yahoo!)</td>
<td>Bluestreak</td>
<td>BrightRoll</td>
<td>BTBuckets</td>
<td>Burst Media</td>
</tr>
<tr>
<td>Casale Media</td>
<td>Chitika</td>
<td>ChoiceStream</td>
<td>ClickTale</td>
<td>Collective Media</td>
<td>comScore VoiceFive</td>
</tr>
<tr>
<td>Coremetrics</td>
<td>Cossette</td>
<td>Criteo</td>
<td>Effective Measure</td>
<td>Eloqua</td>
<td>Eyeblasser</td>
</tr>
<tr>
<td>eXelate</td>
<td>EyeWonder</td>
<td>e-planning</td>
<td>Facilitate Digital</td>
<td>FetchBack</td>
<td>FlashTalking</td>
</tr>
<tr>
<td>Fox Audience Network</td>
<td>FreeWheel</td>
<td>Google</td>
<td>Hurra</td>
<td>interCLICK</td>
<td>Lotame</td>
</tr>
<tr>
<td>Navegg</td>
<td>NextAction</td>
<td>NexTag</td>
<td>Mediaplex (ValueClick Media)</td>
<td>Media 6 Degrees</td>
<td>Media Math</td>
</tr>
<tr>
<td>Microsoft</td>
<td>MindSet Media</td>
<td>Nielsen Online</td>
<td>nugg.ad</td>
<td>Omniture</td>
<td>OpenX</td>
</tr>
<tr>
<td>Outbrain</td>
<td>PointRoll</td>
<td>PrecisionClick</td>
<td>Pulse 360</td>
<td>Quantcast</td>
<td>Quilgo (AOL)</td>
</tr>
<tr>
<td>RichRelevance</td>
<td>Right Media (Yahoo!)</td>
<td>Rocket Fuel</td>
<td>Safecount *</td>
<td>ScanScout</td>
<td>Smart Adserver</td>
</tr>
<tr>
<td>Snoobi</td>
<td>Specific Media</td>
<td>TACODA (AOL)</td>
<td>Tattoo Media</td>
<td>Tealium</td>
<td>TradeDoubler</td>
</tr>
<tr>
<td>Traffic Marketplace</td>
<td>Tribal Fusion / Exponential</td>
<td>TruEffect</td>
<td>Tumri</td>
<td>Turn</td>
<td>Undertone Networks / Zedo</td>
</tr>
<tr>
<td>ValueClick Media</td>
<td>Vizu</td>
<td>Weborama</td>
<td>WebTrends</td>
<td>Yahoo!</td>
<td>[x+1]</td>
</tr>
</tbody>
</table>
Tracking Is Pervasive

64

independent tracking mechanisms in an average top-50 website
Sticky Tracking

Subverting same origin policy
(publisher also runs an ad network)
ad.hi5.com = ad.yieldmanager.com

Flash cookies

Browser fingerprinting

History sniffing
Cookie Syncing

Site A informing site B about user’s identity (via user’s browser)

Allows aggregation across multiple trackers
Tracking Technologies

- HTTP Cookies
- HTTP Auth
- HTTP Etags
- Content cache
- IE userData
- HTML5 protocol and content handlers
- HTML5 storage
- Flash cookies
- Silverlight storage
- TLS session ID & resume
- Browsing history
- window.name
- HTTP STS
- DNS cache
Everything Has a Fingerprint
Fingerprinting Web Browsers

- User agent
- HTTP ACCEPT headers
- Browser plug-ins
- MIME support
- Clock skew

- Installed fonts
- Cookies enabled?
- Browser add-ons
- Screen resolution
Your browser fingerprint **appears to be unique** among the 3,435,834 tested so far.

Only **anonymous data** will be collected by this site.

A paper reporting the statistical results of this experiment is now available: *How Unique Is Your Browser?*, Proceedings of the Privacy Enhancing Technologies Symposium (PETS 2010), Springer Lecture Notes in Computer Science.
84% of browser fingerprints are unique

With Flash or Java, 94% are unique
Programmatic drawing in the browser
- Draw shapes, add text, 3D (via WebGL)

Access to drawn pixels
- Array of RGBA values
- PNG-encoded data URL
Text Rendering ...

<script type="text/javascript">
  var canvas =
      document.getElementById("drawing");
  var context = canvas.getContext("2d");
  context.font = "18pt Arial";
  context.textBaseline = "top";
  context.fillText("Some letters", 2, 2);
</script>
... Text Inspection

<script type="text/javascript">
    var canvas =
        document.getElementById("drawing");
    var context = canvas.getContext("2d");
    context.font = "18pt Arial";
    context.textBaseline = "top";
    context.fillText("Some letters", 2, 2);

    var pixels =
        canvas.toDataURL("image/png");
</script>
WebFonts

- **Problem:** Clients ship with ugly fonts
- **Solution:** Browsers should download fonts from the Internet on demand!

```css
@font-face {
  font-family: 'Sirin Stencil';
  font-style: normal;
  font-weight: 400;
  src:
  url(http://themes.googleusercontent.com/static/fonts/sirinstencil/v1/[...].woff)
  format('woff');}
```
45 Ways To Sirin Stencil

```jsx
context.font = "12pt 'Sirin Stencil'";
```

<table>
<thead>
<tr>
<th>Windows</th>
<th>OS X</th>
<th>Linux</th>
</tr>
</thead>
<tbody>
<tr>
<td>How quickly daft jumping zebras vex. (Also, punctuation: Æ/c.)</td>
<td>How quickly daft jumping zebras vex. (Also, punctuation: Æ/c.)</td>
<td>How quickly daft jumping zebras vex. (Also, punctuation: Æ/c.)</td>
</tr>
<tr>
<td>How quickly daft jumping zebras vex. (Also, punctuation: Æ/c.)</td>
<td>How quickly daft jumping zebras vex. (Also, punctuation: Æ/c.)</td>
<td>How quickly daft jumping zebras vex. (Also, punctuation: Æ/c.)</td>
</tr>
<tr>
<td>How quickly daft jumping zebras vex. (Also, punctuation: Æ/c.)</td>
<td>How quickly daft jumping zebras vex. (Also, punctuation: Æ/c.)</td>
<td>How quickly daft jumping zebras vex. (Also, punctuation: Æ/c.)</td>
</tr>
<tr>
<td>How quickly daft jumping zebras vex. (Also, punctuation: Æ/c.)</td>
<td>How quickly daft jumping zebras vex. (Also, punctuation: Æ/c.)</td>
<td>How quickly daft jumping zebras vex. (Also, punctuation: Æ/c.)</td>
</tr>
<tr>
<td>How quickly daft jumping zebras vex. (Also, punctuation: Æ/c.)</td>
<td>How quickly daft jumping zebras vex. (Also, punctuation: Æ/c.)</td>
<td>How quickly daft jumping zebras vex. (Also, punctuation: Æ/c.)</td>
</tr>
<tr>
<td>How quickly daft jumping zebras vex. (Also, punctuation: Æ/c.)</td>
<td>How quickly daft jumping zebras vex. (Also, punctuation: Æ/c.)</td>
<td>How quickly daft jumping zebras vex. (Also, punctuation: Æ/c.)</td>
</tr>
</tbody>
</table>
Canvas Fingerprinting

Reveals:
- Operating system family
- Browser family
  (except Chrome, Safari on OS X)
- Installed fonts
- Font smoothing parameters

[Mowery and Shacham. “Pixel Perfect”. W2SP 2012]
How Pervasive?

◆ Present in 5.5% of top 100,000 websites
◆ Fingerprinting code comes from 20 different domains
  • addthis.com by far the most popular (95%)

“Don’t Worry, It’s All Anonymous”

◆ Is it?
◆ What’s the difference between
  “anonymous”
  “pseudonymous”
  “identified”
◆ Which technology changed data collection from anonymous to pseudonymous?
How Websites Get Your Identity

Third party is sometimes the site itself

Leakage of identifiers

GET http://ad.doubleclick.net/adj/...
Referer: http://submit.SPORTS.com/...?email=jdoe@email.com
Cookie: id=35c192bcfe0000b1...

Security bugs

XSUH: cross-site URL hijacking

Third party buys your identity
Introduction

Syphilis is a bacterial infection that is usually passed on through having sex with someone who is infected. It can also be passed from an infected mother to her unborn child and, in rare cases, can be caught through injecting drugs. It is extremely rare to catch syphilis through a blood transfusion in the UK as blood donors are carefully screened.

Three stages of disease

Stage 1 (primary syphilis). Symptoms of syphilis begin with a painless but highly infectious sore or sore on the genitals or sometimes around the mouth. It sometime also becomes close contact with the sore, typically during sexual contact, they can also become infected. The sore lasts two to six weeks before disappearing.

Stage 2 (secondary syphilis). Secondary symptoms, such as a skin rash and sore throat, then develop. These symptoms may disappear within a few weeks, after which you experience a latent (hidden) phase with no symptoms, which can last for years. After this, syphilis can progress to its third, most dangerous stage.

Stage 3 (tertiary syphilis). At this stage, it can cause serious damage to the body. The primary and secondary stages are when you are most infectious to other people. In the latent phase (and usually around two years after becoming infected), syphilis cannot be passed onto others but can still cause symptoms. See Symptoms of syphilis for more information on the
History Sniffing

How can a webpage figure out which sites you visited previously?

- Color of links
  - CSS :visited property
  - getComputedStyle()

- Cached Web content timing

- DNS timing
Identity Sniffing

[Wondracek et al. Oakland 2010]

- All social networking sites allow users to join groups
- Users typically join multiple groups
  - Some of these groups are public
- Group-specific URLs are predictable
  - https://www.xing.com/net/[groupID]/forums+
- Intersection of group affiliations acts as a fingerprint
  - Can sometimes infer identity by computing the intersection of group membership lists
Do Not Track

Basics

HTTP header
  - DNT: 1

Standardization

Browser support in FF4, IE9

Beginning to see adoption (AP, NAI)... or not

Privacy protections

No tracking across sites
  – Who is the “third” party?
    Can’t be based on domain
    Example: amazonaws.com, ad.hi5.com ...

No intrusive tracking

Limits on regular log data

Exceptions for fraud prevention, etc.
DNT Adoption Issues

“But the NAI code also recognizes that companies sometimes need to continue to collect data for operational reasons that are separate from ad targeting based on a user’s online behavior. For example, online advertising companies may need to gather data to prove to advertisers that an ad has been delivered and should be paid for; to limit the number of times a user sees the same ad; or to prevent fraud.”

Translation: we’re going to keep tracking you, but we’ll simply call it “operational reasons.”
Today, we are announcing the launch of Atlas. We’ve rebuilt Atlas from the ground up to tackle today’s marketing challenges, like reaching people across devices and bridging the gap between online impressions and offline purchases.

iOS 6 introduces the Advertising Identifier, a non-permanent, non-personal, device identifier, that advertising networks will use to give you more control over advertisers’ ability to use tracking methods. If you

How are these identifiers different from third-party cookies?

Google AdID