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

Publishers

Advertisers

The New York Times

okcupid

Google

slide 5
<table>
<thead>
<tr>
<th>Ad Network</th>
<th>Ad Network</th>
</tr>
</thead>
<tbody>
<tr>
<td>24/7 Real Media</td>
<td>33Across</td>
</tr>
<tr>
<td>Adify</td>
<td>Acerno</td>
</tr>
<tr>
<td>Aggregate Knowledge</td>
<td>Acxiom Relevance-X</td>
</tr>
<tr>
<td>Blue Kai</td>
<td>AdAdvisor</td>
</tr>
<tr>
<td>Blue Lithium (Yahoo!)</td>
<td>AdBrite</td>
</tr>
<tr>
<td>Casale Media</td>
<td>AdInterax (Yahoo!)</td>
</tr>
<tr>
<td>Chitika</td>
<td>AdJuggler</td>
</tr>
<tr>
<td>Criteo</td>
<td>AdShuffle</td>
</tr>
<tr>
<td>Cossette</td>
<td>AdTECH (AOL)</td>
</tr>
<tr>
<td>ChoiceStream</td>
<td>AudienceScience</td>
</tr>
<tr>
<td>Coremetrics</td>
<td>Bizo</td>
</tr>
<tr>
<td>Criteo</td>
<td>BTBuckets</td>
</tr>
<tr>
<td>e-planning</td>
<td>Burst Media</td>
</tr>
<tr>
<td>eXelate</td>
<td>ClickTale</td>
</tr>
<tr>
<td>EyeWonder</td>
<td>Collective Media</td>
</tr>
<tr>
<td>FreeWheel</td>
<td>comScore VoiceFive</td>
</tr>
<tr>
<td>Google</td>
<td>Eloqua</td>
</tr>
<tr>
<td>EyeWonder</td>
<td>Eyeblaster</td>
</tr>
<tr>
<td>Facilitate Digital</td>
<td>FetchBack</td>
</tr>
<tr>
<td>Facilitate Digital</td>
<td>Flashtalking</td>
</tr>
<tr>
<td>Mediaplex (ValueClick Media)</td>
<td>interCLICK</td>
</tr>
<tr>
<td>Media 6 Degrees</td>
<td>Lotame</td>
</tr>
<tr>
<td>Media Math</td>
<td>MindSet Media</td>
</tr>
<tr>
<td>Microsoft</td>
<td>NexTag</td>
</tr>
<tr>
<td>NextAction</td>
<td>Media 6 Degrees</td>
</tr>
<tr>
<td>Nielsen Online</td>
<td>Media Math</td>
</tr>
<tr>
<td>nexTag</td>
<td>Media 6 Degrees</td>
</tr>
<tr>
<td>Media Math</td>
<td>Microsoft</td>
</tr>
<tr>
<td>Media 6 Degrees</td>
<td>Microsoft</td>
</tr>
<tr>
<td>Microsoft</td>
<td>Media Math</td>
</tr>
<tr>
<td>Media 6 Degrees</td>
<td>Microsoft</td>
</tr>
<tr>
<td>Media Math</td>
<td>Microsoft</td>
</tr>
<tr>
<td>Microsoft</td>
<td>Microsoft</td>
</tr>
<tr>
<td>Media 6 Degrees</td>
<td>MSN Media</td>
</tr>
<tr>
<td>MSN Media</td>
<td>Nielsen Online</td>
</tr>
<tr>
<td>Omniture</td>
<td>nugg.ad</td>
</tr>
<tr>
<td>Omnitree</td>
<td>Quantcast</td>
</tr>
<tr>
<td>OpenX</td>
<td>Quantcast</td>
</tr>
<tr>
<td>Outbrain</td>
<td>PrecisionClick</td>
</tr>
<tr>
<td>PointRoll</td>
<td>Pulse 360</td>
</tr>
<tr>
<td>Priceline</td>
<td>Quantcast</td>
</tr>
<tr>
<td>Priceline</td>
<td>Quantcast</td>
</tr>
<tr>
<td>Right Media (Yahoo!)</td>
<td>PrecisionClick</td>
</tr>
<tr>
<td>Rocket Fuel</td>
<td>Pulse 360</td>
</tr>
<tr>
<td>Richrelevance</td>
<td>PrecisionClick</td>
</tr>
<tr>
<td>Right Media (Yahoo!)</td>
<td>Pulse 360</td>
</tr>
<tr>
<td>Safecount *</td>
<td>PrecisionClick</td>
</tr>
<tr>
<td>Snoopi</td>
<td>Pulse 360</td>
</tr>
<tr>
<td>Specific Media</td>
<td>Pulse 360</td>
</tr>
<tr>
<td>Spectrum</td>
<td>Pulse 360</td>
</tr>
<tr>
<td>Traffic Marketplace</td>
<td>Pulse 360</td>
</tr>
<tr>
<td>Tribal Fusion / Exponential</td>
<td>Pulse 360</td>
</tr>
<tr>
<td>TrueEffect</td>
<td>Pulse 360</td>
</tr>
<tr>
<td>Turn</td>
<td>Pulse 360</td>
</tr>
<tr>
<td>Undertone Networks / Zedo</td>
<td>Pulse 360</td>
</tr>
<tr>
<td>ValueClick Media</td>
<td>Pulse 360</td>
</tr>
<tr>
<td>Vizu</td>
<td>Pulse 360</td>
</tr>
<tr>
<td>Weborama</td>
<td>Pulse 360</td>
</tr>
<tr>
<td>WebTrends</td>
<td>Pulse 360</td>
</tr>
<tr>
<td>Yahoo</td>
<td>Pulse 360</td>
</tr>
</tbody>
</table>

(x+1)
2012 DISPLAY ADVERTISING ECOSYSTEM EUROPE

PUBLISHERS
- Data Suppliers
- Data Management Platforms
- Data Exchanges
- Sales Houses
- SSP & Private Ad Exchanges
- Delivery Systems, Tools & Analytics
- Verification & Privacy

ADVERTISERS
- Ad Networks
- Demand Side Platforms
- Agencies
- Agency Trading Desks
- Delivery Systems, Tools & Analytics
- Verification & Privacy
- Published by IMPROVE DIGITAL

Ad Exchanges
- Audience Targeting / Re-Targeting
- Trading Desks
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.

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.
Panopticlick Example

84% of browser fingerprints are unique

With Flash or Java, 94% are unique
<CANVAS>

- 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

```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!

@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

code
context.font = "12pt 'Sirin Stencil'";

## Windows

- How quickly daft jumping zebras vex. (Also, punctuation: &/c.)
- How quickly daft jumping zebras vex. (Also, punctuation: &/c.)
- How quickly daft jumping zebras vex. (Also, punctuation: &/c.)
- How quickly daft jumping zebras vex. (Also, punctuation: &/c.)
- How quickly daft jumping zebras vex. (Also, punctuation: &/c.)

## OS X

- How quickly daft jumping zebras vex. (Also, punctuation: &/c.)
- How quickly daft jumping zebras vex. (Also, punctuation: &/c.)
- How quickly daft jumping zebras vex. (Also, punctuation: &/c.)
- How quickly daft jumping zebras vex. (Also, punctuation: &/c.)
- How quickly daft jumping zebras vex. (Also, punctuation: &/c.)

## Linux

- How quickly daft jumping zebras vex. (Also, punctuation: &/c.)
- How quickly daft jumping zebras vex. (Also, punctuation: &/c.)
- How quickly daft jumping zebras vex. (Also, punctuation: &/c.)
- How quickly daft jumping zebras vex. (Also, punctuation: &/c.)
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%)

Draws Cwm fjordbank glyphs vext quiz into the canvas

Why this text?

“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 on the genitalia or sometimes around the mouth. If somebody else comes into 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
  
  
  http://www.facebook.com/group.php?gid=[groupID]&v=info&ref=nf+
  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
One-Click Fraud

Thank you for your patronage! You successfully registered for our premium online services, at an incredible price of 50,000 JPY. Please promptly send your payment by bank transfer to ABC Ltd at Ginko Bank, Account 1234567. Questions? Please contact us at 080-1234-1234.

Your IP address is 10.1.2.3, you run Firefox 3.5 over Windows XP, and you are connecting from Tokyo.

Failure to send your payment promptly will force us to mail you a postcard reminder to your home address. Customers refusing to pay will be prosecuted to the fullest extent of the law. Once again, thank you for your patronage!
One-Click Fraud

◆ Estimated costs to victims:
  USD 260 million / year

◆ What’s going on here?
◆ Why only Japan?
  • Cultural factors:
    susceptibility to authoritative language
    threat of public shaming

Credible because the website does have your real identity!
Instant Personalization

Creepy is the new normal
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.”
Brave New World?

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.

How are these identifiers different from third-party cookies?

Ad Tracking

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 choose not to receive advertising personalization, you can disable the Advertising Identifier. However, when advertising networks transition to using the Advertising Identifier you may lose targeted ads from other services.