Online Tracking, What Can Be Done About it, and Who’s Doing it

Pete Snyder
Senior Privacy Researcher, Director of Privacy
pes@brave.com
Hi, I’m Pete 👋

- Grew up in Chicago
  …actual Chicago

- Law school -> freelance web stuff
  Started: Anchorage, AK
  Ended: Judge Judy Show invitation

- University of Illinois at Chicago
Me at Brave

- Research at Brave
  ...privacy, blocking, reliability

- Co-Chair of PING
  Privacy committee on W3C

- Research <-> Engineering
  Web compat, filter lists, etc.

- Academic <-> Industry
  Collaborations
Brave in a Slide

- Privacy focused
- Alternative web funding model
  Fix incentive problems
- Research + Engineering
- Not just a browser
  - search.brave.com
  - talk.brave.com
  - VPN
  - more coming…
Overview

- **Why Privacy Matters**
  A sloppy manifesto

- **Defining Tracking**
  Abstracting the problem

- **Tracking in Practice**
  Methods and defenses

- **Privacy Beyond Tracking**
  Other issues and concerns
Overview

- **Why Privacy Matters**
  A sloppy manifesto

- **Defining Tracking**
  Abstracting the problem

- **Tracking in Practice**
  Methods and defenses

- **Privacy Beyond Tracking**
  Other issues and concerns
Why Does Tracking Exist?
100% FREE MAGAZINE SUBSCRIPTIONS!
Welcome the The "First" Banner Ad

Yes, this site is supposed to look this way. After all, this is what most web pages looked like back on October 27, 1994 -- the day that Wired Magazine flipped the switch on its first website, hotwired.com, starting a revolution in web content and advertising that still reverberates today.

This site is dedicated to showing off one of the ads that ran on that site. No, it wasn't the "first" as there were a handful of other ads that ran on various sections of hotwired.com. This site is also here to tell the story of how that ad came to be, how it succeeded beyond anything we had imagined, and how we tried to set an example for how corporations could communicate with their audiences.

This site launched on October 27, 2014. It is being constantly updated, so please check back again soon for more. In the meantime, get started by clicking your mouse in the banner ad above explore these other options.
Disney Lover's Web Ring
CPS STRIKE IS OFFICIALLY ON AS CHICAGO TEACHERS UNION SAYS THERE IS NO LAST-MINUTE DEAL

CPS STRIKE IS ON

CPS strike is officially on as teachers union, Chicago teachers call for work stoppage

SPECIAL SALE
ONLY $2 FOR 20 WEEKS
Get stories that impact you

MORE CPS STRIKE COVERAGE

CPS strike live updates: Chicago teachers reject city offer, will walk off job Thursday

Chicago Park District workers reach contract
Welcome to the World's Worst Website!

This website was designed to graphically demonstrate the most common mistakes made by new web page designers.

Where am I and where are the links to other pages?

An easy to use navigation structure is essential to any well-designed website!
Important information should never be more than 2 clicks away.

As you can see, this text is difficult to read. There needs to more contrast between the background color and the text color. This makes it hard to read.

Keep your backgrounds simple. White or light colors usually work best.
Your background should not compete with the content of the page for the user's attention. If you would like to use a background picture, select a picture that uses muted colors or format your picture as a watermark.
Select text colors which will contrast well with the background picture.

Constantly running animations can be distracting when used excessively.
Identify “expensive” people here

Pay a little to advertise to them here
Summarizing: Why Does Tracking Matter

- Incompatible with dignity
- Power and control
- Transfers wealth from value-creators to attention-attractors
Overview

● Why Privacy Matters
  A sloppy manifesto

● Defining Tracking
  Abstracting the problem

● Tracking in Practice
  Methods and defenses

● Privacy Beyond Tracking
  Other issues and concerns
Definitions

● **Website:** eTLD+1 (determined by public suffix list)
  e.g., brave.com != mozilla.org
  e.g., talk.brave.com == search.brave.com
  e.g., ted.github.io != betty.github.io

● **Origin:** The full DNS host name serving a site

● **First-party:** Site of the top level document

● **Third-party:** any other site
Definitions (more)

- **DOM Storage**: Explicit storage APIs
  e.g., cookies, localStorage, IndexDB

- **Network State**: All other storage
  e.g., caches (v8, DNS, HTTP)
  e.g., Header instructions (HSTS, ALT-SRV, etc)

- **Online Tracking**
  Its trickier…
A Rough Definition of Tracking

- **Linking activities...**
  e.g., being “followed”

- **across boundaries...**
  e.g., temporal, geographic, conceptual

- **In a way not expected or desired.**
  e.g., ignorance or non-consent
Tracking in Context

Bad Internet Things
- Tracking
- Misinformation
- Harassment
- Cyber Crime
A Rough Definition of Tracking

- Linking activities...  
  e.g., being “followed”

- across boundaries...  
  e.g., temporal, geographic, conceptual

- In a way not expected or desired.  
  e.g., ignorance or non-consent
One day…

some-site.example

other-site.example
Question One

some-site.example

One day...

other-site.example

Not linking
Question Two

One day...

some-site.example

some-site.example
Question Three

One day...

some-site.example

some-site.example
One day…

Question Three

some-site.example

Linking (first-party)

some-site.example
Question Four

One day…
Question Four

One day...

Not linking
One day…
Question Five

1. some-site.example

2. other-site.example

One day...

tracker.example

Linking (third-party)
Tracking: Linking...

- **Tying behaviors to same identity**
  Could be pseudonymous, or a “real world” identity

- **Probabilistic or deterministic**
  For some definition of “probable enough”
A Rough Definition of Tracking

- Linking activities...
  e.g., being “followed”

- across boundaries...
  e.g., temporal, geographic, conceptual

- In a way not expected or desired.
  e.g., ignorance or non-consent
Tracking: ...across boundaries...

- **Organizational boundaries**
  e.g., eTLD+1, origin, “first-party set”

- **Temporal boundaries**
  e.g., tying something done last year to something done today

- **Profile boundaries**
  e.g., private browsing, different browsers, accounts
A Rough Definition of Tracking

- Linking activities...
  e.g., being “followed”

- across boundaries...
  e.g., temporal, geographic, conceptual

- In a way not expected or desired.
  e.g., ignorance or non-consent
Tracking: …expectations

- **Expectations differ across platforms**
  Facebook inapp browser vs Tor Browser Bundle

- **Expectations differ across people’s expertise**
  e.g., my dad vs Dworkin

- **Expectations differ across backgrounds**
  e.g., outlook.com vs microsoft.com vs github.com

- **Consent is (sometimes) fuzzy**
  Terms of service ← — — — — — — — — — → Storage Access API
A Rough Definition of Tracking

- Linking activities...
  e.g., being “followed”

- across boundaries...
  e.g., temporal, geographic, conceptual

- In a way not expected or desired.
  e.g., ignorance or non-consent
Overview

- **Why Privacy Matters**
  A sloppy manifesto

- **Defining Tracking**
  Abstracting the problem

- **Tracking in Practice**
  Methods and defenses

- **Privacy Beyond Tracking**
  Other issues and concerns
Tracking Techniques

- Third-party DOM storage
- Network state
- Bounce tracking
- Browser fingerprinting
- IP address
- Personal identifiers
Tracking Techniques

- Third-party DOM storage
- Network state
- Bounce tracking
- Browser fingerprinting
- IP address
- Personal identifiers
Third-party DOM storage

/some-site.example

/tracker.example

cookie=null

cookie=abc
Third-party DOM storage: cookies
Third-party DOM storage: cookies

some-site.example → /pixel.gif cookie=abc → tracker.example

/​​pixel.gif cookie=abc → tracker.example → /pixel.gif cookie=abc → other-site.example
Third-party DOM storage: iframe

```javascript
<iframe src='//tracker.example'>
const LS = localStorage
if (LS['id']) {
  // I re-identified a person
} else {
  // new person, assigning ID
  LS['id'] = Math.random()
}
fetch(`/record?id=${LS['id']}`)
</iframe>
```
Third-party DOM storage: iframe

```javascript
const LS = localStorage;

if (LS['id']) {
  // I re-identified a person
} else {
  // new person, assigning ID
  LS['id'] = Math.random();
}

fetch(`/record?id=${LS['id']}`)
```

```html
<iframe src='//tracker.example'></iframe>
```
Third-party DOM storage: iframe

```javascript
<iframe src='//tracker.example'>
const LS = localStorage
if (LS['id']) {
  // I re-identified a person
} else {
  // new person, assigning ID
  LS['id'] = Math.random()
}
fetch(`/record?id=${LS['id']}`)
</iframe>
```
Third-party DOM storage: partitioning

- Third-party storage is not shared across sites
- Sometimes called “dual-keying”
- Previous:
  \[
  \text{storage\_data} = \text{browser\_storage}[\text{<requested eTLD+1>}]\]
- Partitioning:
  \[
  \text{storage\_data} = \text{browser\_storage}[\text{<first-party eTLD+1>}]\text{[<requested eTLD+1>]}\]
Third-party DOM storage: partitioning

some-site.example

/pixel.gif
cookie=null

cookie=abc

tracker.example
Third-party DOM storage: partitioning

some-site.example

_tracker.example_

/tracker.example/

cookie=null

cookie=abc

_cookie=abc_

_cookie=123_

/cookie=null_

/cookie=123_

other-site.example
## Third-party DOM storage: Defenses

<table>
<thead>
<tr>
<th></th>
<th>Chrome</th>
<th>Safari</th>
<th>Edge</th>
<th>Firefox</th>
<th>Brave</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Block third-party cookies</strong></td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td><strong>Partition storage</strong></td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td><strong>Ephemeral partitions</strong></td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td><strong>List based defenses</strong></td>
<td></td>
<td></td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
</tbody>
</table>
Tracking Techniques

- Third-party DOM storage
- Network state
- Bounce tracking
- Browser fingerprinting
- IP address
- Personal identifiers
Network State Example: HTTP Cache

- Browsers cache things for speed
  Images, JavaScript, etc.

- Caches are generally unpartitioned

- Anything unpartitioned can be a linking key
HTTP Cache Tracking

- some-site.example

- tracker.example

- /pixel/#.gif?action={set, read}
  - action=set: 50% return pixel
    - 50% 404
  - action=read: 100% 404
HTTP Cache Tracking

some-site.example → tracker.example

/pixel/1.gif?action=set → 50% a pixel, 50% 404
/pixel/2.gif?action=set
/pixel/3.gif?action=set
/pixel/4.gif?action=set
/pixel/5.gif?action=set
...
/pixel/32.gif?action=set → 50% a pixel, 50% 404
HTTP Cache Tracking

```javascript
const identifier = []
for (let i = 0; i < 32; i += 1) {
  try {
    const url = `/pixel/${i}.gif?action=read`
    await fetch(url)
    // We hit the cache
    identifier[i] = 1
  } catch (_) {
    // We missed the cache
    identifier[i] = 0
  }
}
// identifier is now a unique 32 bits
```
HTTP Strict Transport Security (HSTS)

- **Website Says “only HTTPS, forever”**
  - e.g. persistent storage

- **Automatic Upgrade**
  - http://example.org -> https://example.org

- **How to leverage?**
HSTS Tracking

- example.org
- a.example.org
- b.example.org
- a.a.example.org
- b.a.example.org
## Network state: Defenses

<table>
<thead>
<tr>
<th></th>
<th>Chrome</th>
<th>Safari</th>
<th>Edge</th>
<th>Firefox</th>
<th>Brave</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Partition network state</strong></td>
<td><img src="image" alt="Partition Network State" /></td>
<td><img src="image" alt="Not Available" /></td>
<td><img src="image" alt="Available" /></td>
<td><img src="image" alt="Available" /></td>
<td><img src="image" alt="Not Available" /></td>
</tr>
<tr>
<td><strong>List based defenses</strong></td>
<td><img src="image" alt="Not Available" /></td>
<td><img src="image" alt="Not Available" /></td>
<td><img src="image" alt="Not Available" /></td>
<td><img src="image" alt="Not Available" /></td>
<td><img src="image" alt="Not Available" /></td>
</tr>
</tbody>
</table>
Tracking Techniques

- Third-party DOM storage
- Network state
- Bounce tracking
- Browser fingerprinting
- IP address
- Personal identifiers
Bounce Tracking

- Response to partitioning
- Third parties use first-parties to track
- Growing in importance as partitioning is more common
Pre-partitioning

1. Record Page View
   cookie=abc

2. Navigate to new page

3. Record Page View
   cookie=abc
Storage partitioning

1. Record Page View
   cookie=abc

2. Navigate to new page

3. Record Page View
   cookie=123
Bounce tracking

1. Attempt to navigate from `some-site.example`

2. Record Page View
   - cookie=123
   - from=some-site...
   - to=other-site...

To `other-site.example`
Navigation tracking

1. Fetch tracking script

2. Attempt to navigate
e.g., https://other-site.example

3. Tracker annotates URL with id=123
e.g., https://other-site.example?id=123

4. Fetch tracking script

5. Reads identifier from URL
# Bounce and Navigation Tracking: Defenses

<table>
<thead>
<tr>
<th></th>
<th>Chrome</th>
<th>Safari</th>
<th>Edge</th>
<th>Firefox</th>
<th>Brave</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Limit storage</strong></td>
<td></td>
<td>heuristic</td>
<td></td>
<td>List</td>
<td></td>
</tr>
<tr>
<td><strong>“Debounce”</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>List</td>
</tr>
<tr>
<td><strong>Warn user</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>List</td>
</tr>
</tbody>
</table>
Tracking Techniques

- Third-party DOM storage
- Network state
- Bounce tracking
- Browser fingerprinting
- IP address
- Personal identifiers
Fingerprinting, contrasted

- **Classic tracking**
  - Website stores an id on the client
  - The client returns the id to the server (cookie or JS)
  - The id is what allows re-identification
  - “Stateful”

- **Fingerprinting / passive tracking**
  - Website finds things different about each visitor
  - Tracker derives the identifier from minor browser differences
  - “Stateless”
Fingerprinting, how?

- Large number of semi-identifiers
  - Browser size
  - Extra fonts
  - Audio hardware
  - Video hardware
  - Installed plugins
  - Color depth

- Add the semi identification up…
All browser users: 5 billion people

You 1 person in 5 billion
All browser users:
5 billion people

Firefox Users
All browser users: 5 billion people

Windows users
All browser users: 5 billion people

Office Fonts
Sending DNT header

All browser users: 5 billion people
All browser users: 5 billion people

Using content blocker
All browser users: 5 billion people

You 1 person in 100
Fingerprinting, abstracted

- Still needs a common value across boundaries
  Sites, sessions, time, etc

- Value needs to be unique
  Otherwise it mixes you up with others

- Value needs to be consistent
  Otherwise it doesn’t (re)identify you
Possible Defenses

- **Try to make browsers look similar**
  Reduce the “bits” available to fingerprinters

- **Try to block bad parties**
  Keep the “bad folks” out

- **Privacy budgets**
  Only allow sites to do so much identifying, e.g., 10 bits but not more

- **Randomization**
  Make browser look intentionally different, within each boundary
# Fingerprinting: Defenses

<table>
<thead>
<tr>
<th></th>
<th>Chrome</th>
<th>Safari</th>
<th>Edge</th>
<th>Firefox</th>
<th>Brave</th>
</tr>
</thead>
<tbody>
<tr>
<td>Restricted hardware</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Feature selection / removal</td>
<td></td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Block fingerprinters</td>
<td></td>
<td></td>
<td>⊗</td>
<td>⊗</td>
<td>⊗</td>
</tr>
<tr>
<td>Randomization</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
</tbody>
</table>
Tracking Techniques

- Third-party DOM storage
- Network state
- Bounce tracking
- Browser fingerprinting
- IP address
- Personal identifiers
IP Addresses are pretty unique!

- ...especially if you look for clusters
  The 3 ips you most commonly connect from is very unique

- IPv6 makes it a lot worse
  Obviously… :-/

- Four general approaches
  - Contracts / promises
  - proxies
  - mix nets
  - block bad parties
## IP Addresses Defenses

<table>
<thead>
<tr>
<th></th>
<th>Chrome</th>
<th>Safari</th>
<th>Edge</th>
<th>Firefox</th>
<th>Brave</th>
</tr>
</thead>
<tbody>
<tr>
<td>Websites promise</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proxies</td>
<td></td>
<td></td>
<td></td>
<td>Optional VPN</td>
<td>Optional VPN</td>
</tr>
<tr>
<td>Mix networks</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Optional Tor</td>
</tr>
<tr>
<td>Block bad parties</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
</tbody>
</table>
Tracking Techniques

- Third-party DOM storage
- Network state
- Bounce tracking
- Browser fingerprinting
- IP address
- Personal identifiers
Personal Identifiers

- Names, email addresses, CCN, etc
  “old school”

- Can be combined with offline sources
  Credit agencies, public legal records, tax documents, etc

- Baked into the web
  Authentication, user accounts, etc
Partitioning to the Rescue (?)

- User holds the “true” value
e.g., true email address

- Browser holds a secret
e.g., secret = rand()

- Derive per site identities
e.g. hash(email + secret + eTLD+1) + @private-email.com

- Applicable to a range of identifiers
Email, CCN, Crypto addresses
# Personal Identifiers: Defenses

<table>
<thead>
<tr>
<th></th>
<th>Chrome</th>
<th>Safari</th>
<th>Edge</th>
<th>Firefox</th>
<th>Brave</th>
</tr>
</thead>
<tbody>
<tr>
<td>Partition email</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>🔍</td>
</tr>
<tr>
<td>Partition Web3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>🔍</td>
</tr>
<tr>
<td>Block scripts</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>🚫</td>
</tr>
</tbody>
</table>
Overview

- **Why Privacy Matters**
  A sloppy manifesto

- **Defining Tracking**
  Abstracting the problem

- **Tracking in Practice**
  Methods and defenses

- **Privacy Beyond Tracking**
  Other issues and concerns
Privacy is more than Absence of Tracking

- Browsers shouldn’t share information unless it’s helpful to the user
e.g., FLoC

- Browsers should serve users first and exclusively
e.g., Reporting API, FLEDGE

- Browsers shouldn’t introduce capabilities that remove user choice
WebBundles

- Browsers shouldn’t confuse users!
First-party sets, SXG

- First-parties are suspect too...
Other privacy protections

● Governments increasing provide legal protections
  GDPR, CCPA, etc

● Browsers can help users assert their privacy rights
  e.g., GlobalPrivacyControl

● Authored by activists, academics, New York Times, DuckDuckGo, Brave
  Implemented in Brave and DDG

● Beware of conflating with “consent management” systems
A final plea…

- You are all plainly, amazingly smart people
- You’ll be able to (mostly) choose your job
- Privacy harms are particularly difficult to remediate
- Consider the privacy implications of a job before you take it
A final plea...

- You are all plainly, amazingly smart people
- You’ll be able to (mostly) choose your job
- Privacy harms are particularly difficult to remediate
- Consider the privacy implications of a job before you take it

Thanks!