

# Web Tracking

An overview

# What is web tracking



**Web tracking** is the practice by which operators of websites collect, store and share information about a particular user's activity on the World Wide Web.

An **HTTP cookie** is a small piece of data sent from a website and stored on the user's computer by the user's web browser while the user is browsing

**First Party:** a domain to which a user goes intentionally, by typing a URL or clicking a link.

**Third Party:** a domain whose content is embedded in a first-party web page.

# Why should we care about web tracking?

Any weird feelings from the eye picture from the previous slide?

Views and edits of privacy-related Wikipedia articles reduced after leaks of NSA surveillance activities\*

People donate more for communal coffee when a picture of watching eyes is posted near the coffee\*\*

source : [https://www.usenix.org/sites/default/files/conference/protected-files/security16\\_slides\\_lerner.pdf](https://www.usenix.org/sites/default/files/conference/protected-files/security16_slides_lerner.pdf)

# Outline

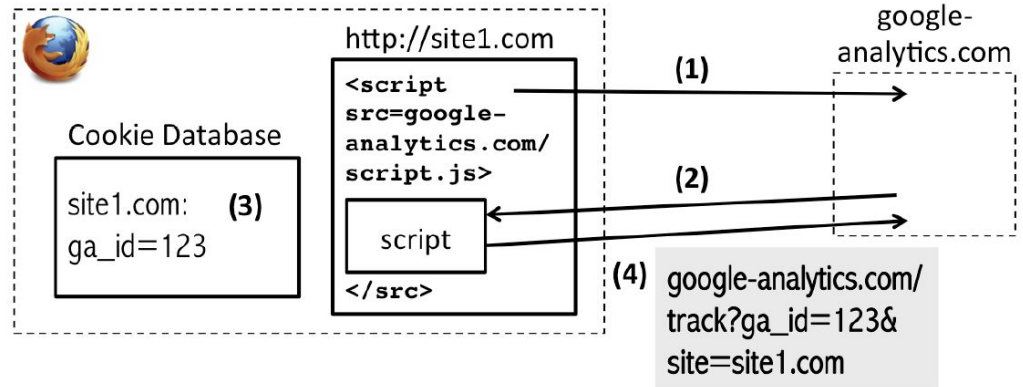
1. Two different types of web tracking
  - a. Stateful (cookie-based)
  - b. Stateless (fingerprinting)
2. Large scale view of web tracking
  - a. OpenWPM
  - b. Result and evaluation
3. Historical view of web tracking
  - a. TrackingExcavator + Wayback Machine
  - b. Result and evaluation

# 1.a Stateful Web tracking

1. Analytics Tracking
2. Vanilla Tracking
3. Forced Tracking
4. Referred Tracking
5. Personal Tracking
6. Referred Analytics Tracking

# Analytics Tracking

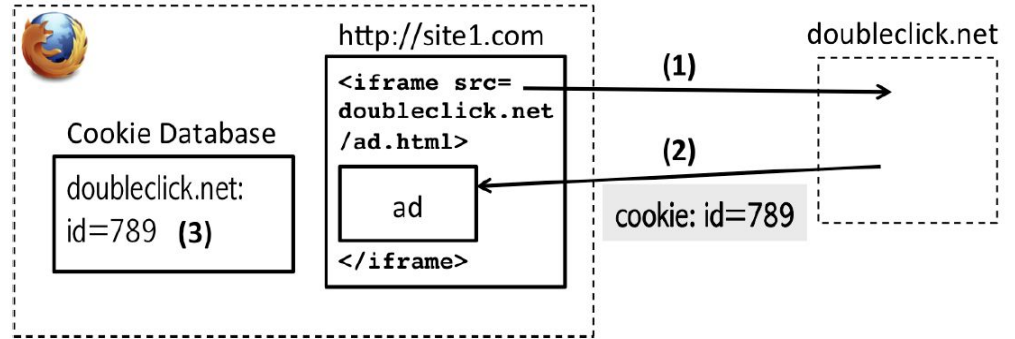
- Serves as a third-party analytics engine
- Can only track **within sites**
- Example: Google Analytics



(1) the website embedding the GA script, which, after (2) loading in the user's browser, (3) sets a site-owned cookie. This cookie is (4) communicated back to GA along with other tracking information.

# Vanilla Tracking

- Uses third-party storage to track users **across sites**.
- Example: DoubleClick



When a website (1) includes a third-party ad from an entity like Doubleclick, Doubleclick (2-3) sets a tracker-owned cookie on the user's browser. Subsequent requests to Doubleclick from any website will include that cookie, allowing it to track the user across those sites.

# Forced Tracking

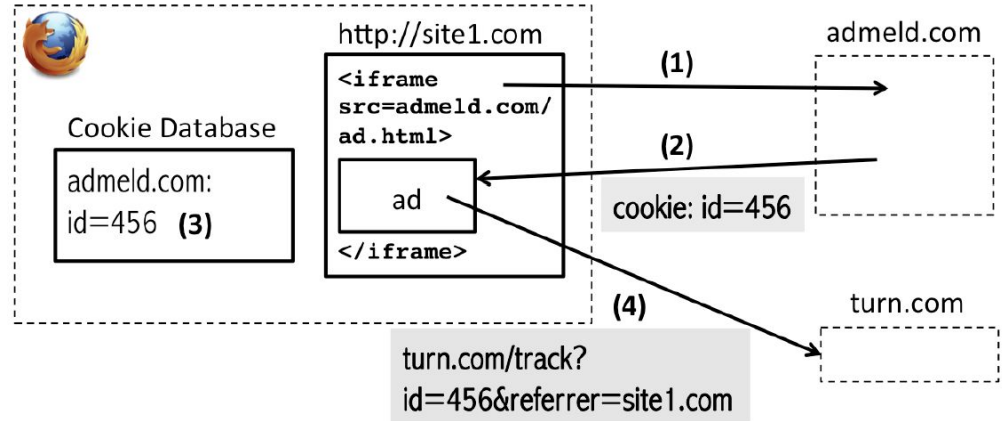
- The cross-site tracker **forces users to visit** its domain directly (e.g., popup, redirect), placing it in a **first-party position**
- Example:  
insightexpressai.com

1. The tracker **forces** the user to visit its domain **directly**, e.g., with a popup or a redirect, allowing it to set its tracker-owned cookie from a first-party position.
2. The tracker sets a tracker-owned cookie, which is then automatically included with any requests to the tracker's domain when allowed by the browser.



# Referred Tracking

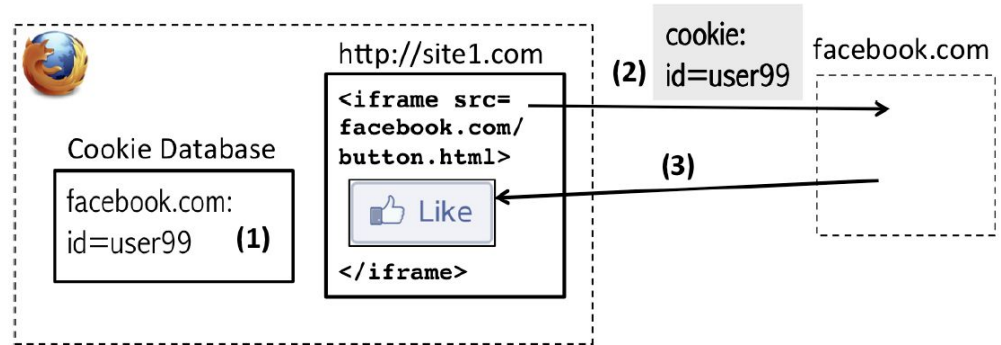
- The tracker relies on other trackers to **leak unique identifiers to it**
- rather than on its own client-side state, to track users across sites.
- Example: invitemediacom



A website (1-2) embeds an ad from Admeld, which (3) sets a tracker-owned cookie. Admeld then (4) makes a request to another third-party advertiser, Turn, and passes its own tracker-owned cookie value and other tracking information to it.

# Personal Tracking

- The cross-site tracker is visited by the user directly in other contexts.
- Example: Facebook “Like” button



Social sites like Facebook, which users visit directly in other circumstances— allowing them to (1) set a cookie identifying the user—expose social widgets such as the “Like” button. When another website embeds such a button, the request to Facebook to render the button (2-3) includes Facebook’s tracker-owned cookie.

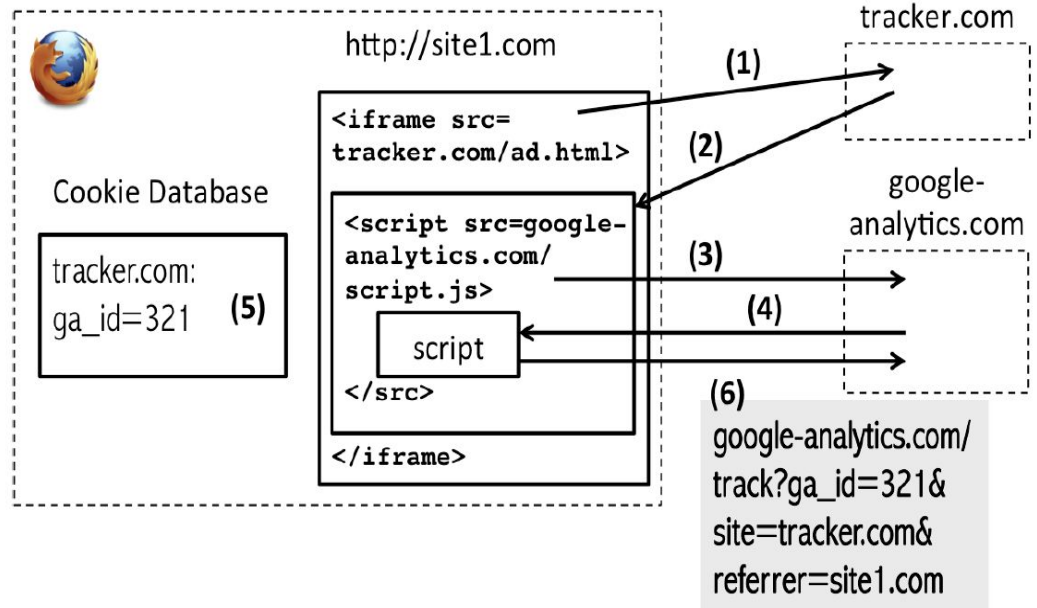
# Referred Analytics Tracking

When **Google Analytics** is embedded by another third-party tracker, rather than by the visited website itself, referred tracking emerges.

The site-owned cookie that GA sets on tracker.com becomes a tracker-owned cookie when tracker.com is embedded on site1.com.

The tracker then passes this identifier to **Google Analytics**, which gains the ability to track the user across all sites on which tracker.com is embedded.

A combination of analytics and referred tracking.



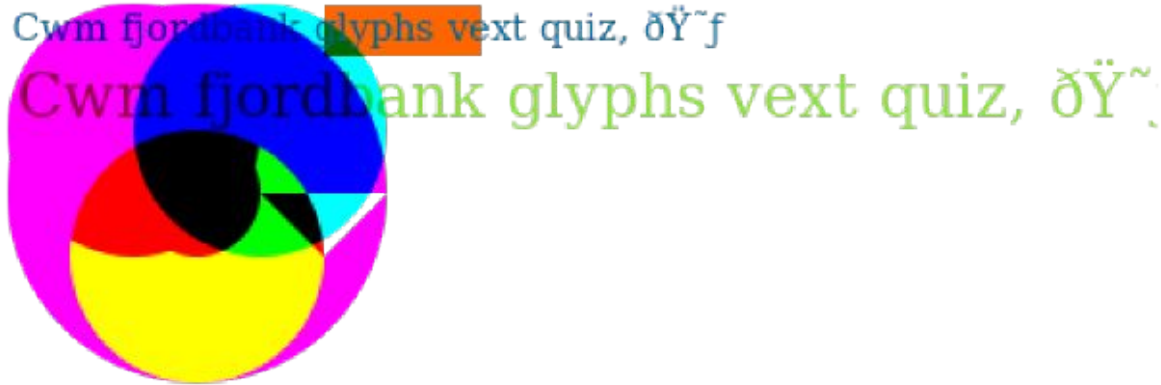
# 1.b Stateless Web tracking

1. Canvas Fingerprinting
2. AudioContext Fingerprinting
3. WebRTC-based Fingerprinting
4. Canvas Font Fingerprinting

# Canvas Fingerprinting

The HTML Canvas allows web application to draw graphics in real time, with functions to support drawing shapes, arcs, and text to a custom canvas element.

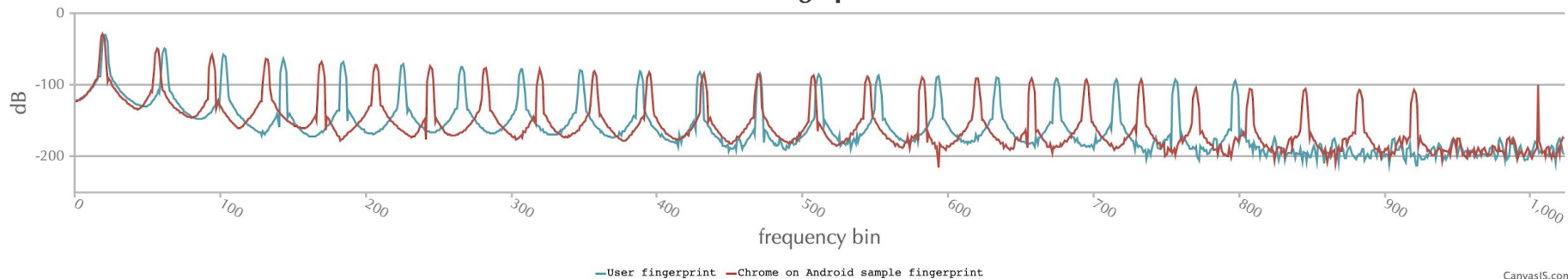
Differences in font **rendering**, **smoothing**, **anti-aliasing**, as well as other device features cause devices to draw the image differently. This allows the resulting pixels to be used a part of a device fingerprint.



# AudioContext Fingerprinting

**VISUALIZATION:**

**Audio Fingerprint**



# AudioContext Fingerprinting

Trackers are attempting to utilize the **Audio API** to fingerprint users in multiple ways.

This **does not require** access to the device's **microphone**, and instead relies on differences in the way the generated signal is processed.

Audio signals processed on different machines or browsers may have **slight differences** due to **hardware or software differences** between the machines, while the same combination of machine and browser will produce the same output.

# Tor

What is Tor again...?

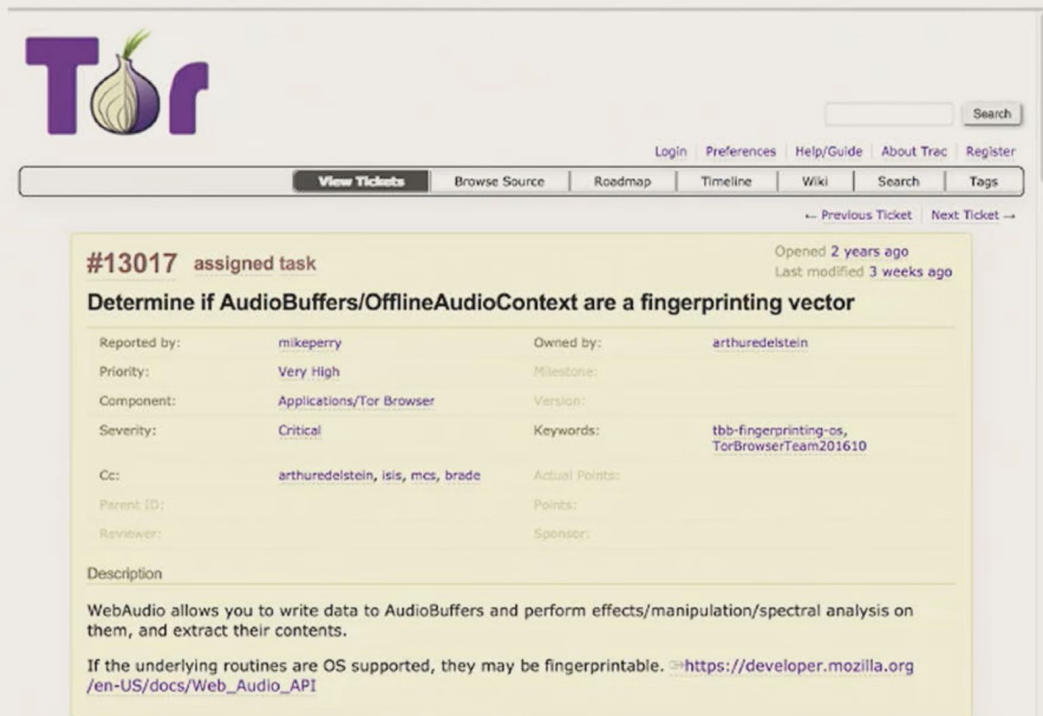
Do you think Tor can block audioContext fingerprinting ...?



# Implications for Tor Browser

271 samples from the Tor Browsers

- 7 distinct fingerprints (2 fingerprints account for 80% of samples)
- Overlap with fingerprints from Firefox shows these largely reveal OS of device



The screenshot shows the Tor Project's bug tracker interface. At the top left is the Tor logo, which features a purple onion. To the right of the logo is a search bar with a 'Search' button. Below the logo and search bar is a navigation menu with links for 'Login', 'Preferences', 'Help/Guide', 'About Trac', and 'Register'. A secondary navigation bar contains 'View Tickets', 'Browse Source', 'Roadmap', 'Timeline', 'Wiki', 'Search', and 'Tags'. Below this is a breadcrumb trail: '← Previous Ticket | Next Ticket →'. The main content area displays a task card for issue #13017, titled 'Determine if AudioBuffers/OfflineAudioContext are a fingerprinting vector'. The card includes the following details:

- #13017 assigned task** (Opened 2 years ago, Last modified 3 weeks ago)
- Reported by:** mikeperry
- Owned by:** arthurelstein
- Priority:** Very High
- Milestone:**
- Component:** Applications/Tor Browser
- Version:**
- Severity:** Critical
- Keywords:** tbb-fingerprinting-os, TorBrowserTeam201610
- Cc:** arthurelstein, isis, mcs, brade
- Actual Points:**
- Parent ID:**
- Points:**
- Reviewer:**
- Sponsor:**

**Description**

WebAudio allows you to write data to AudioBuffers and perform effects/manipulation/spectral analysis on them, and extract their contents.

If the underlying routines are OS supported, they may be fingerprintable. ↗[https://developer.mozilla.org/en-US/docs/Web\\_Audio\\_API](https://developer.mozilla.org/en-US/docs/Web_Audio_API)

source: [https://www.youtube.com/watch?v=\\_mElv9wOkro](https://www.youtube.com/watch?v=_mElv9wOkro)

# WebRTC-based Fingerprinting

WebRTC is a framework for **peer-to-peer** Real Time Communication **in the browser**, and accessible via Javascript.

To discover the best path between peers, each peer collects all available candidate addresses, including addresses from the local network interfaces (such as ethernet or WiFi) and addresses from the public side of the NAT and makes them available to the web application without explicit permission from the user.

WITHOUT user permission, a tracking script that uses WebRTC dataChannel can:

1. Reveal the user's real IP address when behind a VPN
2. Reveal the user's local IP address for each local interface

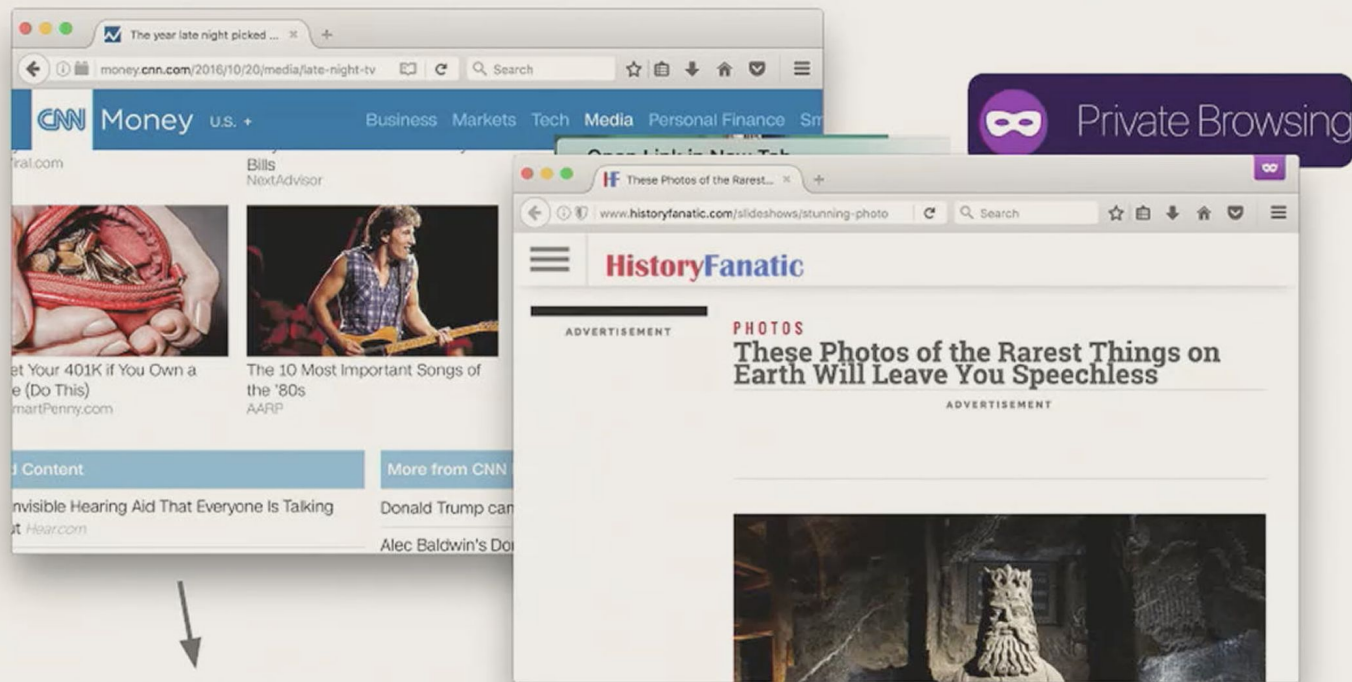
# Canvas Font Fingerprinting

The HTML Canvas API provides a third method to deduce the fonts installed on a particular browser. The canvas rendering interface exposes a `measureText` method, which provides the resulting width of text drawn to canvas.

A script can attempt to **draw text using a large number of fonts** and then **measure the resulting width**.

If the text's width is not equal to the width of the text using a default font (which would indicate that the browser does not have the tested font), then the script can conclude that the browser does have that font available.

# Using Battery Status to Track



Battery Status:  
level: 0.11  
dischargeTime: 12867

The Leaking Battery, Olejnik et. al. (2015)

Battery Status:  
level: 0.11  
dischargeTime: 12867

source: [https://www.youtube.com/watch?v=\\_mElv9wOkro](https://www.youtube.com/watch?v=_mElv9wOkro)

# Question & Discussion

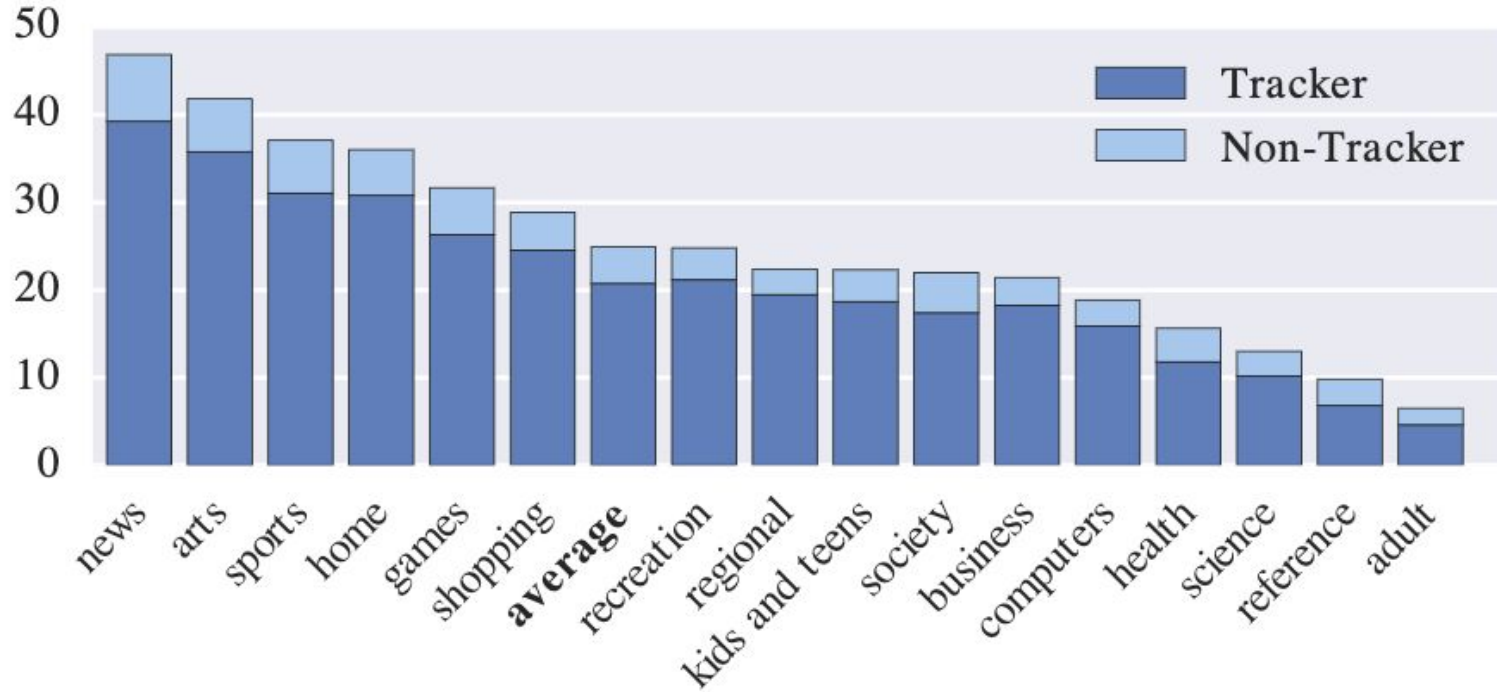
Why go such distance to implement and use web trackers?

## 2. Large Scale View of Web Tracking

Online tracking: A 1-million-site measurement and analysis

- a. Results and finding

# News Sites Have the Most Trackers



# Trend in Fingerprinting

|              |                       |                       |                       |
|--------------|-----------------------|-----------------------|-----------------------|
| Canvas       | Found on 14,371 sites | 400 different scripts | Measured before       |
| AudioContext | Found on 67 sites     | 3 scripts             | New technique by 2016 |
| WebRTC       | Found on 715 sites    | 1 script*             | New technique by 2016 |
| Canvas-font  | Found on 3250 sites   | 6 scripts             | New technique by 2016 |

| Rank Interval | % of First-parties |             |        |
|---------------|--------------------|-------------|--------|
|               | Canvas             | Canvas Font | WebRTC |
| [0,1K)        | 5.10%              | 2.50%       | 0.60%  |
| [1K,10K)      | 3.91%              | 1.98%       | 0.42%  |
| [10K,100K)    | 2.45%              | 0.86%       | 0.19%  |
| [100K,1M)     | 1.31%              | 0.25%       | 0.06%  |

\* calling dataChannel()



# 3. Historical View of Web Tracking

**Internet Jones and the Raiders of the Lost Trackers:  
An Archaeological Study of Web Tracking from 1996 to 2016**

- a. The Wayback Machine
- b. TrackingExcavator
- c. Results

# The Wayback Machine

## Challenges & Opportunities

- ... only provides partial view of third-party requests
- ... considers all third-party requests, in addition to confirmed trackers
- ... allows us to study trends over time
- ... includes traces of popular trackers
- ... provides additional data beyond requests

### Problems:

1. Robots Exclusion
2. Not Archived pages
3. Wayback Escapes
4. Inconsistent archives

2008

updated 4:24 p.m. EDT, Wed September 3, 2008

Make CNN Your Home Page



### Palin's path from city hall to governor's mansion

When she played basketball in high school, Sarah Palin, the soon-to-be Republican vice presidential nominee, earned the nickname "Sarah Barracuda" for her fierce competitiveness. In the 21 months she has served as governor of Alaska, no one is suggesting she's lost her fighting edge. full story

- Palin ready to lead? | Analysts: Her task
- Palin to call for reform in RNC speech
- Will her gender sway women to Palin?

HEALTH MAGAZINE

Secrets to a healthy heart

### Latest News

- Poll measures race in three key states 19 min
- Dems blast Lieberman, say he lied to delegates
- Ticker: McCain greets Palin's daughter, fiancé
- Martin: Keep politics out of Bristol Palin issue
- Biographer: America, meet Sarah Palin
- iReport.com: Can a mom of 5 be VP?
- Cafferty: Should McCain consider different VP?
- CNNMoney: Auto sales plunge
- Gunman's bloody trail leads to six bodies
- Caylee's grandma disputes evidence
- Nagin: Residents won't be turned away
- CNNMoney: Gulf oil production scrambling back
- Hanna roams Caribbean, may hit U.S.
- Pakistan protests 'coalition air attack'
- The cartoonist beloved by GIs and regular guys
- 8 dogs die in hot truck during lunch break
- Review: Google's Chrome needs more polish
- New Marilyn Monroe film scenes found
- Beyonce's sister aims for a path of her own
- CNN Wire: Paraguay coup plot alleged

all news from the past 24hrs » | all most popular »

### Republican National Convention »

- Bernstein: Democrats better take note
- Bush, McCain still an uneasy alliance
- Interactive: How conventions work
- McCain arrives in Minnesota for convention
- Key players Trivia The Forum In Depth

Viewer's Guide

### Video »



Gustav recovery briefing



Dogs die during lunch break 1:39



The RNC in St. Paul

LIVE: Watch coverage of the RNC & Tropical Storm Hanna



ADVERTISEMENT

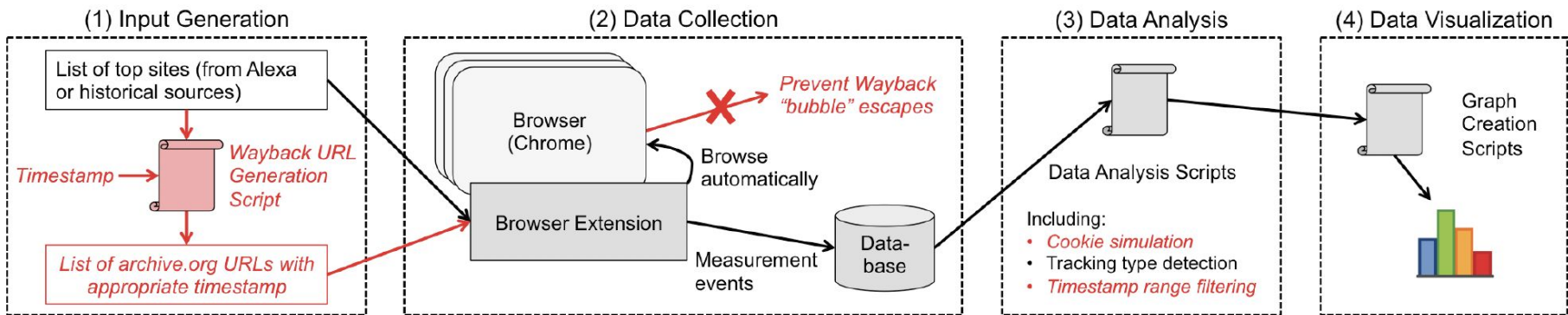
2012

### CNN TV »

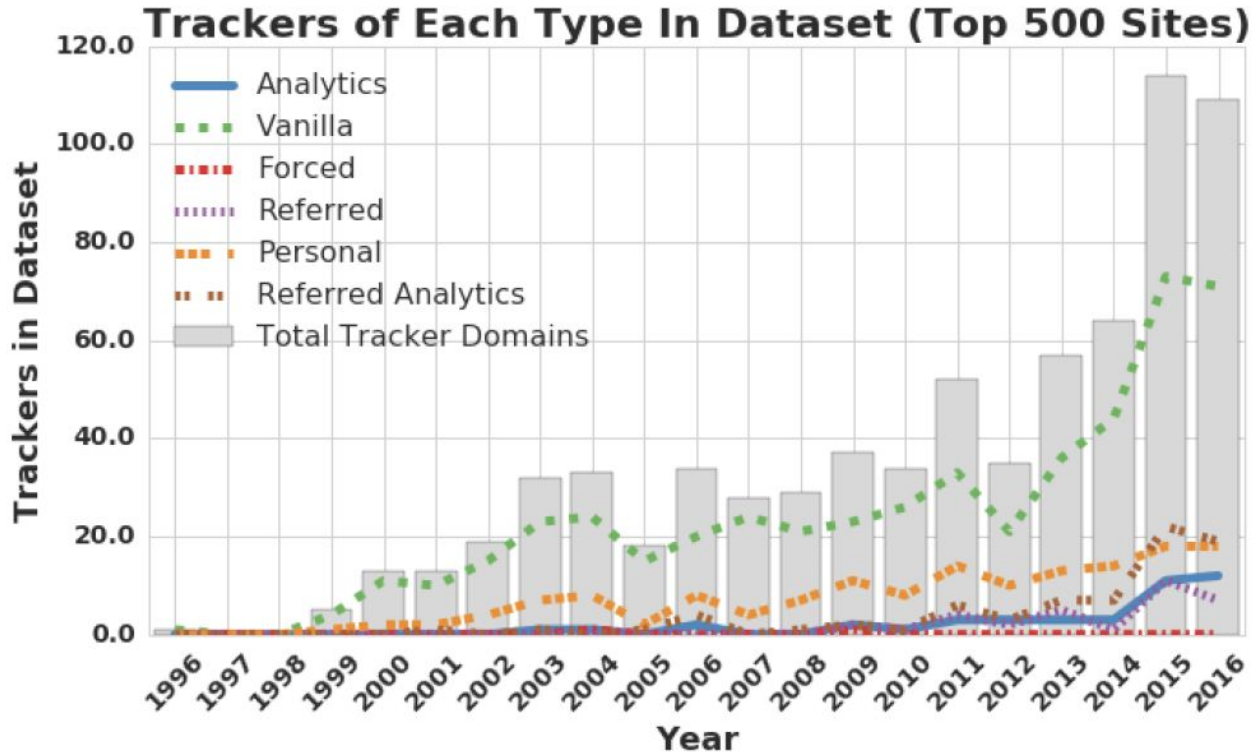


Palin in prime-time

# TrackingExcavator



# Evolution of cookie-based trackers



# Growth in complexity - Cookie Based

| Year | 1Type        | 2Type       | 3Type     | 4Type |
|------|--------------|-------------|-----------|-------|
| 1996 | 100.00% (1)  | 0           | 0         | 0     |
| 1998 | 0            | 0           | 0         | 0     |
| 2000 | 100.00% (13) | 0           | 0         | 0     |
| 2002 | 100.00% (19) | 0           | 0         | 0     |
| 2004 | 96.97% (32)  | 3.03% (1)   | 0         | 0     |
| 2006 | 100.00% (34) | 0           | 0         | 0     |
| 2008 | 100.00% (29) | 0           | 0         | 0     |
| 2010 | 94.12% (32)  | 2.94% (1)   | 2.94% (1) | 0     |
| 2012 | 88.57% (31)  | 11.43% (4)  | 0         | 0     |
| 2014 | 93.75% (60)  | 4.69% (3)   | 1.56% (1) | 0     |
| 2016 | 86.24% (94)  | 11.01% (12) | 2.75% (3) | 0     |

Table 4: Complexity of trackers, in terms of the percentage (and number) of trackers displaying one or more types of tracking behaviors across the top 500 sites.

# Growth in complexity - Fingerprinting

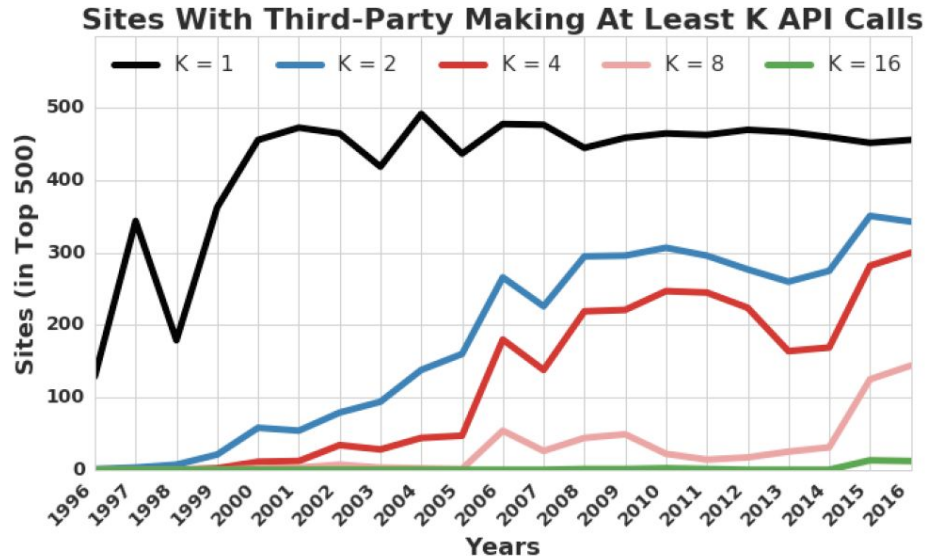


Figure 5: Number of sites in each year with a tracker that calls (on that site) at least K (of our 37) fingerprint-related APIs.

| Year | Most Prolific API-user | Num APIs Used | Coverage |
|------|------------------------|---------------|----------|
| 1998 | reahollywood.com       | 2             | 1        |
| 1999 | go2net.com             | 2             | 1        |
| 2000 | go.com                 | 6             | 2        |
| 2001 | akamai.net             | 8             | 15       |
| 2002 | go.com                 | 10            | 2        |
| 2003 | bcentral.com           | 5             | 1        |
| 2004 | 163.com                | 9             | 3        |
| 2005 | 163.com                | 8             | 1        |
| 2006 | sina.com.cn            | 11            | 2        |
| 2007 | googlesyndication.com  | 8             | 24       |
| 2008 | go.com                 | 12            | 1        |
| 2009 | clicksor.com           | 10            | 2        |
| 2010 | tribalfusion.com       | 17            | 1        |
| 2011 | tribalfusion.com       | 17            | 2        |
| 2012 | imedia.cz              | 12            | 1        |
| 2013 | imedia.cz              | 13            | 1        |
| 2014 | imedia.cz              | 13            | 1        |
| 2015 | aolcdn.com             | 25            | 5        |
| 2016 | aolcdn.com             | 25            | 3        |

# Growth of the top trackers

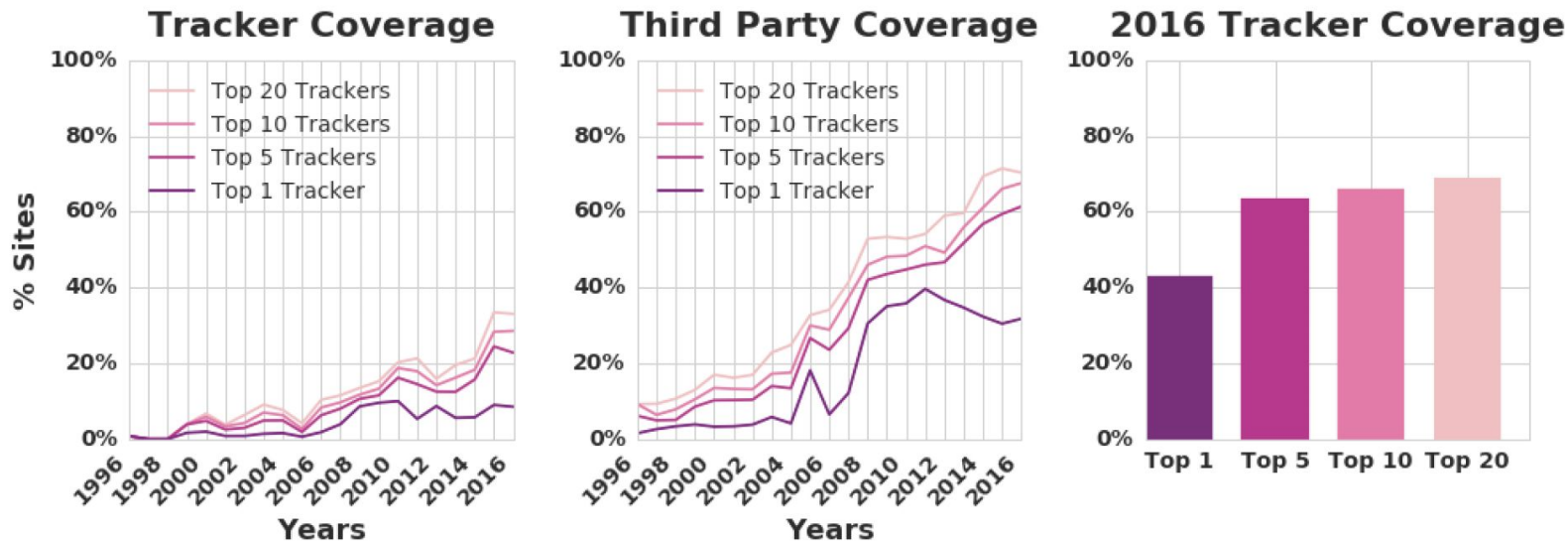


Figure 9: The growth in the coverage (percentage of top 500 sites tracked) of the top 1/5/10/20 trackers for each year is shown in the first and second panels, for all confirmed trackers and for all third parties respectively. The right hand panel shows the values on the live web for confirmed trackers, with the top 5 trackers covering about 70% of all sites in the dataset. Note that top third party coverage in the archive is an excellent proxy for modern confirmed tracker coverage today.





# Questions & Discussion

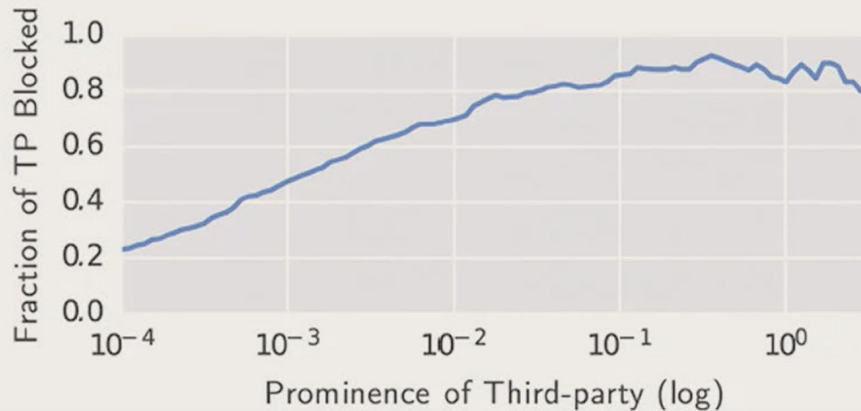
Should web trackers be legal?

How does GDPR play into this?

Any defense against web tracking?

# Privacy tools effectively block stateful tracking

- Third-party cookie blocking
  - 32 out of 50,000 sites work around blocking by redirecting the top-level domain
  - Average number of third-parties per site reduced from ~18 to ~13
- Ghostery
  - Average number of third-parties per site reduced from ~18 to ~3
  - Very few third-party cookies are set



source: [https://www.youtube.com/watch?v=\\_mElv9wOkro](https://www.youtube.com/watch?v=_mElv9wOkro)