18. Tracking on the Web

Blase Ur and David Cash
February 21st, 2022
CMSC 23200 / 33250
The Online Tracking Ecosystem
Online Tracking

- Advertisers want to show you advertisements targeted to your interests and demographics
Online Tracking

• First party = the site you are visiting (whose address is in the URL bar)
• Third party = other sites contacted as a result of your visit to that site
• First-party tracking (e.g., for search)
  – Consider DuckDuckGo and alternatives
Data-Driven Inferences

You might like dogs!
Targeted Advertising

Affinity audiences

Music Lovers
News Junkies
Nightlife Enthusiasts
Outdoor Enthusiasts
Pet Lovers
Targeted Advertising

Dog beds and other pet supplies

PetStuffSite
petstuffsite.com
Targeted Advertising

Dog beds and other pet supplies

PetStuffSite
petstuffsite.com
Mechanics of Tracking

- Canonically, tracking is accomplished via HTTP cookies
  - Third-party cookies
Mechanics of Online Tracking

- JavaScript / images from advertising networks loaded as part of your page
  - In iframes
  - Or sometimes not
  - Why does this matter?

- Let’s discuss: what can an advertising network learn, and how?
Mechanics of Cookie Syncing

Figure 1: Example of advertiser.com and tracker.com synchronizing their cookieIDs. Interestingly, and without having any code in website3, advertiser.com learns that: (i) cookieIDs userABC==user123 and (ii) userABC has just visited the given website. Finally, both domains can conduct server-to-server user data merges.

Browser fingerprinting

• Use features of the browser that are relatively unique to your machine
  – Fonts
  – GPU model anti-aliasing (Canvas fingerprinting)
  – User-agent string
  – *(Often not)* IP address *(Why not?)*
Device Fingerprinting

- Use unique(-ish) combination of device features as an identifier
- [https://panopticlick.eff.org/](https://panopticlick.eff.org/)
Ad Bidding Marketplaces

DATA LEAKAGE IN ONLINE ADVERTISING

This is the current process of real-time bidding that is used in online behavioural advertising.

Legend
- Yellow: Channel of data leakage
- Green: Money
- Red: Personally identifiable information

PageFair
Transparency About Online Tracking
Ubiquity of Online Tracking
Existing Privacy Tools
Existing Privacy Tools

Privacy Badger detected 45 potential trackers on this page. These sliders let you control how Privacy Badger handles each one. You shouldn't need to adjust them unless something is broken.

- weather.api.cnn.io
- rtax.criteo.com
- ad.doubleclick.net
- googleads.g.doubleclick.net
- securepubads.g.doubleclick.net

Disallow Privacy Badger for This Site
Did Privacy Badger break this site? Let us know!
Donate to EFF

Ghostery found 15 trackers on www.cnn.com
14 Blocked

- Advertising
  - Amazon Associates
  - ChartBeat
  - Criteo
  - DoubleClick
  - Google-Publisher-Tags
  - Krux-Digital
  - NetRatings-SiteCensus
  - Outbrain
  - Rubicon
  - ShareThrough

- Site Analytics
  - 2 Trackers
  - 2 Blocked

15 Trackers
Block All
Do not track

- Proposed W3C standard
- User checks a box
- Browser sends “do not track” header to website
- Website stops “tracking”
- W3C working group tried to define what that means
- These days, mostly a no-op (it doesn’t do anything)
Tools to stop tracking, effective?

- **Browser privacy settings**
  - Cookie blocking
  - P3P
  - Tracking Protection Lists
  - Do Not Track
- **Browser add-ons**
- **Opt-out cookies**
- **Digital Advertising Alliance (DAA) AdChoices icon and associated opt-out pages**
Visualization: Connection Graphs
(My Group’s) Tracking Transparency

On The New York Times - Breaking ..., there are 5 trackers.

One of these trackers is Google, which knows about your activity on this page and 3829 others.

In total, 169 trackers have seen you visit 12313 pages. The Tracking Transparency extension has determined that these companies could have inferred your interest in 162 topics.
What are trackers and interests?

When you browse online, your online activity can be tracked by ad networks and analytics companies. We call these trackers.

These companies track your browsing to make guesses about what topics you might be interested in. We call these topics interests.

Companies can personalize your online experience based on these interests. Click on the circles above to learn more.

Your Top Trackers
1. Google
2. Chartbeat
3. Optimizely
4. Microsoft
5. Amazon.com

Your Top Interests
1. Law & Government
2. Online Communities
3. People & Society
4. News
5. Shopping

39 Trackers encountered
7 Pages visited
6 Potential interests

Recent Interests
- Law & Government
- Computers & Electronics
- Shopping
- News
- People & Society

Recent Sites
- twitter.com
- eff.org
- amazon.com
- cnn.com
- nytimes.com
(My Group’s) Tracking Transparency

Who is tracking you?

47 trackers have been present on the sites you visited since installing Tracking Transparency. Your most frequently encountered tracker is Google, which was present on 63.79% of the pages you visited.

Google

Present on 37 pages (63.79% of all pages)

37 Pages 11 Sites 14 Interests

Google was present on 37 pages across 11 sites that you visited since installing Tracking Transparency. From those tracking encounters, they may have guessed that you are interested in 14 topics.

More about this tracker ➤
(My Group’s) Tracking Transparency

1) Categories keywords extraction
   - Google Ads Category
   - Wikipedia API
   - Relevant Articles
   - Text Preprocessing
   - Keywords Extraction
   - Category Keywords List

2) Webpage keywords extraction
   - Given Webpage
   - Text Extraction
   - Text Preprocessing
   - Keywords Extraction
   - Webpage Keywords List

Keywords Matching

Assign Category
(My Group’s) Tracking Transparency

What interests might they think you have?

Trackers collect information about the pages you visit and use this information to identify topics, or interests, that might be relevant to you. These interests are then used to target ads to you and personalize what you see online. Companies don’t usually reveal how they determine your potential interests. Based on the pages you visited, Tracking Transparency’s simulations have identified **52 topics** trackers might think are relevant to you.

For example, you recently visited theverge.com, which Tracking Transparency has determined may be about News.

The chart below shows the interests suggested by your browsing activity. Click a slice of the chart to see more details.

Home Improvement

- **1** Sites
- **8** Trackers

1 of the sites you visited were about Home Improvement. These sites contained a total of 8 trackers.

Home Improvement is a somewhat popular interest.

Other people are often somewhat comfortable with having their interest in this topic being used to personalize their web experience.

More about this interest

Filters

- Recency: ALL, 24 HRS, 7 DAYS
- Popularity: ALL, LESS, MORE
- Comfort: ALL, LESS, MORE
Auditing Online Tracking/Targeting
Alternatives to Cookies for Tracking / Profiling
In-browser Targeting

• The Adnostic research prototype suggested profiling users in-browser and thus choosing ads in-browser from a set
• Key issues include how to properly bill advertisers and how to prevent ad fraud

Google’s FLoC

• Federated Learning of Cohorts

• Clusters users based on their browsing activity and assigns a cohort ID
  – Uses SimHash for clustering
  – Clusters intended to contain 1,000s of users

• Criticisms include fingerprintability, ability to tie cohort to PII, and collapse of different browsing contexts

• (Abandoned in early 2022)
Google’s FLoC

Selecting Interest-based Ads Using FLoC

1. Browsers use a FLoC service to get the mathematical model, consisting of many calculated “cohorts.” In this model, each cohort corresponds to many web browsers having similar recent browsing histories and contains a unique ID.

2. Using that FLoC Model algorithm, your browser calculates your cohort.

3. Let’s say you visited the site of an advertiser abc.com that sells kitchen appliances. Then that site requests the cohort ID from your browser.

4. If you visited additional pages of the advertiser, like searching kitchen utensils, it would record those interests.

5. Advertisers record these cohort activities periodically and share that information with the ad tech company that helps to deliver advertisements.

6. In the same manner, let’s say you visited a publisher site that sells ad space; it will also request your cohort ID.

7. Then the publisher site requests advertisements relevant to that cohort from the ad tech company.

8. The ad tech company combines the data received from the advertiser company about the cohort’s interests and data from the publishing company.

9. Next, the ad tech company chooses suitable ads according to the interests of the cohort.

10. The publisher site then displays the selected advertisement relevant to the interests of the cohort.
Google’s Topics API

Google abandons FLoC, introduces Topics API to replace tracking cookies

Google’s new concept assigns users five interests per week based on web activity

By Emma Roth | Updated Jan 25, 2022, 2:46pm EST
Google’s Topics API

Your browser will store these topics for three weeks before deleting them. Google says that these categories “are selected entirely on your device” and don’t involve “any external servers, including Google servers.” When you visit a website, Topics will show the site and its advertising partners just three of your interests, consisting of “one topic from each of the past three weeks.”

As noted on the Topics API GitHub page, there are currently about 350 available topics in its advertising taxonomy (although Google plans on adding anywhere from “a few hundred” to “a few thousand” eventually). Google says Topics won’t include any “sensitive categories” like race or gender. And if you’re using Chrome, the company is building tools to let you view and delete topics, as well as turn off the feature.