ONLINE TRACKING AND INFERENCING: USING TRANSPARENCY TO OBTAIN USER PERSPECTIVES

Michelle Mazurek, University of Maryland
Tracking and inferencing are pervasive

- Sometimes useful
- Often creepy
- Generally poorly understood
If we want:

- More uptake of PETs
- More effective regulation
- Practices that respect users’ wishes
Then we need:

- More understanding
- More salience
- Less hypothetical imagining!
Our work

- Build tools for transparency of tracking/inferencing
- Ground understanding in real, contextual data
- Measure preferences without hypotheticals
Transparency in practice

- Browser extension: web tracking/inferencing
- Twitter data downloads: behavioral advertising
- Improving transparency/access going forward

* Many slides designed by Miranda Wei
Oh, the Places You’ve Been!

User Reactions to Longitudinal Transparency About Third-Party Web Tracking and Inferencing

By
Ben Weinshel
Miranda Wei
Mainack Mondal
Eurim Choi
Shawn Shan
Claire Dolin
Michelle L. Mazurek
Blase Ur
User-facing tracking controls

- ad- & tracker-blocking extensions
- built-in browser defenses
- privacy dashboards
But these aren’t sufficient

- Some tools show trackers on a given page, but not what this means for you
- Dashboards are generally vague or incomplete

- Instead: Which trackers made *which inferences*?
  - Based on *which browsing activities*?
Our tool
How it works

Arts & Entertainment
> Movies
> Animated Films
Topic modeling

◦ Use Google ad categories as decision classes
◦ Use Wikipedia to train a model
◦ Create a **plausible topic match** per website visited
◦ Fully client-side data collection
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. Facebook
3. comScore
4. Microsoft
5. New Relic

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

300 trackers encountered
38801 pages visited
157 potential interests

Recent Interests
- People & Society
- Law & Government
- Food & Drink
- News
- Travel

Recent Sites
- facebook.com
- Instagram.com
- list-manage.com
- washington.edu
- google.com
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 156 topics trackers might think are relevant to you.

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

Click a slice of the chart to see what trackers think you might be interested in.

Filters

Unpacking Perceptions of Data-Driven Inferences Underlying Online Targeting and Personalization

Claire Dellar, Ben Weisbord, Shawn Shao, Chang Min Hahn, Fairies Cho, Michelle L. Mazurek, Rise Ur
University of Chicago and University of Maryland (1)
(edell, weisbord, shamoo, changm, erin, mazurk, cs.umd.edu (1))
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 **155 topics** trackers might think are relevant to you.

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

**Filters**

- **Recency**: ALL 24 HRS 7 DAYS
- **Popularity**: ALL LESS MORE
- **Comfort**: ALL LESS MORE
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 156 topics trackers might think are relevant to you.

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

Pets & Animals

79 Sites

88 Trackers

79 of the sites you visited were about Pets & Animals. These sites contained a total of 88 trackers.

Pets & Animals is a 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
Your **Pets & Animals** profile

**Pets & Animals** is a popular interest.

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

**Which sites were about Pets & Animals?**

You visited 75 sites that may have been about Pets & Animals. Therefore, trackers may have guessed this is relevant to you. **Click on a bar to learn more.**

![Bar chart showing sites related to Pets & Animals](chart1)

**Which trackers might think you are interested in Pets & Animals?**

89 trackers may have guessed that you are interested in Pets & Animals. **Click on a bar to learn more.**

![Bar chart showing trackers related to Pets & Animals](chart2)
Who is tracking you?

298 trackers have been present on the sites you visited since installing Tracking Transparency. Your most frequently encountered tracker is Google, which was present on 64.34% of the pages you visited.

**Google**

- **Pages**: 24989
- **Sites**: 1349
- **Interests**: 148

Google was present on 24989 pages across 1349 sites that you visited since installing Tracking Transparency. From those tracking encounters, they may have guessed that you are interested in 148 topics.

More about this tracker ➤
Your Google profile

What does Google do?

“Google AdSense is a program run by Google that allows publishers in the Google Network of content sites to serve automatic text, image, video, or interactive media advertisements, that are targeted to site content and audience.” – Source

Based on your browsing, what would Google think your interests are?

Using a machine to assign categories to pages you visit, Google may have guessed that you are interested in 150 topics. Click on a link in the wordcloud to learn more about each interest.

On which sites did Google track you?

Google has tracked you on 1412 sites. Click on a bar to learn more.
Evaluation approach
Iterative usability interviews

13 participants
30 minutes
$10 Amazon gift card
Field study

425 participants, 18+, located in the US, 95% HIT approval rating, use Firefox or Chrome regularly

- **pre-usage survey**
  - [15 minutes, $3]
  - demographics
  - estimates of tracking
  - knowledge and attitudes about OBA

- **use Tracking Transparency for one week**
  - collected telemetry data
  - # of distinct web pages
  - # of trackers
  - inferred topics

- **post-usage survey**
  - [20 minutes, $7]
  - qualitative reactions
  - behavioral intentions
  - estimates of tracking
  - knowledge and attitudes about OBA
Study conditions

Static

Browsing

control

comparison

fully featured

Where were you tracked?

Since installing this browser extension, you have visited 38501 different pages on 1585 sites.

Trackers see which sites you visited through a variety of tracking methods, including third-party cookies, tracking pixels, and browser fingerprinting. When a tracker sees that you have visited multiple sites, they can use that activity to link together your interests.

Most recently visited sites

git.io

github.blog

google.com

github.com
Study conditions

- Static
- Browsing
- Connection Graph
- Tracker List

Control

Comparison

Fully featured
Study conditions

- Static
- Browsing
- Connection Graph
- Tracker List
- No Inferences
- Standard
- Control
- Comparison
- Fully featured

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. Facebook
3. centScore
4. Microsoft
5. New Relic

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

300
38801
157

Recent Interests
- People & Society
- Law & Government
- Food & Drink
- News
- Travel

Recent Sites
- facebook.com
- instagram.com
- b2b-manage.com
- washington.edu
- google.com
Results
Surprised by tracking prevalence

“I was surprised at how much I was tracked by Amazon and Google. I was also surprised to find that I was tracked on over 75% of the pages I visited.” (P369)

“That live.com is the site with the most trackers. That is scary considering I only use live.com for email.” (P140)
Improved knowledge of tracking

“Everything was pretty surprising and it feels like my privacy has been exposed. I never knew that companies tried to create ads supported for me based on my recent searches” (P175)

“It shows my top interest is shopping, which I didn’t figure that to be true, since I usually hate shopping. But it made me realize that I do a lot of shopping online now. That’s new to me. It’s also new that I have 75 potential interests.” (P161)
Improved estimates of tracking

Pre-Usage  Post-Usage

0  40  80  120  160

- Actual
- Standard
- No Inferences
- Connection Graph
- Browsing
- Tracker List
- Static
More likely to use blocking tools

Static
Browsing
Connection Graph
Tracker List
No Inferences
Standard

0.0  0.3  0.5  0.8  1.0

Much more likely
Much less likely
About the same
More likely to use private browsing

- Static Browsing
- Connection Graph
- Tracker List
- No Inferences
- Standard

0.0  0.3  0.5  0.8  1.0

Much more likely  Much less likely  About the same
Takeaways

built a platform to show users data about tracking in their own browsing

tool increased awareness of how inferences are made & ability to quantify tracking

longitudinal visualizations about tracking increased privacy intentions

https://git.io/trackingtransparency
What Twitter Knows
Characterizing Ad Targeting Practices, User Perceptions, and Ad Explanations Through Users’ Own Twitter Data

Miranda Wei, Madison Stamos, Sophie Veys, Nathan Reitinger, Justin Goodman, Margot Herman, Dorota Filipczuk, Ben Weinshel, Michelle L. Mazurek, Blase Ur
Targeted advertising: expectations

Miranda Wei
- **gender**: F
- **location**: Seattle, WA
- **interests**: cats, ramen, Battlestar Galactica

Basic demographics, interests, location
Reality: organic ketchup & more

targetingType: Tailored audiences (lists)
targetingValue:
NCS_PD_04358_Kraft_Organic and natural ketchup buyers_1_26362226
...

targetingType: Keywords
targetingValue: #parenting
...
What ad targeting mechanisms exist and how are they used to target Twitter users?

What do Twitter users think about the mechanisms for...

(a) ad targeting?
(b) transparency?
Study protocol

- Request Twitter data
- Upload ad-related data
- Take customized survey

231 participants
240,000 ads

Files:
- ad-impressions.js
- personalization.js
- twitterAdvertiserList.pdf
Targeting types

demographic:
characteristics about user and their device(s)
provided by user or inferred by Twitter

<table>
<thead>
<tr>
<th>targeting type</th>
<th>uses</th>
</tr>
</thead>
<tbody>
<tr>
<td>language</td>
<td>350,121</td>
</tr>
<tr>
<td>platform</td>
<td>32,351</td>
</tr>
<tr>
<td>location</td>
<td>31,984</td>
</tr>
<tr>
<td>new device</td>
<td>236</td>
</tr>
</tbody>
</table>
## Targeting types

**demographic:**
characteristics of user and their device(s)
- provided by user or inferred by Twitter

<table>
<thead>
<tr>
<th>targeting type</th>
<th>uses</th>
</tr>
</thead>
<tbody>
<tr>
<td>language</td>
<td>350,121</td>
</tr>
<tr>
<td>platform</td>
<td>32,351</td>
</tr>
<tr>
<td>location</td>
<td>31,984</td>
</tr>
<tr>
<td>new device</td>
<td>236</td>
</tr>
</tbody>
</table>

**psychographic:**
user lifestyles, behaviors or attitudes
- provided by user or inferred by Twitter

<table>
<thead>
<tr>
<th>targeting type</th>
<th>uses</th>
</tr>
</thead>
<tbody>
<tr>
<td>behavior</td>
<td>35,008</td>
</tr>
<tr>
<td>interest</td>
<td>25,284</td>
</tr>
</tbody>
</table>
## Targeting types

**demographic:** characteristics of user and their device(s)
- provided by user or inferred by Twitter

<table>
<thead>
<tr>
<th>targeting type</th>
<th>uses</th>
</tr>
</thead>
<tbody>
<tr>
<td>language</td>
<td>350,121</td>
</tr>
<tr>
<td>platform</td>
<td>32,351</td>
</tr>
<tr>
<td>location</td>
<td>31,984</td>
</tr>
<tr>
<td>new device</td>
<td>236</td>
</tr>
</tbody>
</table>

**psychographic:** user lifestyles, behaviors or attitudes
- provided by user or inferred by Twitter

<table>
<thead>
<tr>
<th>targeting type</th>
<th>uses</th>
</tr>
</thead>
<tbody>
<tr>
<td>follower lookalikes</td>
<td>590,502</td>
</tr>
<tr>
<td>conversation</td>
<td>128,005</td>
</tr>
<tr>
<td>behavior</td>
<td>35,008</td>
</tr>
<tr>
<td>interest</td>
<td>25,284</td>
</tr>
</tbody>
</table>

**advertiser:** user information collected offline
- provided by advertiser

<table>
<thead>
<tr>
<th>targeting type</th>
<th>uses</th>
</tr>
</thead>
<tbody>
<tr>
<td>tailored (list)</td>
<td>113,952</td>
</tr>
<tr>
<td>mobile</td>
<td>21,631</td>
</tr>
<tr>
<td>tailored (web)</td>
<td>18,016</td>
</tr>
</tbody>
</table>
Potential policy violations?

prohibited to target by race, religion, sex life, health, politics, and financial status
yet...

- **keyword**: “unemployment,” “Gay,” “#AfricanAmerican,” “latinas”

- **conversation**: “Liberal Democrats (UK)”

- **tailored list**: “YYYY account status: balance due,” “Christian Audience to Exclude,” “LGBT Suppression List”
Opinions of targeting types

OK in the abstract
Creepy in specific (e.g., event targeting)

25% no to general
65% no to specific

accuracy correlated with fairness
... to a point
Comparing ad explanations

- Control: Because advertiser bought an ad
- Current: Facebook, Twitter style

Why am I seeing this ad?

You're seeing this ad because you're on a list Postmates wants to reach on Twitter. When the list was uploaded, Twitter did not learn any new identifying information about you.

Your Data

Postmates
Learn more about Postmates

- March 17, 2019
  Postmates uploaded a hashed list. Twitter matched your information with information on that list.
- April 17, 2019
  You saw this ad from Postmates

There may be other reasons you're seeing this ad, including that Postmates wants to reach people who are ages 18 and up, are located in Phoenix AZ, US, and are Female. This information is based on your Twitter profile and where you've connected to the internet.
Comparing ad explanations

- Control: Because advertiser bought an ad
- Current: Facebook, Twitter style
- Speculative: Detailed text, detailed visual
Comparing ad explanations

- Control: Because advertiser bought an ad
- Current: Facebook, Twitter style
- Speculative: Detailed text, detailed visual
- Speculative: Creepy

creepy

Why am I seeing this ad?

You saw this ad on April 17, 2019 at 11:05 AM on the Twitter app from a(n) Android device, IP address ###.#.#.## (Phoenix AZ, US).

You are seeing this ad because Postmates used your information, such as your email address or phone number, to find you on Twitter.

You are also seeing this ad because Postmates has made the following determinations about you:

- Your information on Twitter was matched with external lists called Suppression (Installs All Time) (email), Suppression (Installs All Time) (Device Id), and Email Suppression List (May 2018)
- You have a lot in common with people who follow @chrishemsworth, and @BarackObama
- You are interested in Health news and general info.
- You are participating in the conversation about Fitness on Twitter.
- You are ages 18 and up, and are Female.
- You are located in or around Phoenix AZ, US.
Ad explanation results

**Useful**
- Control
- Twitter
- Detailed Visual

**Want similar**

% agreement
- strongly agree
- agree
Takeaways

- Data access enables transparency!
  - But is not really all that transparent
- Concerning targeting types are understudied
- How to enforce policy when skirting is trivial?
- Ad explanations should be more detailed
  - People liked “creepy”!
PURSUING USABLE AND USEFUL DATA DOWNLOADS

Sophie Veys, Daniel Serrano, Madison Stamos, Margot Herman, Nathan Reitinger, Michelle L. Mazurek, and Blase Ur
During the Twitter study ...

- Data downloads have lots of useful info
- But they are also undocumented, hard to interpret
- Confusion between rights of access, portability

How should we reimagine data downloads for people?
How do users react to the format and content of data downloads?

What information do they find important? What practical uses do they imagine?

How should data downloads be redesigned to support transparency and other goals?
Method: Focus groups and co-design

- Request data
- Join focus group
- Explore files
- Scavenger hunt
- Data viz 101
- Sketch activity
Focus group details

○ Two sessions for each of six platforms:
  ○ Amazon (orders), Facebook, Google (search), Spotify, Uber, YouTube

○ 3-5 participants per session
  ○ Only one student; only one IT expert per session
  ○ Remotely via Hangouts
  ○ Extensive privacy protections for participants
Results
Surprised by the level of detail

I knew Google is recording everything. It’s just that seeing this in front of me and all the data that has been collected over all the years, it’s like a rude realization. (GA4)

It was like reading a book about myself but not written by myself. (FB1)
Uses and misuses

- Track my own privacy exposure
  - Especially in case of data breach
- Misuse by law enforcement?
Inferences and synthesis

- Desire to synthesize different kinds of data to find patterns
- e.g., Examine spending data or site usage over time

What’s interesting to me is how my online behavior is affecting how this company and all the affiliates see me. And in what category, say, they put me or don’t put me. (FA2)
Disorganized, hard to parse

- Hard to find specific info across multiple files
- Difficulty with JSON files

Most of the interesting data is stored in these files, that as a non-specialist, I can’t read. … We’re effectively illiterate when it comes to reading this additional data that they’ve been collecting. (YA5)
Interactivity desired

- Sort, filter, prioritize
- High-level overview with click/hover to zoom in
- Ability to verify, question, delete in-band
Recommendations: Design

- Organize content
- Allow exploration, filtering, interactivity
- Support aggregation and inferencing (simple scripting?)
- Direct manipulation for participation, erasure

For companies and for third-party organizations
Implications for policy

<table>
<thead>
<tr>
<th>Differentiate</th>
<th>Mandate</th>
<th>Clarify</th>
</tr>
</thead>
<tbody>
<tr>
<td>access from portability</td>
<td>comprehensibility (e.g., via README)</td>
<td>required contents (especially inferences)</td>
</tr>
</tbody>
</table>
Overall takeaways

- **Transparency** and **personalization** help users better understand how their data is collected and used.
- **Concrete examples** lead to meaningful preferences.
- Much **more work needed** to make transparency tools truly **usable and useful**.