# Exposed: Shedding *Blacklight* on Online Privacy*

Lucas Shen† Gaurav Sood‡

July 1, 2025

**Abstract**

To what extent are users surveilled on the web, by what technologies, and by whom? We answer these questions by combining passively observed, anonymized browsing data of a large, representative sample of Americans with domain-level data on tracking from Blacklight. We find that nearly all users ($> 99\%$) encounter at least one ad tracker or third-party cookie over the observation window. More invasive techniques—like session recording, keylogging, and canvas fingerprinting—are less widespread, but over half of the users visited a site employing at least one of these within the first 48 hours. Linking trackers to their parent organizations reveals that a single organization, usually Google, can track over 50% of the average user's web activity. Demographic differences in exposure are modest and often attenuate when we account for browsing volume. However, disparities by age and race remain, suggesting that what users browse—not just how much—shapes their surveillance risk.

**Keywords:** Online Privacy, Online Safety, Digital Divide, KeyLogging, Tracking

*The replication materials are posted on http://github.com/themains/private_blacklight.

†Agency for Science, Technology and Research. lucas@lucasshen.com

‡gsood07@gmail.com

# 1 Introduction

The digital economy increasingly depends on personal data to mediate interactions between users, platforms, and advertisers. As individuals navigate the web, search for information, or engage with apps and services, their activity is routinely logged by a complex ecosystem of tracking technologies. These data flows enable large-scale personalization and behavioral advertising, reshaping the online user experience.

From one perspective, the system has brought real benefits. For consumers, targeted advertising lowers search costs by highlighting products, services, or content that align with their preferences, potentially surfacing relevant options they might not otherwise encounter. For suppliers, especially smaller firms or new entrants, digital targeting offers a cost-effective way to reach relevant audiences without the inefficiencies of mass, untargeted advertising. This improved matching function can expand market reach for niche products and reduce customer acquisition costs. Data shows as much. Disabling cookies can reduce publisher revenue by over 50% (Johnson, Shriver and Du, 2020; Ravichandran and Korula, 2019), with the largest relative losses for small publishers and niche advertisers.

On the flip side, there are real costs to this system. The data that fuels personalization is often collected through opaque and increasingly invasive techniques, ranging from third-party cookies and fingerprinting to session recording and keylogging. These methods power a broader system of surveillance that can result in a wide array of harms. As Citron and Solove (2022) argue, privacy violations can cause physical risks, e.g., stalking, economic losses, e.g., identity theft, psychological harms, e.g., anxiety or loss of trust, and reputational damage. They can also reinforce social inequality through discriminatory and exclusionary practices.

These concerns are magnified by the ease with which ostensibly anonymized data can be re-identified. Even datasets stripped of explicit identifiers can often be traced back to individuals using a small number of behavioral signals—such as search queries, media con-

sumption patterns, or spatio-temporal traces from mobile devices (Achara, Acs and Castelluccia, 2015). When such granular data becomes linkable across contexts, the potential for harm expands.

These risks are not merely hypothetical. In practice, they manifest in the form of predatory or discriminatory targeting. For instance, individuals facing financial hardship are disproportionately targeted with high-interest loans and other exploitative financial products (Christl and Spiekermann, 2016). As recent investigations have shown, users are steered toward more expensive options based on device type, e.g., Mac vs. PC, potentially reducing consumer surplus (Borgesius, 2020; Bujlow et al., 2015; Hannak et al., 2014). Relatedly, some work shows that advertisers and platforms engage in digital redlining, excluding certain users from seeing ads for housing, employment, or credit based on race, location, and other sensitive attributes (Angwin, Tobin and Varner, 2016).

Despite widespread debate over the tradeoffs of online tracking, empirical evidence remains limited on where, how, and to whom these surveillance technologies are deployed. Prior research has typically adopted a site-centric perspective, examining the prevalence of tracking technologies across websites (Acar et al., 2013; Dambra et al., 2022; Englehardt and Narayanan, 2016; Iqbal, Englehardt and Shafiq, 2021; Karaj et al., 2019; Mattu and Sankin, 2020; Niforatos, Zheutlin and Sussman, 2021; Nikiforakis et al., 2013; Sanchez-Rola and Santos, 2018; Sanchez-Rola et al., 2021; Solomos et al., 2020; Zheutlin, Niforatos and Sussman, 2022*b*,*a*). Yet this approach overlooks browsing behavior, and therefore considerably underestimates user-level exposure to tracking (Dambra et al., 2022). How prevalent are advanced tracking techniques like fingerprinting or keylogging at the user level? Which types of users are more likely to encounter such techniques in their everyday browsing? Are certain populations, by virtue of the sites they visit, more exposed to surveillance than others?

This paper addresses these questions by combining two complementary data sources. We begin with passively collected, anonymized browsing data and sociodemographic profiles

for a large, nationally representative sample of American adults, obtained from YouGov. These data provide granular insight into the websites people actually visit, enabling us to assess real-world exposure to tracking technologies rather than relying on stated privacy attitudes or a sample of highly visited sites. Each visited domain is then linked to privacy audit data from Blacklight, a tool developed by The Markup that scans websites for the presence of third-party cookies, device fingerprinting, session recording, keylogging, and redirect-based surveillance. This combined dataset enables us to assess the actual privacy risks that users face online and to quantify disparities in exposure across various demographics, including gender, race, education, and age.

Our analysis offers three key contributions. First, we document the user-centric prevalence of sophisticated surveillance tools across the modern web. Second, we show how exposure varies across demographic groups, revealing new dimensions of digital inequality. Third, we provide a framework for measuring and monitoring privacy harms using passively collected behavioral data—a critical step toward evidence-based privacy policy and accountability.

## 2 Research Design, Data, and Measures

To quantify users' exposure to online tracking, we combine two data sources: (1) a month-long, passively collected, anonymized dataset of domain-level web traffic from a nationally representative panel of 1,200 U.S. adults—covering over six million visits—and (2) domain-level audits from Blacklight, a real-time scanning tool developed by The Markup that detects seven types of tracking technologies, including more invasive techniques like session recording and canvas fingerprinting (Section 2.2).

We construct two complementary measures of user-level exposure (Section 2.3). The first is cumulative exposure, the total number of tracker encounters during the observation window. The second is a rate-adjusted measure that normalizes by browsing volume, captur-

ing the average number of trackers per visit. This distinction allows us to separate exposure due to time spent online from that driven by browsing choices. A very small number of panelists have no observed traffic during the study period and are excluded from the analyses. We assume these cases are missing completely at random. Similarly, not all visited domains return successful analyses from Blacklight, due to technical issues like temporary errors and redirects. These instances are excluded from the exposure computations and again assumed to be missing completely at random. We revisit these assumptions in Section 4.

Beyond domain-level exposure, we assess how much of a user's browsing trail is observable by the parent organization, e.g., Meta. To measure this surveillance capacity, we link third-party services to their parent firms and calculate the share of a user's browsing history accessible to any one organization (Section 2.4).

Lastly, we analyze demographic disparities in exposure (Section 2.5), examining how age, race, gender, and education correlate with both the volume and rate of exposure.

## 2.1 Browsing data

Our browsing data comes from YouGov, which maintains a large panel of US adults and uses matched sampling to construct representative samples. This involves drawing a random population from a large synthetic representative sampling frame (Rivers and Bailey, 2009), who are then invited to take a survey. Non-respondents are replaced with similar individuals. Our study sample consists of 1,200 such American adults who have volunteered to install a passive metering software, RealityMine, on their device in lieu of rewards, which collects de-identified web browsing data over a one-month period in June 2022 (Shen and Sood, 2025; Sood, 2022; Sood and Shen, 2024). This software logs visits to web domains with anonymized URLs (e.g., *https://www.google.com/search?ANONYMIZED* or *https://mail.google.com/mail/u/0/?ANONYMIZED*) and visit timestamps regardless of browser type or privacy settings. All participants gave informed consent and were fully aware

of the data collection process, including passive web tracking, which they could opt out of at any time. Personal data such as passwords or secure form entries was excluded, with all data anonymized, including URLs as we described above (please see Appendix A).

**Table 1.** Overview of data

| **A. Sample size** | n | (%) |
|---|---|---|
| No. individuals | 1,132 | — |
| No. domains | 64,074 | — |
| No. visits | 6,297,382 | — |
| No. domains, Blacklight | 34,078 | (53.2%) |
| No. visits, Blacklight | 4,767,099 | (75.7%) |
| **B. Demographics** | n | (%) |
| Female | 635 | (52.9%) |
| Male | 565 | (47.1%) |
| White | 762 | (63.5%) |
| Hispanic | 176 | (14.7%) |
| Black | 152 | (12.7%) |
| Other | 61 | (5.1%) |
| Asian | 49 | (4.1%) |
| High school diploma or below | 427 | (35.6%) |
| Some College education | 350 | (29.2%) |
| College Graduate | 272 | (22.7%) |
| Postgraduate | 151 | (12.6%) |
| $<$ 25 years old | 97 | (8.1%) |
| 25–34 years old | 222 | (18.5%) |
| 35–49 years old | 298 | (24.8%) |
| 50–64 years old | 301 | (25.1%) |
| 65+ years old | 282 | (23.5%) |

Note: Percentages in Panel A represent the proportion of total domains or total visits covered by each tracking tool. Percentages in Panel B indicate the proportion of individuals in each demographic category.

Overall, our data of digital traces includes over 6 million web visits to over 64,000 unique domains from 1,134 individuals over a month (Table 1). 65 individuals had no online activity on their device in the entire month, an additional individual had all visits without relevant metadata such as the URL, and two more had domains with no tracking data.

Our sample also includes individual-level demographics, summarized in Panel B of Table 1 such as gender, race (Black, Hispanic, White, Other), education level (high school diploma or below, some college education, college degree, postgraduate college degree), and age, which we bin into five groups: $<$ 25, 25–34, 35–49, 50–64, and 65+ years old. Our

panel is representative of the US adult population, with the gender, race, education, age, and geography (five regions) closely resembling that of the same-year Current Population Survey (Shen and Sood, 2025).

## 2.2 Measuring Tracking on Domains

Blacklight is an on-demand privacy inspection tool that simulates a fresh user visiting a website and scans for seven types of stateful and stateless tracking methods. Blacklight identifies tracking through browser automation, network request monitoring, and behavioral script analysis. We submitted 64,074 unique domains visited in our sample to Blacklight and obtained results for 34,078 domains (53.25%), covering 76% of all visits in our dataset (Table 1). Specifically, Blacklight detects these seven tracking methods (see Appendix B for more details):

- **Ad Trackers:** Detected via outgoing requests matched to DuckDuckGo's "Ad Motivated Tracking" list.
- **Third-party Cookies:** Detected by analyzing 'Set-Cookie' headers on requests to third-party services.
- **Facebook Pixel and Google Analytics:** Collect granular behavioral data for ad targeting and analytics.
- **Session Recording Scripts:** Detected based on script behavior and a known list of URLs for session replay services.
- **Keylogging:** Identified by typing known values into form fields and monitoring network activity for exfiltration of those exact keystrokes.
- **Canvas Fingerprinting:** Detected by inspecting '`<canvas>`' behavior and analyzing pixel-level script outputs.

Of these, session Recording, keylogging, and canvas fingerprinting are especially invasive (Acar et al., 2014; Karaj et al., 2019; Mattu and Sankin, 2020; Mowery and Shacham,

2012; Senol et al., 2022). These techniques also raise privacy risks beyond conventional tracking, as they bypass commonly proposed hygiene measures such as ad blockers and cookie deletion.

## 2.3 Measuring Exposure to Tracking Methods

To quantify the extent of user-level exposure to online tracking, we link users' browsing data (Section 2.4) with Blacklight scans for domain-level tracking (Section 2.2). Each individual $i$ has a set of site visits $\mathcal{V}_i$, where each visit $v$ corresponds to a timestamped instance of visiting a webpage from domain $d$. Let $d(v)$ denote the domain associated with visit $v$. $|\mathcal{V}_i|$ is the total number of visits for that individual in the month. We compute exposure to one of the tracking methods $s$ detected by Blacklight (Section 2.2) by aggregating tracker counts based on the domain of each visit (Equation (1)). To adjust for varying browsing intensity, we compute a rate-normalized exposure rate, normalizing cumulative exposure by the user's total number of visits (Equation (2)).

$$\text{Cumulative Exposure}_i^{(s)} = \sum_{v \in \mathcal{V}_i} \left| trackers_{d(v)}^{(s)} \right|, \tag{1}$$

$$\text{Exposure Rate}_i^{(s)} = \left( \frac{1}{|\mathcal{V}_i|} \cdot \text{Cumulative Exposure}_i^{(s)} \right) \tag{2}$$

These measures approximate the cumulative volume and rate of behavioral data collected on an individual, reflecting the size of their digital footprint. We use these metrics to examine the extent of privacy exposure online and disparities across demographic groups, leveraging self-reported characteristics collected alongside the browsing data (Section 2.5). In subsequent analyses, we use both measures to examine the extent of individual privacy exposure and its variation across demographic subgroups (Section 2.5).

## 2.4 Measuring Tracking by Organizations: Browsing History

$$|\text{Organizations}_i| = \left| \bigcup_{v \in \mathcal{V}_i} O_{iv} \right| \tag{3}$$

$$\text{Tracking share}_{ij} = \frac{\sum_{v \in \mathcal{V}_i} \mathbf{1}(j \in O_{iv})}{|\mathcal{V}_i|} \tag{4}$$

To measure the breadth and depth of tracking by organizations, we link domain-level metadata from the Blacklight analyses, which identifies the third-party domains (e.g., *connect.facebook.net*) embedded on the private domains, to parent organizations (e.g., *Facebook, Inc.*) using the DuckDuckGo Tracker Radar data (https://github.com/duckduckgo/tracker-radar). The Tracker Radar maps over 38,000 third-party domains to over 19,000 distinct organizations. We then link these parent organizations ($O$) to the visit-level data ($\mathcal{V}$) via the detected third-party domains. This allows us to quantify: (i) the number of distinct organizations tracking each user (Equation (3)) and (ii) how much of a user's browsing activity is visible to any organization $j$ (Equation (4)). Organizations owning multiple third-party domains on the same private domain are counted only once.[1]

## 2.5 Demographic Differences

To estimate disparities in online tracking, we model cumulative exposure and exposure rate as a function of a person's demographics. Specifically,

$$\text{y}_i = \alpha + \beta_1 \text{women}_i + \beta_2^k \text{race}_i + \beta_3^k \text{education}_i + \beta_4^k \text{age group}_i^k + \varepsilon_i, \tag{6}$$

[1] We also compute organizations' shares weighted by dwelling time ($t$):

$$\text{Tracking share}_{ij}^{(\text{dur})} = \frac{\sum_{v \in \mathcal{V}_i} \mathbf{1}(j \in O_{iv}) \cdot t_{iv}}{\sum_{v \in \mathcal{V}_i} t_{iv}}, \tag{5}$$

as an alternative measure of Equation (4), and reach similar findings (Appendix D).

where the outcome measure is individual $i$'s exposure to each of the seven tracking methods from Blacklight. All models are estimated using ordinary least squares with Huber-White robust standard errors. Demographic covariates include gender (woman; ref: man), race/ethnicity (African American, Asian, Hispanic, Other; ref: White), education (some college, college degree, postgraduate; ref: high school or less), and age group (25–34, 35–49, 50–64, 65+; ref: 18–24). Since all demographic predictors are represented as indicator variables, their coefficients can be compared directly.

# 3 Results

The results section is structured as follows. First, we report the prevalence and speed of exposure to the seven tracking technologies. Second, we examine how exposure varies by demographics. Third, we quantify the extent to which a single tracking organization can observe a user's online activity. Finally, we examine demographic differences in the depth of tracking by organizations.

## 3.1 Exposure to Different Kinds of Tracking

**Table 2.** Summary of cumulative exposure

| | Cumulative exposure | | | | | | | Percentage encountering | |
|---|---|---|---|---|---|---|---|---|---|
| | Mean (1) | Std. dev. (2) | Min. (3) | 25p (4) | Median (5) | 75p (6) | Max. (7) | At least 1 (8) | At least 10 (9) |
| Ad Trackers | 27,407 | 48,279 | 0 | 2,620 | 9,738 | 29,240 | 517,968 | 99.6% | 99.1% |
| Third-Party Cookies | 32,325 | 55,184 | 0 | 3,133 | 11,757 | 35,647 | 700,142 | 99.4% | 99.1% |
| Facebook Pixel | 383 | 657 | 0 | 40 | 147 | 463 | 5,808 | 94.7% | 87.3% |
| Google Analytics | 35 | 104 | 0 | 0 | 8 | 29 | 1,619 | 72.4% | 46.5% |
| Session Recording | 155 | 353 | 0 | 10 | 54 | 165 | 5,788 | 89.7% | 76.0% |
| Keylogging | 309 | 935 | 0 | 4 | 26 | 148 | 10,315 | 84.9% | 65.9% |
| Canvas Fingerprinting | 320 | 697 | 0 | 18 | 84 | 288 | 7,643 | 91.7% | 81.0% |

Note: Cumulative exposure to trackers is defined in Equation (1). Columns (8)–(9) report the percentage of people encountering at least one and at least ten trackers within the month.

**Table 3.** Summary of exposure rate

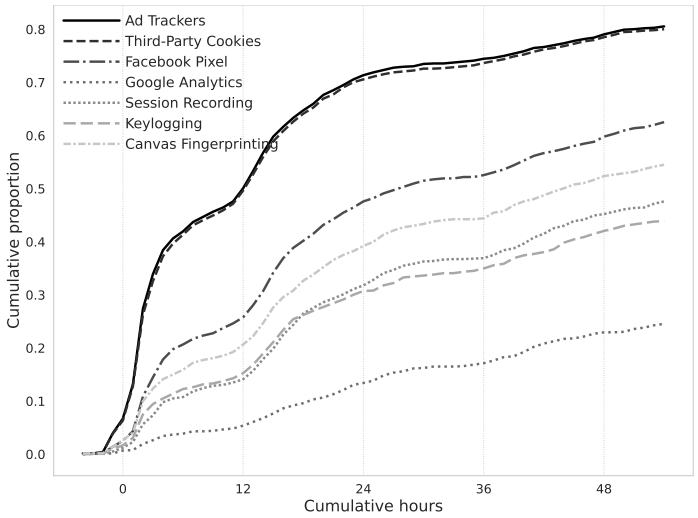
| | Mean (1) | Std. dev. (2) | Min. (3) | 25p (4) | Median (5) | 75p (6) | Max. (7) |
|---|---|---|---|---|---|---|---|
| Ad Trackers | 4.98 | 3.64 | 0.00 | 2.66 | 3.99 | 6.28 | 31.41 |
| Third-Party Cookies | 6.12 | 5.05 | 0.00 | 3.26 | 4.83 | 7.36 | 53.42 |
| Facebook Pixel | 0.08 | 0.09 | 0.00 | 0.03 | 0.06 | 0.11 | 1.00 |
| Google Analytics | 0.01 | 0.04 | 0.00 | 0.00 | 0.00 | 0.01 | 0.99 |
| Session Recording | 0.03 | 0.05 | 0.00 | 0.01 | 0.02 | 0.04 | 0.59 |
| Keylogging | 0.04 | 0.07 | 0.00 | 0.00 | 0.01 | 0.04 | 0.58 |
| Canvas Fingerprinting | 0.06 | 0.09 | 0.00 | 0.01 | 0.04 | 0.07 | 1.00 |

Note: Exposure rates to trackers are defined in Equation (2).

Tracking is near universal, with ad trackers and third-party cookies the most common methods. During the month-long observation period, 99.1% of users encountered more than ten ad trackers or third-party cookies (see Table 2). On average, users encountered 27,407 ad trackers ($\hat{\sigma}$ = 48,279) and 32,325 third-party cookies ($\hat{\sigma}$ = 55,184). The corresponding medians—9,738 and 11,757 (Table 2)—suggest heavily right-skewed distributions. Normalizing by the number of visits dramatically reduces the skew. Users are exposed to, on average, 5 ad trackers ($\hat{\sigma} = 3.6$) and 6.1 third-party cookies ($\hat{\sigma} = 5.1$) per visit (Table 3) with medians of 4 and 4.8 respectively. A tighter spread and lower skew suggest that most of the variation in total exposure is driven by differences in how much users browse.

More invasive tracking methods—session recording, keylogging, and canvas fingerprinting—can be found on nearly 9% of the domains but are encountered less frequently. Users encounter session recording on 3% of visits ($\hat{\sigma}$ = 0.05), keylogging scripts on 4% of visits ($\hat{\sigma}$ = 0.07), and fingerprinting scripts on 6% of visits ($\hat{\sigma}$ = 0.09) (Table 3). This suggests users are likelier to browse domains without invasive tracking. Despite the low rates, the cumulative exposure is non-trivial. For instance, 91.7% of users encountered canvas fingerprinting at least once, and over 65% encountered all three at least ten times (Table 2).

Because tracking is pervasive, exposure is rapid. Using browsing timestamps, we identify when each user first encountered each tracking method. Half of the users encounter




**(a)**

| | 0h | 12h | 24h | 36h | 48h |
|---|---|---|---|---|---|
| Ad Trackers | 0.067 | 0.501 | 0.714 | 0.745 | 0.791 |
| Third-Party Cookies | 0.064 | 0.496 | 0.706 | 0.737 | 0.785 |
| Facebook Pixel | 0.026 | 0.258 | 0.476 | 0.526 | 0.598 |
| Google Analytics | 0.007 | 0.054 | 0.134 | 0.171 | 0.23 |
| Session Recording | 0.012 | 0.141 | 0.318 | 0.369 | 0.452 |
| Keylogging | 0.017 | 0.153 | 0.307 | 0.35 | 0.42 |
| Canvas Fingerprinting | 0.027 | 0.207 | 0.392 | 0.444 | 0.524 |

**(b)**

**Figure 1.** The proportion of users who had encountered a particular tracker by a particular time. We start measuring at 6 PM on 31 May (due to time zones) when at least 50 users have logged browsing activity. The table reports the cumulative proportions at the specified hours.

an ad tracker or a third-party cookie within the first 12 hours of the start of measurement (see Figure 1). By 48 hours, nearly 80% have encountered at least one tracker or cookie. Even the more intrusive techniques—session recording, keylogging, and canvas fingerprinting—reach nearly half the users within 48 hours.

## 3.2 Demographic Differences in Exposure to Tracking Methods

Table 4 and Table 5 (columns (1)–(7)) report regression estimates of demographic differences in cumulative exposure and exposure rate for different tracking methods.

Controlling for other demographic factors, gender is not a strong predictor of net exposure to tracking—except, women encounter canvas fingerprinting significantly more than

**Table 4.** Demographic differences in cumulative exposure

| | Tracking mechanisms | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Ads | Cookies | FB Pixel | GA | Keyloggers | Session rec | Canvas FP | Max share |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Woman | −35.4 | −30.5 | −38.1 | 2.1 | −0.96 | −2.8 | 92.9** | −5.4** |
| | (27.5) | (31.4) | (38.8) | (6.0) | (55.3) | (20.7) | (39.1) | (2.7) |
| Race: African American | −18.6 | −27.6 | −1.8 | −7.0 | −32.8 | −3.3 | −39.0 | −2.5 |
| | (41.4) | (45.0) | (69.5) | (7.7) | (83.7) | (26.3) | (63.0) | (4.2) |
| Race: Asian | 11.3 | 34.5 | 21.8 | 39.6 | −141.1* | −58.7** | 16.8 | 13.2 |
| | (69.9) | (83.0) | (88.3) | (33.8) | (76.5) | (25.5) | (80.4) | (9.2) |
| Race: Hispanic | −40.6 | −41.7 | −76.9** | −17.2*** | −53.8 | −19.7 | −24.2 | −2.5 |
| | (30.6) | (35.3) | (37.3) | (5.3) | (72.6) | (24.9) | (51.0) | (3.8) |
| Race: Other | −13.9 | −2.9 | −10.8 | −12.3 | −137.8** | −33.0 | 191.8 | −4.4 |
| | (57.9) | (75.2) | (90.7) | (7.8) | (67.5) | (27.8) | (146.2) | (4.8) |
| Educ: Some college | 9.9 | 27.9 | 46.8 | 11.2 | −17.1 | 4.8 | 32.7 | 4.1 |
| | (29.6) | (35.2) | (44.2) | (7.1) | (65.1) | (19.9) | (43.8) | (3.0) |
| Educ: College | 126.3*** | 160.9*** | 76.8 | 15.4** | 169.4* | 87.8** | 87.2* | 13.0*** |
| | (43.4) | (49.9) | (48.4) | (7.0) | (87.4) | (35.9) | (52.7) | (3.8) |
| Educ: Postgraduate | 89.9* | 87.6* | 173.0** | 13.0 | 6.2 | 68.2* | 160.9* | 11.1** |
| | (52.1) | (52.4) | (88.1) | (13.3) | (79.6) | (35.4) | (94.2) | (4.9) |
| Age: 25–34 | 20.2 | 22.4 | 31.1 | −8.4 | −95.8 | 0.27 | 35.6 | 2.9 |
| | (25.1) | (31.7) | (54.1) | (13.6) | (116.5) | (32.2) | (51.4) | (5.4) |
| Age: 35–49 | 99.0*** | 97.9*** | 75.7 | 6.3 | 94.6 | 37.4 | 84.2** | 2.5 |
| | (30.6) | (34.9) | (50.9) | (13.2) | (127.2) | (32.5) | (42.3) | (4.8) |
| Age: 50–64 | 185.9*** | 217.8*** | 175.8*** | −4.2 | 146.8 | 75.5** | 152.5*** | 7.5 |
| | (39.3) | (46.0) | (57.3) | (12.0) | (133.8) | (36.9) | (45.7) | (5.1) |
| Age: 65+ | 309.3*** | 351.7*** | 320.3*** | 3.3 | 287.9** | 136.0*** | 358.6*** | 13.7*** |
| | (37.5) | (44.7) | (65.0) | (12.0) | (132.1) | (36.7) | (59.6) | (4.9) |
| Constant | 110.0*** | 125.5*** | 217.4*** | 28.3*** | 188.1* | 73.8** | 69.4* | 21.5*** |
| | (29.2) | (33.9) | (50.3) | (10.4) | (110.9) | (30.0) | (41.6) | (4.5) |
| Dependent variable mean | 274.1 | 323.3 | 383.3 | 35.1 | 309.1 | 155.4 | 319.8 | 30.1 |
| $R^2$ | 0.07 | 0.07 | 0.04 | 0.02 | 0.03 | 0.04 | 0.05 | 0.04 |
| Observations | 1,134 | 1,134 | 1,134 | 1,134 | 1,134 | 1,134 | 1,134 | 1,134 |

*Note:* Each column reports coefficients from estimating Equation (6), where the outcome is the cumulative exposure (Equation (1)) to the seven tracking mechanisms and the number of visits tracked by the top organization in column (8), defined as the tracker organization associated with the highest share of a user's total web visits (Section 2.4). Ad trackers (Ads) and third-party cookies (columns 1–2), and the max share of visits tracked (column 8) are scaled by a factor of 1/100, such that a coefficient of 1 corresponds to 100 tracking instances. Please see Figure C.1 for an alternative visualization of the estimates. Significance levels: * 0.1 ** 0.05 *** 0.01.

men ($\hat{\beta} = 93$, $\widehat{\text{SE}} = 39.1$, $p < .05$) (see Table 4). Racial differences are also limited: Asians are less exposed to keylogging and session recording, those categorized as 'Others' are less exposed to keylogging, and Hispanic users are tracked less frequently by Facebook Pixel and Google Analytics.

**Table 5.** Demographic differences in exposure rate

| | Tracking mechanisms | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Ads | Cookies | FB Pixel | GA | Keyloggers | Session rec | Canvas FP | Max share |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Woman | −0.203 | −0.101 | 0.002 | −0.0009 | 0.000 | 0.004 | 0.011** | −0.017* |
| | (0.211) | (0.291) | (0.005) | (0.002) | (0.004) | (0.003) | (0.005) | (0.010) |
| Race: African American | −0.035 | −0.453 | 0.004 | 0.0003 | 0.004 | 0.009 | −0.0007 | −0.018 |
| | (0.339) | (0.430) | (0.010) | (0.003) | (0.007) | (0.006) | (0.008) | (0.015) |
| Race: Asian | −1.20*** | −1.49*** | −0.020** | −0.002 | −0.018*** | −0.013*** | −0.014* | 0.006 |
| | (0.299) | (0.436) | (0.009) | (0.003) | (0.005) | (0.004) | (0.007) | (0.030) |
| Race: Hispanic | 0.088 | 0.040 | 0.0006 | −0.001 | 0.001 | 0.000 | 0.003 | −0.0002 |
| | (0.322) | (0.452) | (0.008) | (0.003) | (0.007) | (0.004) | (0.007) | (0.015) |
| Race: Other | −0.279 | −0.093 | −0.008 | −0.004** | −0.009 | 0.002 | 0.012 | −0.010 |
| | (0.435) | (0.672) | (0.008) | (0.002) | (0.008) | (0.006) | (0.011) | (0.021) |
| Educ: Some college | 0.192 | 0.188 | −0.002 | −0.0002 | 0.000 | 0.003 | 0.008 | 0.023* |
| | (0.265) | (0.362) | (0.007) | (0.003) | (0.005) | (0.004) | (0.007) | (0.012) |
| Educ: College | 0.490* | 0.761* | −0.010 | −0.003 | 0.006 | 0.002 | 0.001 | 0.039*** |
| | (0.294) | (0.421) | (0.007) | (0.003) | (0.006) | (0.004) | (0.007) | (0.013) |
| Educ: Postgraduate | 0.245 | 0.265 | −0.007 | −0.005 | −0.0002 | 0.004 | 0.017* | 0.054*** |
| | (0.315) | (0.440) | (0.008) | (0.003) | (0.007) | (0.005) | (0.010) | (0.016) |
| Age: 25–34 | 0.375 | 0.412 | −0.013 | −0.004 | 0.0008 | 0.002 | 0.004 | −0.035 |
| | (0.279) | (0.427) | (0.014) | (0.005) | (0.008) | (0.006) | (0.009) | (0.022) |
| Age: 35–49 | 1.82*** | 1.95*** | 0.006 | 0.003 | 0.018** | 0.012* | 0.016 | −0.032 |
| | (0.315) | (0.480) | (0.014) | (0.006) | (0.008) | (0.006) | (0.010) | (0.020) |
| Age: 50–64 | 1.92*** | 2.12*** | 0.0004 | −0.003 | 0.026*** | 0.015** | 0.009 | −0.071*** |
| | (0.312) | (0.431) | (0.013) | (0.005) | (0.008) | (0.007) | (0.009) | (0.019) |
| Age: 65+ | 2.81*** | 3.07*** | 0.006 | −0.005 | 0.035*** | 0.014** | 0.033*** | −0.066*** |
| | (0.318) | (0.449) | (0.013) | (0.005) | (0.009) | (0.006) | (0.009) | (0.019) |
| Constant | 3.27*** | 4.20*** | 0.085*** | 0.014*** | 0.021*** | 0.020*** | 0.036*** | 0.587*** |
| | (0.264) | (0.400) | (0.014) | (0.005) | (0.007) | (0.006) | (0.009) | (0.019) |
| Dependent variable mean | 5.0 | 6.1 | 0.08 | 0.010 | 0.04 | 0.04 | 0.06 | 0.55 |
| $R^2$ | 0.07 | 0.05 | 0.01 | 0.01 | 0.03 | 0.02 | 0.03 | 0.03 |
| Observations | 1,134 | 1,134 | 1,134 | 1,134 | 1,134 | 1,134 | 1,134 | 1,134 |

*Note:* Each column reports coefficients from estimating Equation (6), where the outcome is the exposure rate (Equation (2)) to the seven tracking mechanisms and the share of users' visits tracked by the top organization in column (8), defined as the tracker organization associated with the highest share of a user's total web visits (Section 2.4). Please see Figure C.2 for an alternative visualization of the estimates. Significance levels: * 0.1 ** 0.05 *** 0.01.

In contrast, differences by education and age are more pronounced. College-educated users encounter more trackers than those with a high school diploma or less. For instance, they encounter 16,090 more third-party cookies ($p < .01$) and 88 more session recorders ($\widehat{\text{SE}} = 35.9$, $p < .05$). Users with a postgraduate degree show similar patterns to those with a college degree. Age also plays a large role: older users are most exposed, with those 65 and above encountering significantly more trackers across all tracking methods, except for Google Analytics.

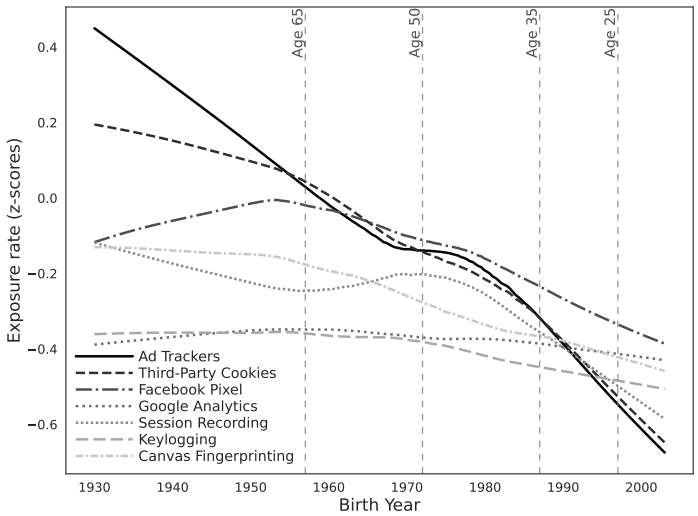



**Figure 2.** Exposure rate by birth year. Lines represent LOWESS-smoothed standardized rates (z-scores) of the exposure rates by the seven tracking methods. Values are winsorized at the 95th percentile. Vertical dashed lines correspond to the age groups.

Adjusting for browsing volume suggests that some demographic differences in tracking exposure reflect how much people browse, not which sites they visit (see Table 5). For example, the large gaps by education mostly vanish after normalization, suggesting that more educated users are online more often—not browsing more heavily tracked sites.

Some differences, however, remain. The gender gap in canvas fingerprinting remains: women encounter one additional fingerprinting script per 100 visits ($\widehat{\text{SE}} = 0.005$, $p < .05$). Age gradients in exposure also remain. Older users—especially those 65 and above—continue to experience higher exposure rates to ad trackers, third-party cookies, session recording, keylogging, and canvas fingerprinting (see Figure 2).[2]

Some differences sharpen after normalization. Asian users, who had lower cumulative exposure only to session recording and keylogging, now show lower exposure rates across nearly every method but Google Analytics. This suggests that, once online activity is held constant, they tend to visit less heavily tracked sites.

[2] Many demographic differences in exposure rates are significant even after correcting for multiple comparisons. Applying a Bonferroni correction for the 12 demographic predictors tested ($p < .00416$), all coefficients with unadjusted $p < .01$ in Table 5 remain significant.

Taken together, these results help pinpoint the sources of demographic gaps in tracking. Some reflect how often people go online; others reflect where they go. However, it is important to note that demographics explain little on their own: across all models, they account for less than 8% of the variation in exposure Tables 4 to 5), pointing to the dominant role of individual browsing habits.

## 3.3 Tracking by Organizations

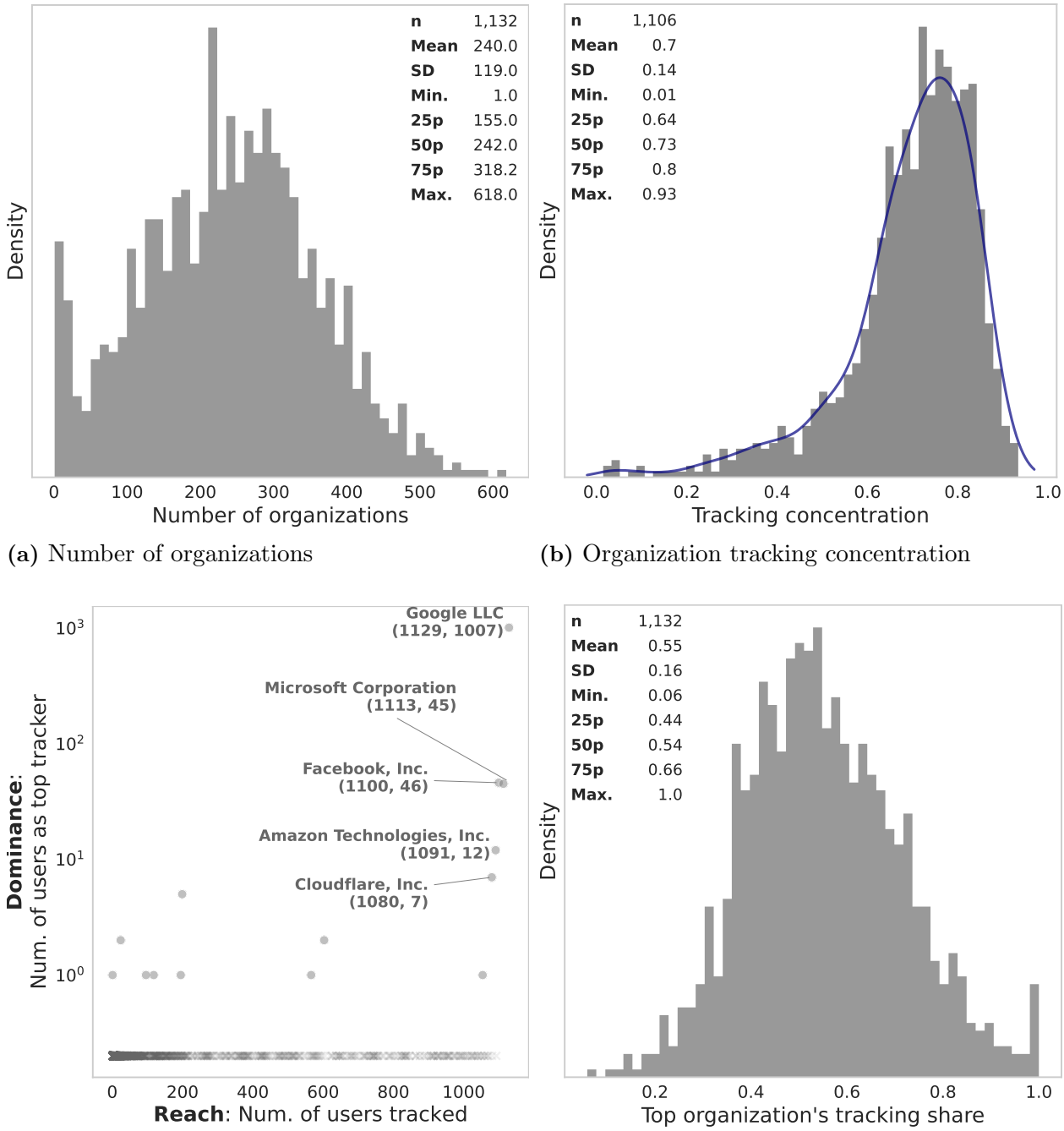
Mapping third-party services to parent organizations, we assess both the number of organizations tracking each user (Equation (3)) and the share of users' browsing histories tracked by each organization (Equation (4)).

Figure 3a shows that users are typically tracked by 155 to 318 organizations, with a median of 242. Despite this breadth, exposure is highly concentrated. Figure 3b shows that for users tracked by at least ten organizations, exposure is dominated by a handful of organizations, with the median Gini coefficient of 0.73.

Figure 3c plots organizations' tracking *dominance*—the number of users for whom it has the largest share of browsing history—against tracking *reach*—the number of users it tracks at least once, highlighting organizations with near-ubiquitous presence. Google towers over all in both reach and dominance, being the top organization for 99.6% of the sample. Other prominent organizations are Microsoft, Facebook, Amazon, and Cloudflare.[3]

Figure 3d shows the distribution of the maximum share of browsing history of a user tracked by an organization. On average, 55% of a user's browsing history is tracked by a single organization ($\hat{\sigma} = 0.16$). The median user has similar exposure, with 54% of their browsing history tracked by any single organization. At the 75th percentile, the top

[3]Likewise, tracking exposure is highly concentrated among a handful of domains (Appendix E), with many sites embedding multiple types of tracking technologies. Financial and e-commerce platforms are particularly prominent in contributing to the tracking via session recording and keylogging, while other big tech and social media companies, such as Microsoft and TikTok, are prominent in canvas fingerprinting.




**(a)** Number of organizations

**(b)** Organization tracking concentration

**(c)** Dominance vs. Reach

**(d)** User-level browsing history tracked by the top organization

**Figure 3.** The share of browsing history tracked by parent organizations. Panel (a) reports the number of organizations tracking each user's browsing history. Panel (b) reports the concentration of users' browsing history exposure across organizations (Gini coefficients, for those tracked by $\geq 10$ organizations). Panel (c) plots each organization's dominance—the number of users for whom it tracked the largest share of browsing history—against its Reach, the total number of users it tracked. The parentheses report the corresponding numbers. Panel (d) reports the largest share of each user's browsing history tracked by a single organization.

organization's share is 66%. Defining organizations' tracking share using the time spent online (Equation (5)) yields similar measures (Appendix D).

## 3.4 Demographics Differences in Tracking by Organizations

Lastly, we consider how the share of a user's browsing activity visible to the single most dominant tracking organization varies by demographics.

Whereas Section 3.2 examines demographic differences exposure to the seven tracking technologies detected by Blacklight, here we examine demographic differences in (i) the cumulative share of total visits observed by the top organization (column (8), Table 4) and (ii) the rate-normalized proportion of total visits observed by the top organization (column (8), Table 5).

Women have a slightly lower depth of exposure than men, while those with a college degree or postgraduate education have a greater depth of exposure compared to those with a high school diploma or below. These differences hold even when normalized by total visits (see column (8) of Table 5). Women have a 1.7 percentage point lower maximum share of visits ($\widehat{\text{SE}} = 1.0\%$, $p < .1$), while college-educated and postgraduate users have 3.9 (SE = 1.3%, $p < .01$) and 5.4 ($\widehat{\text{SE}} = 1.6\%$, $p < .01$) percentage point higher shares, respectively (column (8) of Table 5).

Interestingly, for age, the coefficients flip between the cumulative and rate (column (8) of Table 4). Older users (65+) have more of their visits tracked overall than younger users (18–24), according to the cumulative measure (column (8), Table 4). But when we look at the share of visits tracked, older users (50+) are less exposed than younger users—by at least 6.6 percentage points ($p < .01$). The difference reflects differences in browsing patterns by age.[4] These findings reinforce the theme in Section 3.2, where nearly all users are tracked

[4] As with Section 3.2, the demographic differences in the depth of tracking by organizations for education levels and age groups persist after correcting for multiple demographic tests (see Footnote 2).

online, but the intensity and structure of that tracking vary systematically by demographic characteristics. The depth of tracking by big organizations (e.g., Google, Microsoft, Facebook) reflects not just differences in online behavior but also deeper patterns of the digital gap.

# 4 Discussion

By linking digital traces from a representative sample of American adults with domain-level tracking audits, this study estimates individuals' exposure to online tracking. It also identifies who collects this information and how much of a user's web activity they can observe. The analysis advances the literature on online privacy in several ways.

First, unlike prior research that largely focused on audits of the most visited or most prominent websites (Englehardt and Narayanan, 2016; Karaj et al., 2019; Mattu and Sankin, 2020; Niforatos, Zheutlin and Sussman, 2021; Sanchez-Rola and Santos, 2018; Sanchez-Rola et al., 2021; Zheutlin, Niforatos and Sussman, 2022*b*,*a*), this study leverages passively observed browsing data from a large, representative sample. This allows for a more accurate estimate of actual user-level tracking exposure across the population (Dambra et al., 2022). Dambra et al. (2022) take a foundational step toward user-centric measurement by combining antivirus telemetry with custom web crawls, finding that user-level exposure is more concentrated than that measured from the trackers' perspective. Our study complements and extends this approach by linking domain-level tracking data—covering a wide range of tracking technologies—to observed browsing behavior from a representative panel of American adults with demographic data, allowing us to examine unequal exposure by demographics.

Second, the findings confirm that tracking on the web is nearly universal. Virtually all users in the sample encountered ad trackers and third-party cookies, with a median exposure in the tens of thousands. Like Dambra et al. (2022), we find that these encounters

occur rapidly. Most users were exposed to these trackers within the first 48 hours of the month-long observation period. We further show that more invasive technologies—such as session recording, keylogging, and canvas fingerprinting—appear less frequently but are still widespread, with over 40% of users encountering each of them within the first two days.

Third, exposure is not evenly distributed across the population. Users with more formal education, for instance, tend to experience higher levels of tracking. However, much of this disparity is explained by differences in browsing intensity. When exposure is normalized by the number of visits, demographic differences attenuate substantially, suggesting that more educated users are tracked more in part because they are online more often.

Yet, not all disparities vanish after accounting for browsing volume. In particular, older users consistently exhibit higher exposure rates per visit. This suggests that differences in exposure are not solely driven by time spent online, but also by the types of websites visited and the trackers embedded within them.

Despite these patterns, demographics explain only a small share of the variation in tracking exposure. Across both cumulative and normalized measures, the explanatory power of demographic variables is limited, with R-squared values of less than 8 percent in all specifications.

Finally, we examine the concentration of tracking across organizations. Although users may encounter hundreds of trackers, exposure is highly concentrated. We identify the same top three tracking organizations as Dambra et al. (2022), which analyzes the top tracking organizations by aggregating over all visits. As with Dambra et al. (2022), we find Google the most pervasive. Our estimates indicate that Google alone captures the largest share of browsing history for nearly 90% of users, with a median share of 54% of visits. The next closest organizations—Microsoft and Facebook—are the dominant trackers for only about 4% of users each, underscoring the extent to which a few firms dominate the tracking ecosystem. Our analysis of organization tracking aggregates across users, identifying the

single organization that observes the largest share of their browsing activity. This user-level measure of organizational dominance further allows us to examine how concentration varies across demographic groups, revealing, for instance, that younger users have a higher proportion of their browsing history visible to a single organization.

Several limitations of the study warrant discussion. First, while the digital traces include activity from mobile phones, they do not cover the tracking ecosystems within mobile applications, which often rely on embedded software development kits (SDKs) not detectable via browser-based methods (Achara, Acs and Castelluccia, 2015; Binns et al., 2018).

Second, the tracking audit tool, Blacklight, analyzes domains in real-time but has important blind spots. It does not detect more obfuscated forms of tracking, such as CNAME cloaking, nor does it capture server-side tracking that occurs outside the browser—even when users block cookies. Moreover, Blacklight focuses exclusively on client-side methods and may miss less visible forms of tracking. It also does not differentiate between benign and potentially harmful tracking; for example, session recording or canvas fingerprinting may be used for bot detection or UX testing, not necessarily surveillance (Mattu and Sankin, 2020; Senol et al., 2022).

Third, tracking audits were successful for only about half of the visited domains. These successfully scanned domains account for more than 75% of total visits, suggesting that failed scans occurred on less-visited sites. Additionally, a small subset of participants had no recorded web activity during the study period and were excluded from the analysis. In both cases, we assume that the missingness is unrelated to tracking exposure. While the high coverage of visits and low participant attrition reduce this concern, the possibility remains that tracking patterns differ systematically in the unobserved cases.

Fourth, the data rely on passive metering, and users' awareness of being observed—despite consenting to monitoring—may suppress true behavior. This could lead to an underestimation of actual tracking exposure, making our estimates conservative lower bounds (Bosch

et al., 2024; Penney, 2016; Shen and Sood, 2025; Sood and Shen, 2024).

Finally, our exposure measures reflect potential visibility to third-party organizations, not confirmed data transfers or behavioral profiling, though the presence of trackers is widely used as a proxy for privacy risk (Dambra et al., 2022; Karaj et al., 2019; Mattu and Sankin, 2020; Niforatos, Zheutlin and Sussman, 2021; Zheutlin, Niforatos and Sussman, 2022*b*,*a*).

# References


Acar, Gunes, Christian Eubank, Steven Englehardt, Marc Juarez, Arvind Narayanan and Claudia Diaz. 2014. The Web Never Forgets: Persistent Tracking Mechanisms in the Wild. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*. CCS '14 New York, NY, USA: Association for Computing Machinery p. 674–689.
**URL:** https://doi.org/10.1145/2660267.2660347

Acar, Gunes, Marc Juarez, Nick Nikiforakis, Claudia Diaz, Seda Gürses, Frank Piessens and Bart Preneel. 2013. FPDetective: dusting the web for fingerprinters. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security*. CCS '13 New York, NY, USA: Association for Computing Machinery p. 1129–1140.
**URL:** *https://doi.org/10.1145/2508859.2516674*

Achara, Jagdish Prasad, Gergely Acs and Claude Castelluccia. 2015. On the unicity of smartphone applications. In *Proceedings of the 14th ACM Workshop on Privacy in the Electronic Society*. pp. 27–36.

Angwin, Julia, Ariana Tobin and Madeleine Varner. 2016. "Facebook Lets Advertisers Exclude Users by Race." *ProPublica* . Accessed: April 2, 2025.
**URL:** *https://www.propublica.org/article/facebook-advertising-discrimination-housing-race-sex-national-origin*

Binns, Reuben, Ulrik Lyngs, Max Van Kleek, Jun Zhao, Timothy Libert and Nigel Shadbolt. 2018. Third Party Tracking in the Mobile Ecosystem. In *Proceedings of the 10th ACM Conference on Web Science*. WebSci '18 New York, NY, USA: Association for Computing Machinery p. 23–31.
**URL:** https://doi.org/10.1145/3201064.3201089

Borgesius, Frederik Zuiderveen. 2020. "Price discrimination, algorithmic decision-making, and European non-discrimination law." *European Business Law Review* 31(3).

Bosch, Oriol J., Patrick Sturgis, Jouni Kuha and Melanie Revilla and. 2024. "Uncovering Digital Trace Data Biases: Tracking Undercoverage in Web Tracking Data." *Communication Methods and Measures* 0(0):1–21.

Bujlow, Tomasz, Valentín Carela-Español, Josep Solé-Pareta and Pere Barlet-Ros. 2015. "Web tracking: Mechanisms, implications, and defenses." *arXiv preprint arXiv:1507.07872* .

Christl, Wolfie and Sarah Spiekermann. 2016. "Networks of control." *A report on corporate surveillance, digital tracking, big data & privacy Facultas* .

Citron, Danielle Keats and Daniel J Solove. 2022. "Privacy harms." *BUL Rev.* 102:793.

Dambra, Savino, Iskander Sanchez-Rola, Leyla Bilge and Davide Balzarotti. 2022. When Sally Met Trackers: Web Tracking From the Users' Perspective. In *31st USENIX Security Symposium (USENIX Security 22)*. Boston, MA: USENIX Association pp. 2189–2206.
**URL:** https://www.usenix.org/conference/usenixsecurity22/presentation/dambra

Englehardt, Steven and Arvind Narayanan. 2016. Online Tracking: A 1-million-site Measurement and Analysis. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. CCS '16 New York, NY, USA: Association for Computing Machinery p. 1388–1401.
**URL:** https://doi.org/10.1145/2976749.2978313

Hannak, Aniko, Gary Soeller, David Lazer, Alan Mislove and Christo Wilson. 2014. Measuring price discrimination and steering on e-commerce web sites. In *Proceedings of the 2014 conference on internet measurement conference*. pp. 305–318.

Iqbal, Umar, Steven Englehardt and Zubair Shafiq. 2021. Fingerprinting the Fingerprinters: Learning to Detect Browser Fingerprinting Behaviors. In *2021 IEEE Symposium on Security and Privacy (SP)*. pp. 1143–1161.

Johnson, Garrett A, Scott K Shriver and Shaoyin Du. 2020. "Consumer privacy choice in online advertising: Who opts out and at what cost to industry?" *Marketing Science* 39(1):33–51.

Karaj, Arjaldo, Sam Macbeth, Rémi Berson and Josep M. Pujol. 2019. "WhoTracks.Me: Shedding light on the opaque world of online tracking.".
**URL:** https://arxiv.org/abs/1804.08959

Mattu, Surya and Aaron Sankin. 2020. "How We Built a Real-Time Privacy Inspector.". Accessed: 2025-02-25.
**URL:** https://themarkup.org/blacklight/2020/09/22/how-we-built-a-real-time-privacy-inspector

Mowery, Keaton and Hovav Shacham. 2012. Pixel Perfect: Fingerprinting Canvas in HTML5. In *Proceedings of W2SP 2012*, ed. Matt Fredrikson. IEEE Computer Society.

Niforatos, Joshua D, Alexander R Zheutlin and Jeremy B Sussman. 2021. "Prevalence of third-party data tracking by US hospital websites." *JAMA Network Open* 4(9):e2126121–e2126121.

Nikiforakis, Nick, Alexandros Kapravelos, Wouter Joosen, Christopher Kruegel, Frank Piessens and Giovanni Vigna. 2013. Cookieless Monster: Exploring the Ecosystem of Web-Based Device Fingerprinting. In *2013 IEEE Symposium on Security and Privacy*. pp. 541–555.

Penney, Jonathon W. 2016. "Chilling effects: Online surveillance and Wikipedia use." *Berkeley Tech. LJ* 31:117.

Ravichandran, Deepak and Nitish Korula. 2019. "Effect of disabling third-party cookies on publisher revenue." *Google Report* .

Rivers, Douglas and Delia Bailey. 2009. Inference from matched samples in the 2008 US national elections. In *Proceedings of the Joint Statistical Meetings*. pp. 627–639. www.asasrms.org/Proceedings/y2009/Files/303309.pdf.

Sanchez-Rola, Iskander and Igor Santos. 2018. Knockin' on Trackers' Door: Large-Scale Automatic Analysis of Web Tracking. In *Detection of Intrusions and Malware, and Vulnerability Assessment*, ed. Cristiano Giuffrida, Sébastien Bardin and Gregory Blanc. Cham: Springer International Publishing pp. 281–302.

Sanchez-Rola, Iskander, Matteo Dell'Amico, Davide Balzarotti, Pierre-Antoine Vervier and Leyla Bilge. 2021. Journey to the Center of the Cookie Ecosystem: Unraveling Actors' Roles and Relationships. In *2021 IEEE Symposium on Security and Privacy (SP)*. pp. 1990–2004.

Senol, Asuman, Gunes Acar, Mathias Humbert and Frederik Zuiderveen Borgesius. 2022. Leaky Forms: A Study of Email and Password Exfiltration Before Form Submission. In *31st USENIX Security Symposium (USENIX Security 22)*. Boston, MA: USENIX Association pp. 1813–1830.
**URL:** *https://www.usenix.org/conference/usenixsecurity22/presentation/senol*

Shen, Lucas and Gaurav Sood. 2025. "Bad Domains: Exposure to Malicious Content Online.".
**URL:** *https://github.com/themains/bad_domains*

Solomos, Konstantinos, Panagiotis Ilia, Sotiris Ioannidis and Nicolas Kourtellis. 2020. "Clash of the Trackers: Measuring the Evolution of the Online Tracking Ecosystem.".
**URL:** *https://arxiv.org/abs/1907.12860*

Sood, Gaurav. 2022. "YouGov Pulse Data for 1200 people for June 2022.". DOI: 10.7910/DVN/VIV4TS.

Sood, Gaurav and Lucas Shen. 2024. "Holier Than Thou? No Large Partisan Gaps in the Consumption of Pornography Online." *Journal of Quantitative Description: Digital Media* 4:n/a. DOI: 10.51685/jqd.2024.011.

Zheutlin, Alexander R., Joshua D. Niforatos and Jeremy B. Sussman. 2022*a*. "Data-Tracking Among Digital Pharmacies." *Annals of Pharmacotherapy* 56(8):958–962. PMID: 34978215.
**URL:** *https://doi.org/10.1177/10600280211061757*

Zheutlin, Alexander R., Joshua D. Niforatos and Jeremy B. Sussman. 2022*b*. "Data-Tracking on Government, Non-profit, and Commercial Health-Related Websites." *Journal of General Internal Medicine* 37(5):1315–1317.
**URL:** https://doi.org/10.1007/s11606-021-06695-8

# Supporting Information

## A Participant consent and data privacy

Before enrolling in a YouGov panel, people receive detailed information about the nature and scope of the data collection. Potential participants are informed about the types of data that will be collected, such as visited domains, and what will not be collected, including any information entered into secure forms, such as usernames, passwords, or payment details (See YouGov's FAQ, https://today.yougov.com/about/faq).

Only after reviewing this information do individuals consent to participate. Participation is entirely voluntary, and panelists can pause or uninstall the tracking software at any time. (See pages 3 and 4 of the installation guide for Terms and Conditions and Privacy Policy made known to participants.)

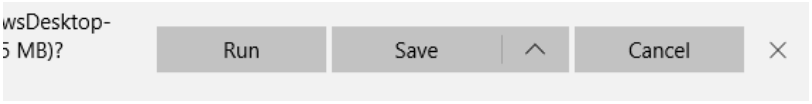
The browsing data collection application—YouGov Pulse—is developed in partnership with RealityMine and is available as a browser extension or mobile app. The app ensures anonymity: researchers never have access to identifying information, and no data is shared with third parties.

To encourage participation, panelists earn points through YouGov's reward system– 2,000 points upon joining and an additional 1,000 points for completing a full month of activity.

## Installing YouGov Pulse on Windows

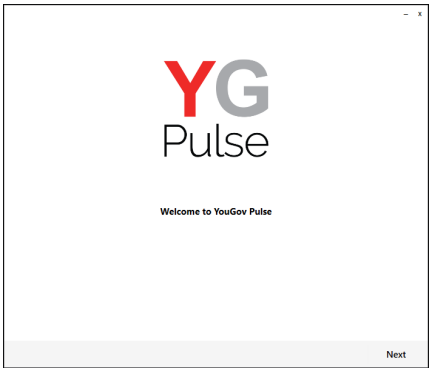
**Step 1**

Open the link provided from either the survey or the email. Download the software and click “Run” (depending on your browser, this might look slightly different) or open the file.



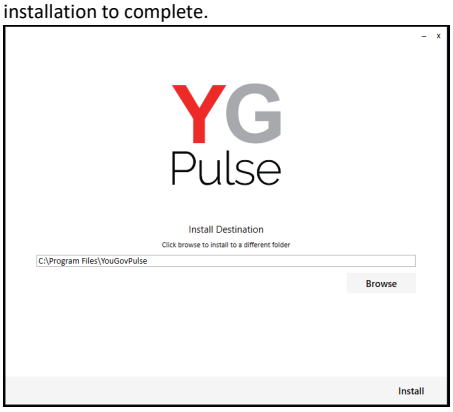

**Step 2**

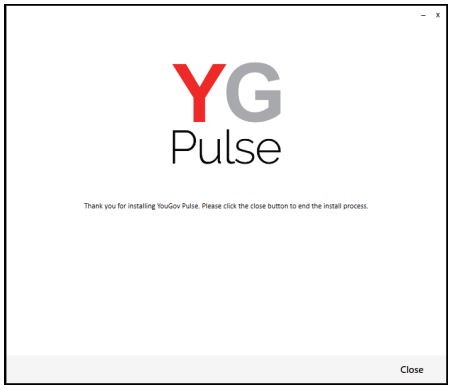
Start the installation process and click “Next”




**Step 3**

Choose installation destination and click “Install”. Accept any prompts from Windows to allow installation to complete.



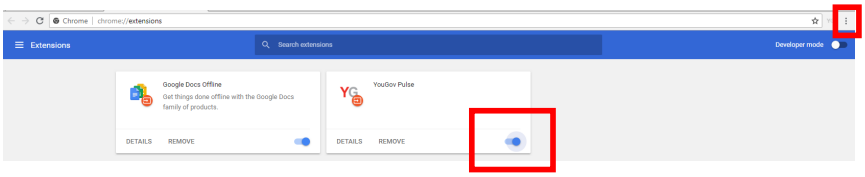



**Step 4**

Install the Google Chrome and/or Mozilla Firefox browser extension(s) by clicking 'OK' on the box that pops up – if either of them are open (Note: The browser will close.)

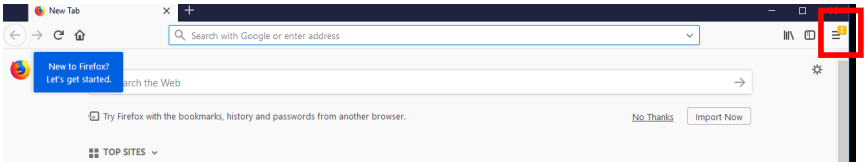
If the browsers are not open, you will not see the message box. Instead, you will notice that the extension has been added next time you open the browser.

***On Chrome-***
Either after you have clicked 'OK' or next time you open Google Chrome, click "Enable Extension" on the window in the top right. If you can't see this notification, click on the three dots next to the URL bar and select "More Tools > Extenstions". Ensure that the YouGovPulse Extension is enabled by moving the slider to the right if necessary:

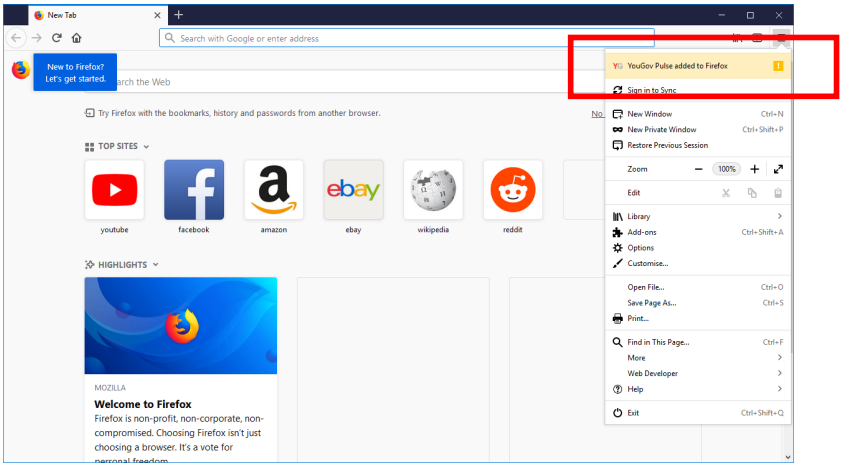



Once it is installed and enabled, you will see the YouGov Pulse icon to the right of the URL bar.

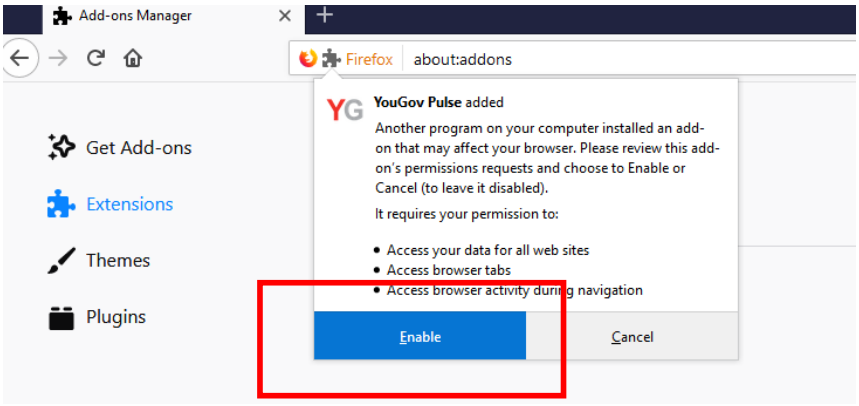
***On Firefox-***
Either after you clicked "OK" or next time you open Mozilla Firefox, click the yellow exclamation mark under the "Open Menu" icon:




Click on this, then select "YouGov Pulse added to Firefox"

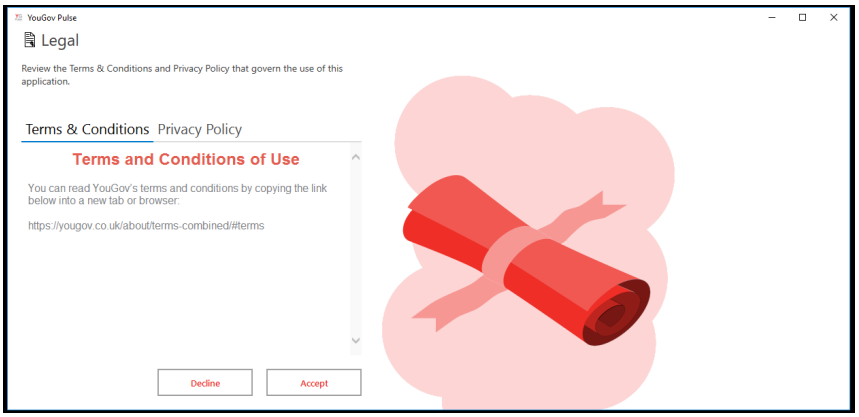


and “Enable” to activate the Add-On.



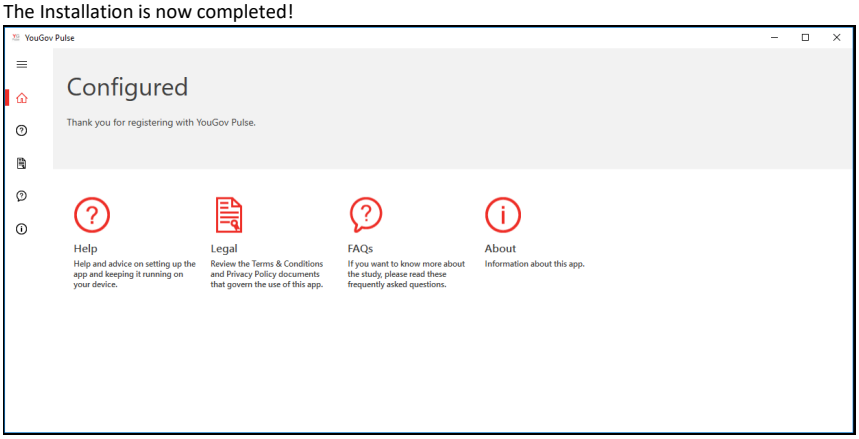

**Step 5**

Open the YouGov Pulse App, read the Terms and Condition and the Privacy Policy and select “Accept”.



**Completed!**

The Installation is now completed!




**IMPORTANT NOTE**: *Please make sure that the software is running in the background at all times. If the app stops running, you'll stop earning your points. You can delete the App at any time if you decide to no longer be part of the YouGov Pulse project.*

# B Tracking methods

This appendix summarizes the seven tracking methods that Blacklight detects on the homepage of the domain and one additional randomly selected internal page (Mattu and Sankin, 2020). Blacklight analyses are retrieved from a 24–48-hour cache when available or performed in real-time if no recent results exist.

- **Ad Tracking:** Ad trackers are third-party scripts embedded in websites that collect user browsing behavior and send it to advertising networks. These scripts help build user profiles for targeted advertising or retargeting across websites. Blacklight detects ad tracking by identifying network requests to known advertising domains (domains under "Ad Motivated Tracking", https://github.com/duckduckgo/tracker-radar/blob/main/docs/CATEGORIES.md) in the DuckDuckGo Tracker Radar list.

- **Third-party Cookies:** Cookies are small text files stored in the user's browser. Third-party cookies originate from domains other than the one being visited and are widely used to track users across websites.

- **Facebook Pixel:** Facebook Pixel is a tracking script that monitors user behavior–such as page views, button clicks, and purchases—and sends this data to Facebook for ad targeting and conversion analytics. It links off-site behavior to user profiles across the Facebook ecosystem, even if users are not logged in to Facebook. Blacklight detects Facebook Pixel by identifying network requests to Facebook domains and inspecting URL query parameters for data patterns that match Pixel's documented schema.

- **Google Analytics:** Another major tracking tool operated by a major tech company is Google Analytics, which uses JavaScript tags and cookies to monitor user behavior such as session duration, navigation, and referrals. Blacklight detects it by flagging requests to known Google Analytics endpoints, such as `http://stats.g.doubleclick.net`.

- **Session Recording:** Session replay scripts record user activity on a website, including mouse movements, scrolling, and form inputs—often in real time (Senol et al., 2022). These recordings can be replayed by website owners, revealing detailed behavioral data and potentially sensitive information. Blacklight detects session recording by monitoring network requests for URL substrings known to be associated with session

replay tools (https://web.archive.org/web/20210830151649/https://gist.github.com/gunesacar/0c67b94ad415841cf3be6761714147ca).

- **Keylogging:** A potentially more invasive subset of session recording, keylogging captures every keystroke a user makes—including input into masked fields like passwords and credit card forms—before submission. This technique can reveal highly sensitive user data. Blacklight enters pre-determined text into input fields and monitors network requests for the same outgoing data.

- **Canvas Fingerprinting:** This method leverages the HTML5 canvas element to render invisible graphics and analyze subtle rendering differences based on the user's hardware and software configuration (Acar et al., 2014; Mowery and Shacham, 2012). These differences can be used to create a persistent, stateless identifier for tracking users across sessions (Karaj et al., 2019; Mattu and Sankin, 2020). Blacklight infers that canvas fingerprinting is used for tracking if scripts silently draw meaningful content on a sufficiently large canvas, do not use it for interactivity, and then extract pixel-level data in a way consistent with generating unique user identifiers.

# C Alternative Visualization of Estimates

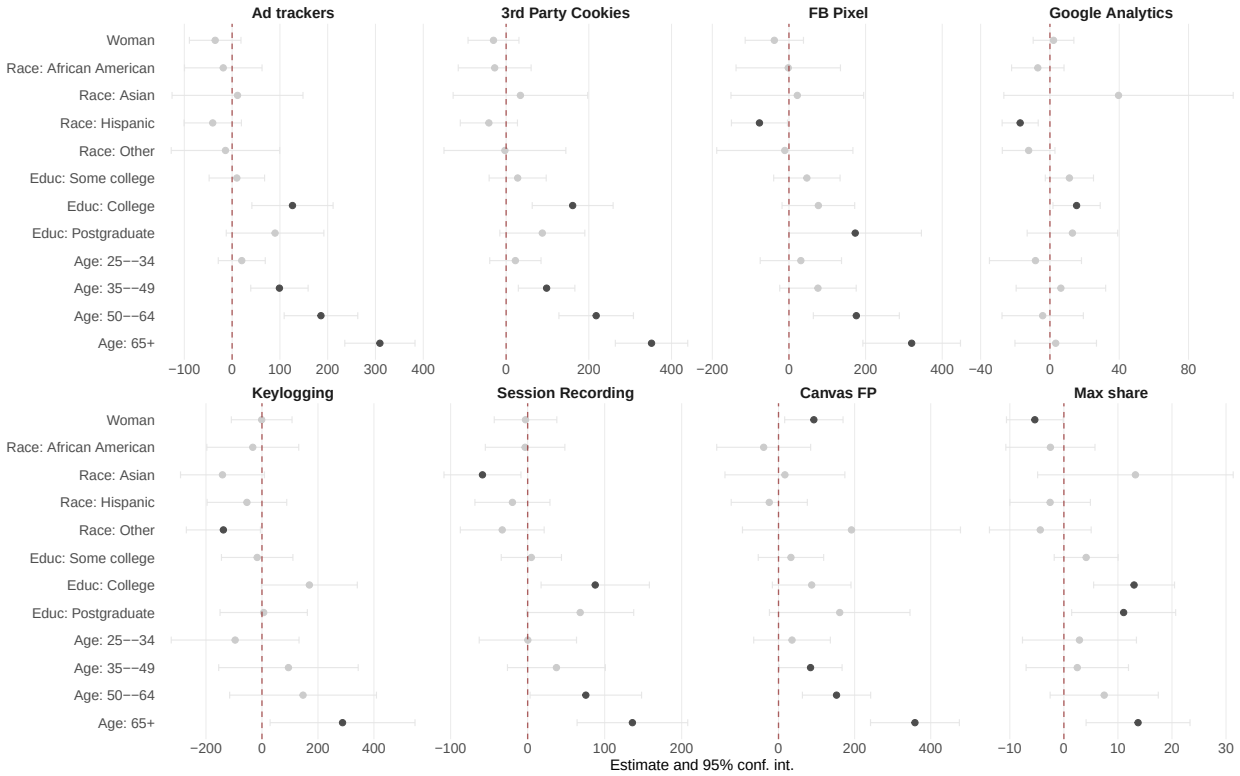



**Figure C.1.** Estimated coefficients in *cumulative exposure* by demographic group. Corresponds to Table 4. Each panel shows the estimated effect (makers) and 95% confidence intervals (horizontal lines) from OLS regressions associating exposure to one of the seven tracking methods and the total number of visits tracked by a single organization. Black markers indicate statistically significant estimates at $p < .05$; gray markers indicate non-significant estimates.

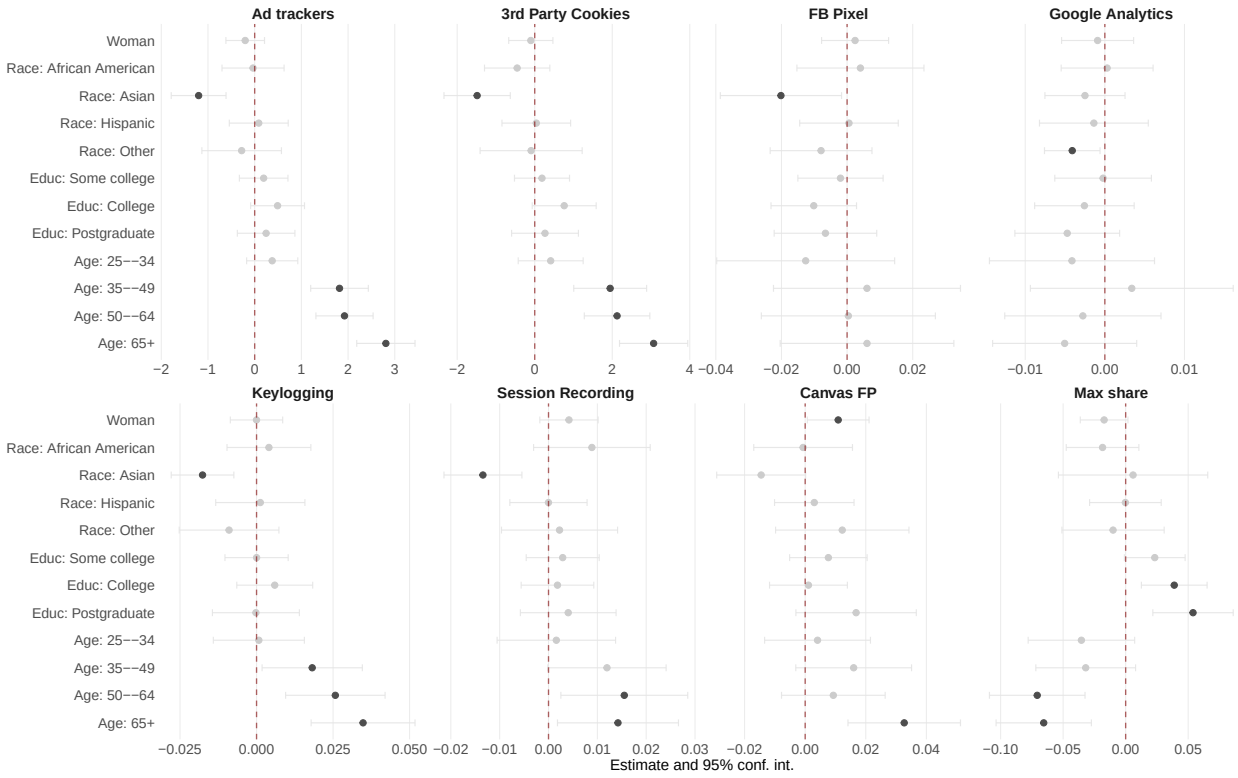



**Figure C.2.** Estimated coefficients in *exposure rate* by demographic group. Corresponds to Table 5. Each panel displays the OLS estimate and 95% confidence intervals for regressions of either the exposure rate to the tracking technology or the proportion of visits tracked by a single organization. Black markers indicate statistically significant estimates at $p < .05$; gray markers indicate non-significant estimates.

# D Organization tracking weighted by time

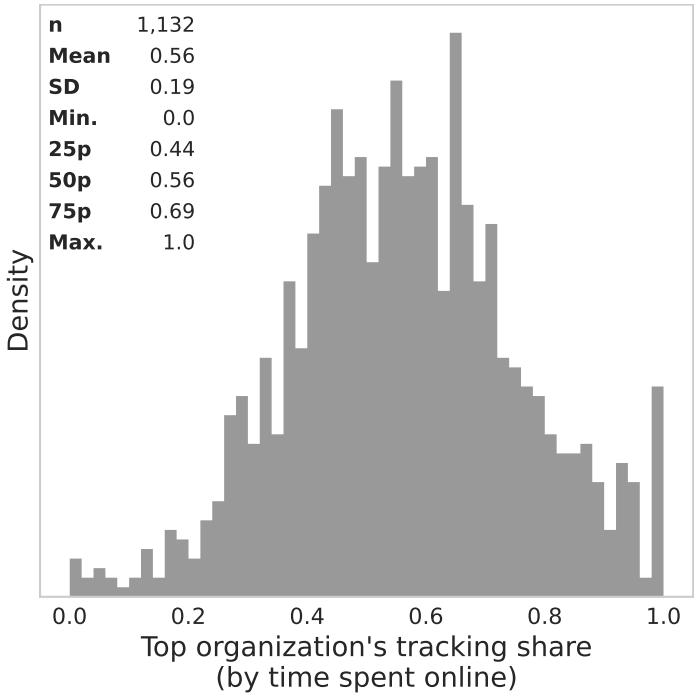



**Figure D.1.** The largest share of each user's browsing time online tracked by a single organization (Equation (5)):

$$\text{Tracking share}_{ij}^{(\text{dur})} = \frac{\sum_{v \in \mathcal{V}_i} \mathbf{1}(j \in O_{iv}) \cdot t_{iv}}{\sum_{v \in \mathcal{V}_i} t_{iv}}.$$

See Figure 3d for the corresponding figure for tracking shares by site visits.

# E Top Tracking Domains

**Table E.1.** Top domains contributing to exposure

| | Ads (1) | Cookies (2) | FB Pixel (3) | GA (4) | Session rec (5) | Keyloggers (6) | Canvas FP (7) |
|---|---|---|---|---|---|---|---|
| 1 | yahoo.com (246k) | yahoo.com (246k) | ebay.com (30k) | kohls.com (2.7k) | xfinity.com (10k) | yahoo.com (246k) | live.com (80k) |
| 2 | google.com (987k) | google.com (987k) | capitaloneshopping.com (23k) | force.com (2.1k) | capitalone.com (9.9k) | capitaloneshopping.com (23k) | microsoft.com (26k) |
| 3 | live.com (80k) | live.com (80k) | chase.com (14k) | pixiv.net (1.9k) | cbssports.com (6.2k) | smugmug.com (10k) | capitaloneshopping.com (23k) |
| 4 | aol.com (47k) | bing.com (236k) | rakuten.com (12k) | mheducation.com (1.4k) | dell.com (5.5k) | weather.com (3.8k) | linkedin.com (19k) |
| 5 | microsoft.com (26k) | microsoft.com (26k) | hulu.com (11k) | tupperware.com (1.4k) | att.com (4.9k) | activemeasure.com (3.6k) | rakuten.com (12k) |
| 6 | cbssports.com (6.2k) | cbssports.com (6.2k) | xfinity.com (10k) | thriftbooks.com (1.0k) | earthlink.net (4.1k) | venatusmedia.com (3.5k) | hulu.com (11k) |
| 7 | xfinity.com (10k) | xfinity.com (10k) | usps.com (9.7k) | adp.com (977) | venatusmedia.com (3.5k) | revenueuniverse.com (3.0k) | xfinity.com (10k) |
| 8 | youtube.com (233k) | msn.com (39k) | nielseniq.com (9.4k) | equitybank.com (888) | homedepot.com (3.0k) | doceree.com (2.9k) | tiktok.com (10.0k) |
| 9 | ebay.com (30k) | ebay.com (30k) | netflix.com (7.0k) | priceline.com (808) | doceree.com (2.9k) | spot.im (2.9k) | capitalone.com (9.9k) |
| 10 | imdb.com (7.5k) | weather.com (3.8k) | wellsfargo.com (6.8k) | webtoons.com (705) | kohls.com (2.7k) | yelp.com (2.3k) | washingtonpost.com (8.0k) |
| 11 | washingtonpost.com (8.0k) | dynata.com (22k) | dell.com (5.5k) | ourfamilywizard.com (614) | ancestry.com (2.6k) | attn.tv (2.2k) | espn.com (6.7k) |
| 12 | rakuten.com (12k) | imdb.com (7.5k) | nextdoor.com (5.1k) | coupons.com (597) | discover.com (2.5k) | westlaw.com (2.1k) | target.com (5.9k) |
| 13 | cnn.com (4.4k) | nielseniq.com (9.4k) | iheart.com (5.1k) | yaysavings.com (574) | zoosk.com (2.4k) | kroger.com (2.0k) | bankofamerica.com (5.7k) |
| 14 | weather.com (3.8k) | cnn.com (4.4k) | 9gag.com (4.8k) | meetup.com (570) | attn.tv (2.2k) | ex.co (1.8k) | dell.com (5.5k) |
| 15 | usps.com (9.7k) | youtube.com (233k) | earthlink.net (4.1k) | narvar.com (560) | cmix.com (2.0k) | dropbox.com (1.8k) | biggerbooks.com (4.0k) |
| 16 | 9gag.com (4.8k) | twitter.com (111k) | biggerbooks.com (4.0k) | overdrive.com (557) | prizerebel.com (2.0k) | pnc.com (1.5k) | citi.com (3.9k) |
| 17 | nielseniq.com (9.4k) | nytimes.com (6.0k) | activemeasure.com (3.6k) | managebuilding.com (529) | zleague.gg (1.9k) | morningjournal.com (1.1k) | cbsi.com (3.3k) |
| 18 | nytimes.com (6.0k) | centurylink.net (1.8k) | venatusmedia.com (3.5k) | wootric.com (511) | trendmicro.com (1.9k) | 53.com (1.0k) | homedepot.com (3.0k) |
| 19 | hulu.com (11k) | kohls.com (2.7k) | productreportcard.com (3.4k) | evergage.com (479) | phoenix.edu (1.9k) | thriftbooks.com (1.0k) | samsclub.com (2.8k) |
| 20 | iheart.com (5.1k) | civicscience.com (7.4k) | cbsi.com (3.3k) | udemy.com (474) | verizon.com (1.8k) | trulia.com (995) | zulily.com (2.8k) |
| 21 | kohls.com (2.7k) | dell.com (5.5k) | ups.com (3.2k) | fox.com (339) | jcpenney.com (1.5k) | qvc.com (991) | kohls.com (2.7k) |
| 22 | foxnews.com (3.5k) | foxnews.com (3.5k) | homedepot.com (3.0k) | hobbylobby.com (311) | tupperware.com (1.4k) | dynatrace.com (922) | discover.com (2.5k) |
| 23 | capitaloneshopping.com (23k) | aol.com (47k) | honeygain.com (2.9k) | daisous.com (309) | wurflcloud.com (1.4k) | newspapers.com (892) | adobe.com (2.4k) |
| 24 | chase.com (14k) | rakuten.com (12k) | spot.im (2.9k) | wgal.com (308) | playsugarhouse.com (1.2k) | kaiserpermanente.org (890) | kroger.com (2.0k) |
| 25 | dell.com (5.5k) | 9gag.com (4.8k) | samsclub.com (2.8k) | epsilon.com (292) | copart.com (1.1k) | mapquest.com (819) | shein.com (2.0k) |
| 26 | dynata.com (22k) | google.co.uk (18k) | airbnb.com (2.8k) | noom.com (284) | veritonic.com (1.1k) | upmc.com (769) | trendmicro.com (1.9k) |
| 27 | centurylink.net (1.8k) | chase.com (14k) | kohls.com (2.7k) | bizpacreview.com (274) | dominos.com (1.0k) | e-rewards.com (764) | aliexpress.com (1.8k) |
| 28 | espn.com (6.7k) | capitalone.com (9.9k) | ancestry.com (2.6k) | factor75.com (245) | emi-rs.com (1.0k) | offerup.com (726) | pnc.com (1.5k) |
| 29 | msn.com (39k) | morningjournal.com (1.1k) | discover.com (2.5k) | tbdliquids.com (234) | slickdeals.net (998) | odysee.com (700) | jcpenney.com (1.5k) |
| 30 | linkedin.com (19k) | linkedin.com (19k) | adobe.com (2.4k) | reverbnation.com (224) | fidelity.com (975) | vccs.edu (653) | navyfederal.org (1.4k) |
| 31 | twitter.com (111k) | nascar.com (903) | zoosk.com (2.4k) | avant.com (223) | newspapers.com (892) | forter.com (571) | coursera.org (1.4k) |
| 32 | democraticunderground.com (14k) | spot.im (2.9k) | duolingo.com (2.4k) | mtsac.edu (218) | kaiserpermanente.org (890) | meetup.com (570) | ea.com (1.4k) |
| 33 | zillow.com (19k) | investing.com (839) | vidyard.com (2.3k) | njlottery.com (211) | etrade.com (889) | blueconic.net (418) | nordstrom.com (1.3k) |
| 34 | navyfederal.org (1.4k) | adobe.com (2.4k) | experian.com (2.2k) | examfx.com (198) | equitybank.com (888) | sutherlandglobal.com (403) | newyorklife.com (1.3k) |
| 35 | oregonlive.com (1.4k) | zoho.com (15k) | attn.tv (2.2k) | gerberlife.com (197) | grabpoints.com (882) | reserveohio.com (399) | hp.com (1.2k) |
| 36 | hideout.co (11k) | venatusmedia.com (3.5k) | kroger.com (2.0k) | higherincomejobs.com (193) | bhg.com (841) | bandcamp.com (375) | pusherapp.com (1.2k) |
| 37 | zoosk.com (2.4k) | navyfederal.org (1.4k) | prizerebel.com (2.0k) | clover.com (182) | investing.com (839) | netspend.com (361) | playsugarhouse.com (1.2k) |
| 38 | civicscience.com (7.4k) | huffpost.com (1.1k) | zleague.gg (1.9k) | onlygreatjobs.com (178) | gofundme.com (839) | freefarmtowngiftshop.com (339) | booking.com (1.1k) |
| 39 | huffpost.com (1.1k) | trendmicro.com (1.9k) | trendmicro.com (1.9k) | kmov.com (177) | mcafee.com (817) | connatix.com (329) | expedia.com (1.1k) |
| 40 | kitco.com (2.0k) | paycor.com (1.6k) | verizon.com (1.8k) | pushwoosh.com (172) | medallia.com (813) | hibid.com (329) | copart.com (1.1k) |
| 41 | adobe.com (2.4k) | iheart.com (5.1k) | ex.co (1.8k) | truegloryhair.com (164) | adidas.com (783) | hobbylobby.com (311) | truist.com (1.0k) |
| 42 | google.co.uk (18k) | cbsnews.com (865) | grizly.com (1.6k) | mintmobile.com (158) | chegg.com (767) | opentable.com (306) | 53.com (1.0k) |
| 43 | capitalone.com (9.9k) | office.com (18k) | paycor.com (1.6k) | quantilope.com (157) | opera.com (757) | twinspires.com (306) | slickdeals.net (998) |
| 44 | foodnetwork.com (1.0k) | attn.tv (2.2k) | allrecipes.com (1.6k) | walmart.com.mx (155) | wishpond.com (740) | ms.gov (296) | qvc.com (991) |
| 45 | reddit.com (61k) | vidyard.com (2.3k) | westernjournal.com (1.5k) | yummybazaar.com (153) | neu.edu (724) | partycentersoftware.com (294) | adp.com (977) |
| 46 | nascar.com (903) | bonvoyaged.com (751) | pnc.com (1.5k) | foxsports.com (148) | salemove.com (710) | pinnbank.com (259) | fidelity.com (975) |
| 47 | morningjournal.com (1.1k) | doceree.com (2.9k) | mheducation.com (1.4k) | everyplate.com (146) | adam4adamsfw.com (697) | eyebuydirect.com (241) | citibankonline.com (971) |
| 48 | ups.com (3.2k) | verizon.com (1.8k) | pandora.com (1.4k) | uscellular.com (144) | vergic.com (694) | centercode.com (236) | barclaycardus.com (928) |
| 49 | discover.com (2.5k) | wellsfargo.com (6.8k) | oregonlive.com (1.4k) | pubnub.com (135) | pearson.com (672) | edx.org (216) | npr.org (922) |
| 50 | westernjournal.com (1.5k) | meetup.com (570) | wurflcloud.com (1.4k) | guard.io (129) | oldnational.com (670) | chicoryapp.com (201) | michaels.com (900) |

*Note:* This table reports the top 50 domains (rows) contributing to individual-level exposure for each of the seven tracking methods (columns). A domain $d$'s contribution to individual-level exposure is computed as:

$$\text{Contribution}_d^{(s)} = \sum_i \sum_{v \in \mathcal{V}_{id}} |trackers_d^{(s)}|,$$

based on all individual-domain visit instances, weighted by the number of trackers of type $s$ present on domain $d$. Parentheses report the total number of visits.