# Stop tracking me Bro! Differential Tracking of User Demographics on Hyper-Partisan Websites

Pushkal Agarwal
King's College London, UK
pushkal.agarwal@kcl.ac.uk

Sagar Joglekar
King's College London, UK
sagar.joglekar@kcl.ac.uk

Panagiotis Papadopoulos
Brave Software Inc.
ppapadopoulos@brave.com

Nishanth Sastry
King's College London, UK
nishanth.sastry@kcl.ac.uk

Nicolas Kourtellis
Telefonica Research, Spain
nicolas.kourtellis@telefonica.com

## ABSTRACT

Websites with hyper-partisan, left or right-leaning focus offer content that is typically biased towards the expectations of their target audience. Such content often polarizes users, who are repeatedly primed to specific (extreme) content, usually reflecting hard party lines on political and socio-economic topics. Though this polarization has been extensively studied with respect to content, it is still unknown how it associates with the online tracking experienced by browsing users, especially when they exhibit certain demographic characteristics. For example, it is unclear how such websites enable the ad-ecosystem to track users based on their gender or age.

In this paper, we take a first step to shed light and measure such potential differences in tracking imposed on users when visiting specific party-line's websites. For this, we design and deploy a methodology to systematically probe such websites and measure differences in user tracking. This methodology allows us to create user personas with specific attributes like gender and age and automate their browsing behavior in a consistent and repeatable manner. Thus, we systematically study how personas are being tracked by these websites and their third parties, especially if they exhibit particular demographic properties. Overall, we test 9 personas on 556 hyper-partisan websites and find that right-leaning websites tend to track users more intensely than left-leaning, depending on user demographics, using both cookies and cookie synchronization methods and leading to more costly delivered ads.

## CCS CONCEPTS

• **Information systems** → **Online advertising**; **Web log analysis**; • **Social and professional topics** → User characteristics.

## KEYWORDS

hyper-partisan news, web cookies, advertisements, web tracking.

## 1 INTRODUCTION

In the era of mass Web monitoring, users are being tracked and their behavioral data collected and used, typically for *ad-targeting purposes*. Recent research has explored the impact of ad-tracking on user privacy [37, 39], and showed the scale and sophistication of the ad-ecosystem, and ad-pricing used during the real-time bidding protocol (RTB) [40] or header bidding protocol [35] to serve ad-slots to webpages. In fact, a deeper examination of the ad-ecosystem reveals a set of complex techniques such as synchronization of cookies across websites [2, 34, 39] and fingerprinting of user devices [33, 36], in order to perfect the user profiles being targeted. Indeed, a highly precise profile allows ad-platforms to effectively match ads with target audiences. However, the side effect of these highly sophisticated profiling techniques is a massive breach of user's privacy. Without any consent, a profiled user is unknowingly disclosing her interests and (dis)likes to the benefit of ad-platforms.

Recently, such user profiling techniques have also been used *for targeted opinion shaping* such as Cambridge Analytica's U.S. presidential campaigns for Donald Trump [41] and Ted Cruz [48], UK-based campaigns for the Brexit referendum [20], and Russia's social media operations during the 2016 U.S. presidential election [8]. All these campaigns have one common motif: use of fake and partisan disinformation campaigns. In fact, a slew of websites hosting such content have emerged since the start of Trump's campaign [9]. These pages allow fake and partisan news to be shared unchecked on the social media, garnering ever increasing partisan audiences. Some examples are Infowars, Breitbart and Fox News, for the right, and Occupy Democrats, Bipartisan Report and MSNBC for the left.

At such a juncture, we are compelled to ask: *Do these fake and Hyper-partisan news websites (HPWs) – which have been shown to have highly selective audiences – exhibit any particular differential behavior when it comes to tracking their online visitors?* i.e., are websites on the left (right), differentially tracking the left (right)-leaning visitors more than the right (left) leaning ones?

To answer this multi-faceted question, we first establish a methodology for understanding how HPWs and their embedded third parties track different user demographics. We create 9 carefully crafted personas representing different genders and age groups. We load browsers with these personas and visit a list of 556 verified HPWs [9], to observe differences in the way these personas are

being tracked via different types of cookies placed by HPWs and the third party ad-ecosystem.

We search for large-scale patterns by co-clustering both the personas and websites visited, using non-negative matrix factorization [27, 29, 57]. We examine the extent to which the ad-ecosystem performs unbalanced tracking and cookie synchronization [2, 39] on users of particular demographics, when a persona visits the right- or left-leaning HPWs. Finally, we intercept and analyze RTB-based ads placed when visiting HPWs with particular personas, and establish differences between tracking practices on right and left HPWs, as well as differences in the monetary value of placed ads.

With this study, we make the following main contributions:

(1) We design the first to our knowledge methodology to detect possible unbalanced tracking performed from HPWs and their collaborating third parties of users that belong to different demographics.

(2) We extend an existing crawling tool with our methodology, and open source our framework[1] in order to enable fellow researchers, policy makers or even end-users to audit websites on how they personalize tracking technology based on the visitor's web profile.

(3) Our results show that in HPWs, *advertisers set the majority of cookies* on users' browsers. *These advertisers are among the top in the overall ad-ecosystem*, and are over-represented on HPWs compared to their presence on the general web. In fact, some of the top trackers are twice as prevalent on HPWs as on the general web. Also, having an established persona from a particular demographic (with cookies obtained from visiting stereotypical websites for users of that demographic) results in *up to 15% more cookies* stored than for a baseline with no set persona. Furthermore, popular or highly-ranked HPWs track users more intensely than lower-ranked websites. More importantly, our results show that right-leaning websites, in general, track users with *up to 25% more cookies* than left-leaning websites. They also enable more data sharing with advertisers and trackers by facilitating *up to 50% more cookie synchronizations* between these third parties, than in left-leaning websites. This intense tracking leads *up to 5x higher cost ads* than left-leaning websites. We discuss implications of our work for online users' privacy.

## 2 OVERVIEW OF METHODOLOGY

### 2.1 Third party tracking on the Web

Websites nowadays consist of several components that may originate from many different domains other than the one the user visited. These components provide functionality like widgets, analytics, targeted ads, and recommendations. In order to provide as much personalized content as possible, these third parties keep track of user's personal information (e.g., geolocation, browsing device, gender) and preferences (e.g., purchases, searches, etc.) locally in a cookie placed in the user's browser, and in their servers' database using the same cookie ID.

Although several policies (e.g., Intelligent Tracking Prevention [53] and Same Origin Policy [54]) have been proposed to mitigate the

---

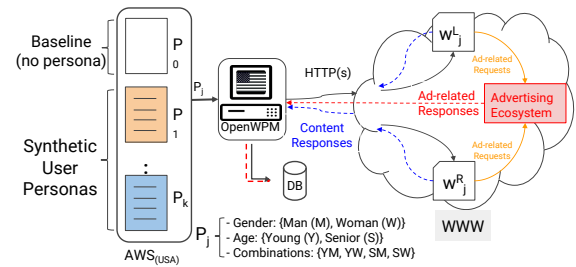[1]Data and code available from http://tiny.cc/partisan-tracking



**Figure 1: Crawling methodology and framework for measuring tracking of different personas by hyper-partisan websites visited, and third parties embedded in them.**

privacy intrusion from such pervasive tracking, there are several sophisticated techniques that allow trackers to bypass such mechanism (e.g., Cookie Synchronization [7, 39], Web Beacons [21], Cross-Device Tracking [46], etc.) and re-identify a user across different websites and create detailed profiles about her preferences and interests (e.g., sexual preferences, political beliefs, etc.) [24]. Then, this data can be sold to anyone interested [49], or handed over to agencies [4].

### 2.2 Overview of Crawling Methodology

A general illustration of our methodology and implemented framework is shown in Figure 1. To obtain consistent user behavior that allows repeatable and systematic web measurements, the proposed framework leverages a methodology with carefully curated user personas that have a history built on Web traffic that could be expected from users of specific demographic groups. Using each of these defined personas, we then visit HPW, which have been carefully categorized (manually) as left- or right-leaning by domain experts (journalists and fact checkers at Buzzfeed News [44, 45]). Then, we log and monitor first and third party (e.g., ad-ecosystem) tracking performed during these visits, for later analysis and comparison between personas and baseline or null users. We summarize various terms used in the paper in Table 1.

### 2.3 Incremental Persona Building

Our goal is to visit different websites with the "persona" of a user from a particular demographic group, based on age or gender. We build personas by visiting top Alexa ranking websites for different demographics. The intuition is that the third party ecosystem that enables tracking and personalized targeting [3, 6, 11] builds up a persona or profile of a user based on the type of websites they visit. Thus, by visiting websites which are highly popular in a particular demographic, we establish a persona of that demographic. Previous works have inferred that there are certain demographic and geographical features which are highly associated with a person's partisan lean [9], including age, gender and location (whether from a right- or left-leaning US state). Therefore, we build our personas

**Table 1: Terms and notations used in our methodology.**

| Terms | Notations |
|-------|-----------|
| General persona<br>Simple persona<br>Rich persona | $P_j$, where $P_0$ is baseline (i.e., does not have any profile or user history loaded),<br>$P_{1 \leq j \leq 4}$ are 1-feature personas,<br>$P_{5 \leq j \leq 8}$ are 2-feature, "rich history" personas |
| Hyper-partisan websites (Left)<br>Hyper-partisan websites (Right) | $W^{L_i}$ are left-leaning websites; $1 \leq i \leq 164$<br><br>$W^{R_i}$ are right-leaning websites; $165 \leq i \leq 556$ |
| Tracking Domains (Left)<br>Tracking Domains (Right) | $D^{L_i}$ are domains in all left-leaning websites; $1 \leq i \leq 164$<br>$D^{R_i}$ are domains in all right-leaning websites; $165 \leq i \leq 556$ |
| Crawl (SQLite Database) | $C_n^{P_j}$ represents crawl database file for persona $P_j$, $0 \leq j \leq 8$ and $1 \leq n \leq 5$, 15 runs of the same setup depending if it is a simple persona or a rich persona. |

based on these characteristics, focusing on age and gender, and training from different locations in the USA.

We use `alexa.com` which gives the list of websites that are most popular with different demographics[2], as captured in October 2018. However, an important question is: *How many of the websites should be visited, before a user profile becomes "stable"?* To check this, we visit the top websites in each category in a random order[3], and count the number of distinct third party cookies obtained after each new visit. Figure 2 shows the number of third party cookies for personas with different demographics, after visiting the top websites for each category. We observe a large increase in the number of cookies when the first 2–3 websites are visited, and then the increase in absolute numbers of distinct cookies tapers off. Following the average number of 30 cookies found per website in a study of 1 million Alexa-ranked websites [17], we say that a persona has "matured" once it has reached $\geq 50$ unique third party cookies. We observe that visiting 6–7 websites is sufficient for a persona to mature, except in the case of Youth, which does not receive more than 100 cookies (possibly due to laws such as COPPA [13] which safeguard against collecting personal information on very young users) even after visiting 20 websites. However, even in the case of the Youth demographic, we observe that the first 2–3 website visits already expose the user to the majority of distinct third party cookies, and by the time 6–7 websites are visited, well over 50% of the eventual total number of cookies have been collected.

Therefore, we conservatively assume a persona with a specific demographic being "stable" or "established" by visiting the top 10 websites associated with users of that demographic, and storing the third party cookies obtained. Also, visiting the same websites from different locations produced similar numbers of cookies. When
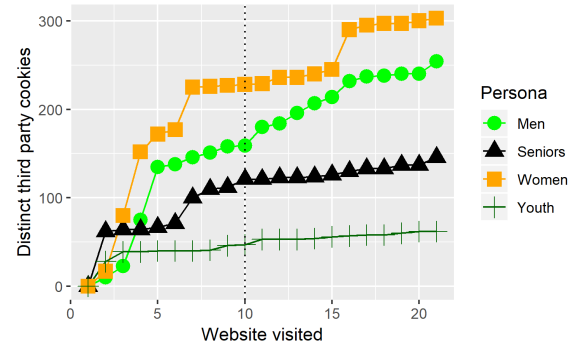


**Figure 2: Numbers of distinct third parties observed after visiting a set of top `alexa.com` websites for building persona per category (man, woman, young, senior).**

**Table 2: Examples of websites from each category of `alexa.com` for creating personas with specific demographics.**

| Demographic | Sample websites |
|-------------|-----------------|
| Youth (Y)<br>Senior (S) | (Student, Kidzworld).com, Voicesofyouth.org<br>Aarp.org, (Medicare, Cms).gov |
| Woman (W)<br>Man (M) | (Cosmopolitan, Womansday, Sheknows).com<br>(Menshealth, Mensjournal, Esquire).com |

visiting other websites (e.g., the HPWs in §2.5), we mimic a specific demographic by loading cookies obtained for that user group. Similar numbers of websites have been used in other studies to establish a profile of cookies. For example, in [28] they used the top 10 websites of each of 16 categories for analyzing ad content.

Note that we also use compound personas with more than one demographic characteristic (e.g., young man, or senior woman). To establish compound personas, we visit and store cookies from the top 10 websites from each demographic (e.g., 10 of the top websites for seniors and 10 for women collectively establish the "Senior Woman" persona), in a random order. We encode each persona with the initial letter of the feature incorporated in the persona. For example 'SW' signifies a persona representing *Senior-Women*. Similarly a YM persona represents the persona of a Young-Man. The personas used are listed in Table 2.

We visit all websites in a persona-specific list as a stateful crawl that stores the user history in a browser. After the visits are finished, we dump the browser state accumulated as an archive file per persona. This persona state corresponds to stored cookies, HTTP calls done, location, etc. Each archived file can then be used to bootstrap a browser with the said persona, ensuring our measurements per persona for each visit of a HPW are bootstrapped with the same browser persona state. Next, we discuss more details on automation and browser settings of the crawling approach.

## 2.4 Crawling Engine of Framework

We make use of *OpenWPM*[1], a popular tool for measurements and automating web browsers, which is developed by Princeton University under the Web Transparency & Accountability Project [17].

---

[2]Under https://www.alexa.com/topsites/category
[3]We have verified that the actual order of visit does not affect the results

Using OpenWPM, we create user personas and bootstrap parallel sessions of *Firefox* browsers to visit HPWs with each persona. In order to be able to compare browsing experience of a persona with a baseline, we load null personas (i.e., a user with no website visit history) to the browser and re-execute the website auditing. Also, using the same tool, we collect cookies set both by javascript calls (i.e., transactional cookies) and HTTP calls (regular cookies set in the browser, logged as Profile cookies). Transactional cookies include all logs of cookies set in the browser, including the deleted instances of cookies. In our analysis, we consider both types and refer to them as *Cookies*.

We kept most of the default settings of OpenWPM's browser instance and updated a few for our use. The default settings include loading a Firefox browser (only this is supported by the tool), having always enabled third party (tp_cookie) tracking, keeping blockers like donottrack, disconnect and ghostery, https-everywhere, adblock-plus, ublock-origin and tracking protection as false, so that nothing is blocked. The updates include enabling the http_instrument which logs HTTP responses, requests and redirects, using the selenium headless browser to perform crawling and, setting js_instrument, cookie_instrument and save_javascript to true to store javascript cookies table, regular (profile) cookies table and all javascript snippets loaded and executed by website respectively.

A HPW crawl with a loaded persona is stateless, i.e., each HPW website visit is independent. We repeated each HPW crawl 5 times per persona and 2 times per baseline, to account for infrequent, but unavoidable network errors. In later sections, we label a persona visiting a left- or right- HPW, using 'L' and 'R' respectively, and append this with the persona's acronym. For example, 'L:Y' is for a *Youth* persona visiting *left* HPWs; similarly 'R:SM' is for a *Senior-Man* persona visiting *right* HPWs.

Since each browser instance is independent, multiple browser instances can be initiated on a crawling server (depending on its available resources). This parallelization allows for multiple, simultaneous crawls of personas and baselines, scaling our auditing to many websites at the same time from one location. This also allows us to capture the same tracking and advertising effort from third parties, before ad-campaign budgets change.

Our crawls were executed in Nov-Dec 2018. Our crawling engine was set up on 10 Amazon Web Services (AWS) instances (each: 1 GB memory, 1 core, 8 GB storage, Ubuntu 14.04) in the USA, for instantiating parallel crawls. The crawling was orchestrated by OpenWPM, loading a Firefox 52.0 browser instance and visit websites using Selenium. Each AWS instance uses a parallel and independent OpenWPM instance loaded with a different persona profile. Each crawl gives us ~500MB SQLlite database across the 10 instances. This database stores information about the HTTP calls, cookies, visit sequence, and other meta information based on the settings described earlier.

## 2.5 Datasets Collected

We implemented our methodology into the framework illustrated in Figure 1 for user tracking on 667 HPWs. Each website is marked as left- or right-leaning ($W^L$ and $W^R$, respectively) based on the description and self attestation from the website's 'about' page, or

**Table 3: Crawls for persona building (P:) and visits to HPWs (HPW:). Second column: total count of requests or responses across crawls; Third column: average count per website; Fourth: standard deviation (SD) per website; Fifth: median count per website.**

| Response Type | Total Count | Mean per website | SD per website | Median per website |
|---|---|---|---|---|
| P: HTTP Requests | 53.2k | 190 | 186 | 135 |
| P: HTTP Responses | 49.6k | 177 | 155 | 132 |
| P: HTTP Redirects | 7k | 25 | 46 | 9 |
| P: Cookies | 36.8k | 131 | 192 | 67 |
| HPW: HTTP Requests | 8.3 M | 159 | 194 | 98 |
| HPW: HTTP Responses | 8.1 M | 144 | 181 | 93 |
| HPW: HTTP Redirects | 2.5 M | 65 | 149 | 40 |
| HPW: Cookies | 3.5 M | 60 | 132 | 16 |

facebook page description. Interestingly, out of 667 websites curated in 2016, 111 websites were not active in Nov-Dec 2018 when the crawls took place. Thus, in our study we use 556 active websites: 164 left- and 392 right-leaning HPWs. We normalize our results to account for the imbalance between the two lists. We visit HPWs using baseline (null) and persona profiles. We store the cookies served to each persona profile or new user and analyze their distributions, type of first or third party sending them, popularity of the HPW involved, etc. A summary of our dataset regarding HTTP calls and cookies set during persona building and final crawls for all personas and HPWs is given in Table 3. The results on baseline profiles and personas are analyzed in the next two sections, respectively.

## 3 BASIC USER TRACKING IN HPWs

Our modus operandi is to use the previous methodology with our deployed framework to visit various HPW, and examine differences in the tracking performed by the first and third parties included in each. In this section, we focus on null profiles to measure basic tracking performed on new (null) users. We conduct an investigation on the differences in tracking, by comparing left and right HPWs (Sec. 3.1), by looking into the top trackers on the Web and whether they are over/under represented in left and right HPWs (Sec. 3.2), and finally, by studying popularity of HPWs and checking if this associates with differences in tracking (Sec. 3.3).

## 3.1 Who facilitates more tracking: left or right?

The volume of cookies in a website is a good proxy for measuring how much the website tracks a user directly, and how much it enables third parties in its page to track this user. To understand how much of this tracking is happening on HPWs, in Figure 3, we plot the Cumulative Distribution Function (CDF) of number of cookies set on a baseline user's profile when visiting $W^L$ and $W^R$. On average, $W^R$ set 9 more cookies than $W^L$ on baseline users.

Next, we study the type of cookies set on baseline users, using a list of Disconnect.me [15]. In Figure 4, we breakdown the cookies into six types: *first-party, advertising, analytics, content, social, and other*. Third party cookies from advertising entities significantly
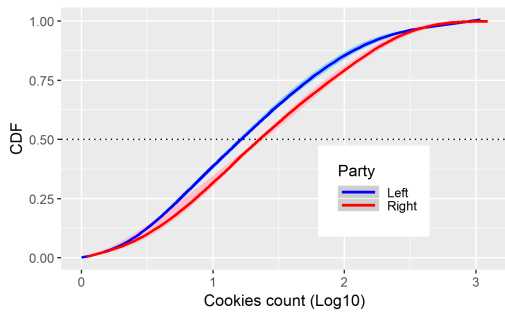
**Figure 3: CDF of total number of cookies stored per website, when a baseline user visits left and right-leaning HPWs.**
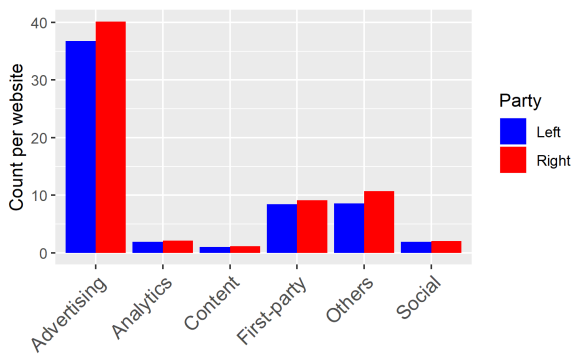


**Figure 4: Average number of cookies per website, per type of domain sending them (using the `Disconnect.me` categories).**
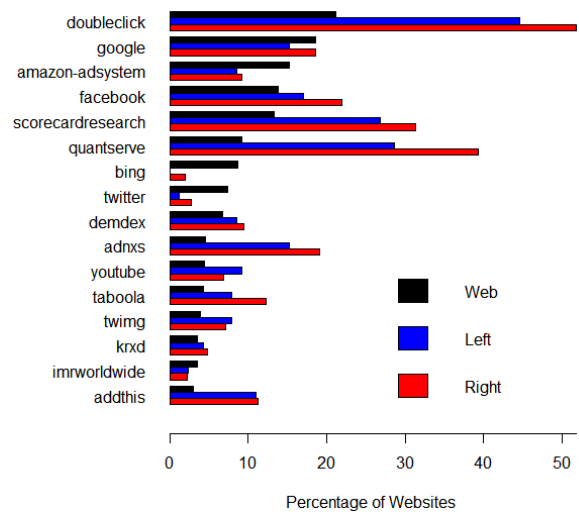


**Figure 5: Percentage of websites with cookies set (x-axis) from top 16 tracking domains on the Web (y-axis), for left and right-leaning HPWs, as well as in the Web, for a baseline user. Domain ranking based on live list of `whotracks.me`.**
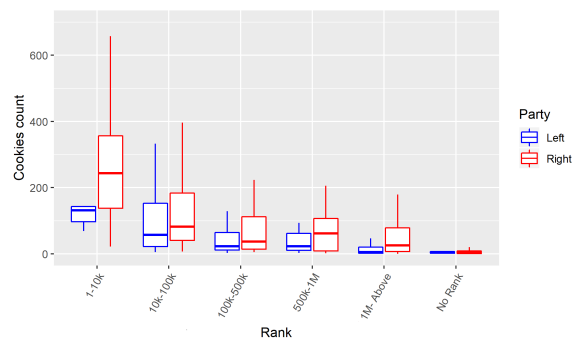


**Figure 6: Cookies count for baseline users visiting $W^R$ and $W^L$, vs. the `alexa.com` global ranking of these websites captured on 25/02/19. Top rank $W^R$ have highest median of cookie count and the median decreases as the rank increases. Websites with no ranking have very few cookies.**

outnumber all other types. Furthermore, $W^R$ place more cookies of all types, and especially for advertising, in comparison to $W^L$.

## 3.2 Is this tracking more than the general Web?

Earlier, we showed that HPWs enable tracking of users by third parties, and that $W^R$ do so in higher intensity than $W^L$. But who are the top trackers in these partisan websites? And how different is the tracking they do, when a baseline user visits such HPWs? We extract the top trackers on the Web using a live list maintained by `whotracks.me` [52] on 25/09/19, and compare this list against the trackers detected in the two groups of websites, when they are visited by a baseline user. We find that 72% of third parties match between the list and the ones on HPWs. However, there are also differences among the HPW tracking ecosystem and the Web with respect to intensity in tracking from each party. Figure 5 shows a histogram of percentage of websites that specific trackers appear in, and drop cookies on their users. Many of these top trackers are over-represented in the HPWs examined, in comparison to the Web. For example, *Doubleclick, Scorecardresearch, Quantserve, and Adnxs,* appear in at least twice as many websites (proportionally) than in the general Web. Furthermore, many of them appear in many more websites on the right ($W^R$) than the left ($W^L$), demonstrating an interest in tracking users visiting right-leaning HPWs.

## 3.3 Is tracking associated with site popularity?

We observed differences between $W^R$ and $W^L$ in number of cookies they drop on users, type of third parties involved, and top trackers in the Web that also appear in these websites. Next, we investigate how the tracking intensity of HPWs associates with the popularity of each HPW. In Figure 6, we present the cookie counts of HPWs dropped on baseline users, against the respective rank range of each website based on Alexa. Again, we find that regardless of the rank range, $W^R$ serve more cookies than $W^L$. Interestingly, the top ranked $W^R$ (i.e., 1-10K) demonstrate the highest median count, with that count decreasing for lower ranks.

In the next section, we compute various measures to provide statistical evidence of differences in personas and their demographics vs. intensity of tracking facilitated by Hyper-partisan websites.

(a) Third parties & persona building third party overlap　　(b) KS-statistic for comparison of third parties　　(c) Third party difference of baseline & personas
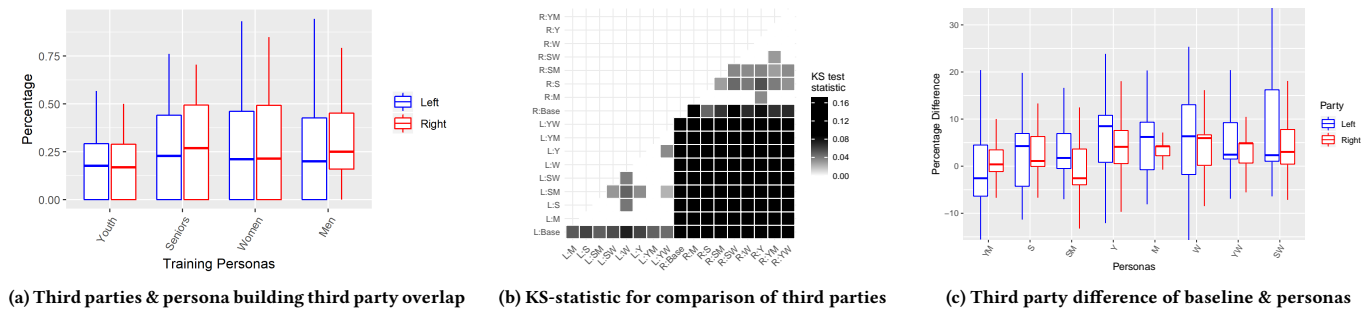
**Figure 7: (a) Overlap between third parties dropping cookies during persona building, and when the same personas visit HPWs. (b) Heat-map of KS statistic test for all pairwise comparisons between distributions of numbers of unique third parties serving cookies. All cells with $p \geq 0.01$ are whited out; only cells with $p < 0.01$ are colored. (c) Percentage difference between third parties serving cookies to baseline and loaded personas. The x-axis is sorted on medians of all personas.**

## 4 USER PERSONA TRACKING IN HPWs

The central question of this study is whether online personas exhibiting specific partisan browsing patterns are targeted in a differentiated fashion by the ads and analytics ecosystem. We use our deployed framework to investigate this question. Using the built personas, we first examine the overlap between third parties for distinct personas across partisan aisles. This helps us ascertain whether our framework's personas are being treated as valid by the tracking ecosystem (Sec. 4.1). Then, we measure statistical differences between third parties across personas and popularity of domains (Sec. 4.2). We also use a matrix-based clustering technique to investigate whether trackers behave preferentially across partisan lines (Sec. 4.3). Then, we investigate the information flows between trackers, by measuring cookie synchronizations happening across webpage visits (Sec. 4.4). Finally, we evaluate whether the ad-ecosystem differentiates between our partisan personas in terms of the ad-pricing offered across different HPWs (Sec. 4.5).

### 4.1 Third Parties in User Personas and HPWs

To produce systematic results regarding tracking of personas, we first need to validate whether our personas are being tracked by the HPWs ad-ecosystem. Validating this step imparts confidence on our results. To that end, we select the personas of youth, seniors, women and men, and separately visit left and right leaning HPWs, to check the overlap of common third parties across HPWs and personas. For this, we measure the total number of common third parties in personas and left or right HPWs. As seen in Figure 7a, all comparisons yield upwards of 25% overlap in domains of third party cookies. This signifies that our personas can be actively tracked by trackers that also exist in left and right-leaning HPWs.

### 4.2 Which demographic is tracked more?

Next, we compare the number of third party cookies from different domains stored for different personas (Figure 7b). Again, we see that $W^R$ have many more third parties embedded in their pages. They also have higher number of cookies than $W^L$, perhaps expectedly given the larger number of third parties. This result points to a tendency from $W^R$ to monitor more intensely users across the

Web with tracking technologies, with the purpose of (re)targeted advertising.

Although fewer third parties are interested in $W^L$, the top ones that are involved have a higher coverage than the left HPWs covered. As example, most of the top third parties such as *Google* and *Adobe* have similar coverage of both $W^L$ and $W^R$. However, providers such as *Pubmatic* and *Adsupply* have a higher coverage of and specialization in $W^L$ or $W^R$, respectively. In cookies count, we also find extreme cases from $W^R$ (e.g., spectator.org, realclearpolitics.com) and $W^L$ (e.g., newshounds.us, opednews.com) which set more than a *thousand* cookies.

Next, in Figure 7c, we compare the difference in numbers of third parties and cookies, when comparing a baseline persona with one of the curated personas. A few personas, such as Youth Man (YM), seem to experience a slight fall in number of cookies relative to baseline. However, most others see an increase in the number of third parties setting cookies. From Figure 7c, we see that, in general, single-feature personas such as *Woman, Man* and *Youth* see the highest increase in number of cookies placed, regardless of political leaning of the website. Comparing this with Figure 2, where *Youth* hardly gets ~50 cookies even after visiting 10 websites, this increase in tracking over baseline suggests unusual behavior on HPWs. From the same figure, for $W^L$, we see the highest increase in tracking is for *Youth* and for $W^R$ the highest tracking increase is for *Woman*.

As a special case, we also study two popular websites: *MSNBC* ($W^L$) and *Fox News* ($W^R$). Figure 8 shows the cookies placed by these two websites on visitors with different personas. These examples highlight that there is a bias in the way partisan news websites change their tracking behaviors, depending on the demographic of the user visiting them. Also note that *Fox News* ($\in W^R$) serves more cookies than *MSNBC* ($\in W^L$) for all personas.

### 4.3 Hyper-partisan Clustering

We test our framework on self declared left or right leaning websites [9]. Hence, we hypothesize that the trackers being deployed on these HPWs would converge onto preferentially tracking personas conducive to the prime audiences of these websites. For example, a
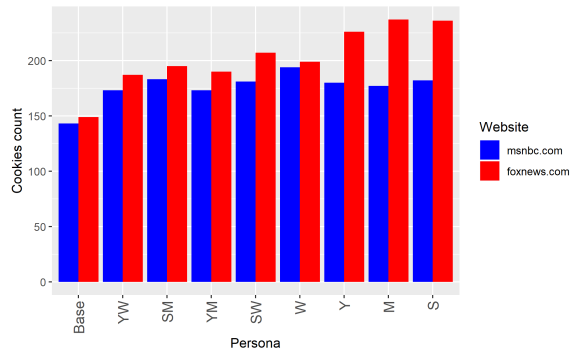
**Figure 8: Variation in cookies count for foxnews.com (top ranked $W^R$) and msnbc.com (top ranked $W^L$) with various personas. In both websites, the cookie count is lowest for the baseline persona. X-axis is sorted by increasing count of cookies on both sides.**

tracker deployed on a right leaning website would prefer deploying the tracking cookies on a right leaning persona as opposed to a left leaning persona.

To test this hypothesis, we use Non-Negative Matrix Factorization (NMF) [29] to cluster the websites, as well as the targeted demographics into the most viable clusters. The members of these clusters exhibit more similar behaviors to each other, compared to members outside these clusters.

NMF decomposes any non square matrix $A$ into two components such that $A \approx B \times C$, where $B$ is called the factor *basis* matrix with the dimensions $p \times k$ , and $C$ is called the *coefficient* matrix with dimensions $k \times c$, for any choice of $k$ clusters. For a given choice of $k$, the NMF algorithm tries to solve the optimization problem:

$$min||A - BC||^2_F, \text{ such that } B \geq 0 \text{ and } C \geq 0 \tag{1}$$

where $F$ represents the *Frobenius* norm. To setup this optimization problem, we first need to compose the matrix $A$, such that it reflects the cookie profile observed by any given persona. We create a $p \times c$ dimensional matrix, which we call $A$, where $p$ represents the rows corresponding to all the personas we curate, and $c$ represents the columns which correspond to the different third party. We normalize each row, such that $A_{ij}$ now represents the percentage of cookies injected for a user of the $P_i$ persona by the $j^{th}$ domain.

As described in Section 3.1, we have distinct set of cookies for each persona viz. *first-party, advertising, analytics, content, social and others.* Hence, we create 6 distinct matrices corresponding to the 6 different types of cookies. We further factorize each of these 'persona cookie profile' matrices $A^t$, into corresponding basis $B^t$ and coefficient $C^t$ matrices where $t \in (1, 6)$ using NMF.

According to [10, 23], clusters should be chosen to have high cophenetic correlation and small residuals. After experimenting with various values of $k (\in [2, 10])$, we find that the best clustering is obtained when $k = 3$, giving us the maximum cophentic correlation (0.982) and minimum residual value (9).

The results of this clustering have two implications. First, the well-defined clusters for $k = 3$ validate our hypothesis that the trackers deployed on HPWs preferentially converge onto specific

partisan personas. Second, it allows us to further pinpoint how left- and right-leaning personas are associated with different kinds of third parties, ranging from advertising to analytics. The value of $k = 3$ also matches the intuition that two of the three clusters could be assigned to the left and right partisan positions, respectively. The third cluster can be interpreted as the middle ground (if any) between party lines. This can be seen in Figure 9. Figure 9a (first party) shows that our profiles clearly fall into either the Left or Right cluster. Figure 9b (first party) shows only the top 10 websites for clarity (rather than all HPWs). Here again, we can see websites fall into either the left or right category. After we computed the NMF on the different matrices $A^t$, we get the various $B^t$ and $C^t$ factors in Figure 9, for advertising, analytics, etc. Figure 9a shows that 'first party' and 'other' domains essentially take clearly the extreme far clusters, as they have in cookies set from $W^L$ and $W^R$ (Figure 9b). Personas exposed to $W^L$ are easier to cluster for advertising, analytics and content, whereas social is less clear. Furthermore, from the $C^t$ factors in Figure 9b, we see that some first party domains are clearly clustered in the left (e.g., *The Nation*) and others as right (e.g., *Fox News*). Also, for analytics and advertising, many domains are heavy on the right-wing cluster. For content, many domains including *google, microsoft* and *adobe* appear to be clustered on the left-wing cluster, apart from *oberon*[4]. Interestingly, for social, *Facebook* and *reddit* are clustered in the right-wing cluster.

## 4.4 User information flows and Cookie Synchronization between trackers

So far, we studied individual cookies set for each user (persona) in our methodology, and their potential impact on tracking of the said user. In this section, we set out to explore the possible discrepancies in the flows of the user information among the different tracking and advertising entities. The scope of these information flows is not only to share directly information about the current user, but more importantly to synchronize the different aliases by which a user may be known to different entities, building this way a common identifier. This identifier will later allow the server-to-server exchange of user data in the background [2], or data purchase in user data markets [16, 47]. This alias synchronization is the point where all user data collected by different third party trackers get attributed to a very single rich user profile.

The most common alias (or user ID) used on the web for user identification is the cookie ID, set on the user side the first time a domain meets the said user. However, cookies are domain-specific. So according to the same-origin policy, a domain cannot see the cookies set by another domain. To bypass this and join together the different sets of user IDs, trackers and advertisers invented the *Cookie Synchronization* (CS) technique, which they use in order to pass information from one to the other tagged with the ID of the user, thereby merging their databases in the background.

CS is performed when a third party $D1$, embedded in website $W$, redirects or issues a new HTTP request to a different collaborating third party $D2$, while piggybacking in the request the cookie ID $userD1$ it has set for the given user. This way, $D2$ learns that the user just visited $W$ (information obtained from the referrer field) and is also known on the web as *user D*1. After that, $D2$ can drop

---

[4]oberon-media.com is a multi-platform firm that gives gaming solutions
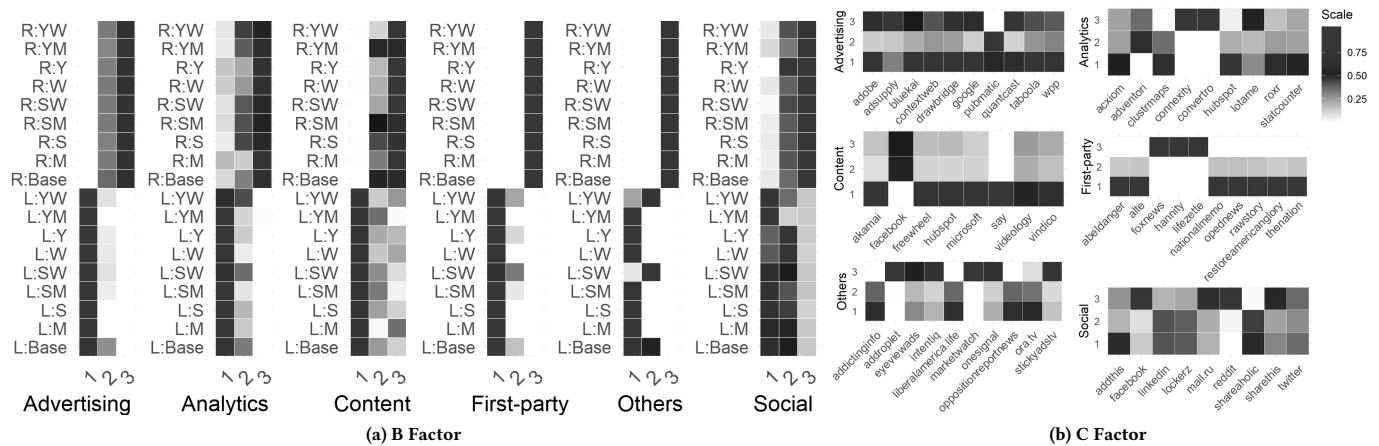
(a) B Factor



(b) C Factor

**Figure 9:** $B^t$ and $C^t$ factors from NMF clustering of $A^t$ for six categories and 3 clusters (i.e., $k = 3$). We show the scale of all the figures in top-right corner, which is normalized from 0 to 1.
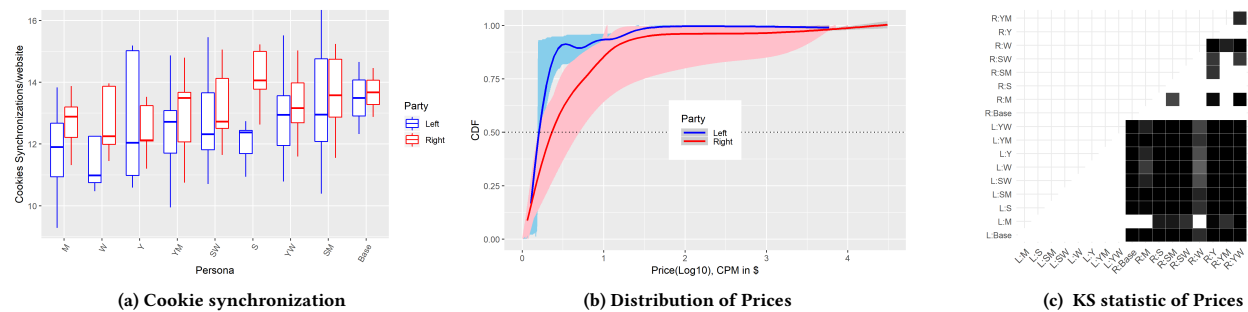


(a) Cookie synchronization



(b) Distribution of Prices



(c) KS statistic of Prices

**Figure 10: (a) Cookie synchronizations per HPWs for the different personas. For the vast majority of personas, the persona visiting $W^R$ receives more cookie synchronizations than the one visiting $W^L$. (b) Distribution of prices in CPM paid by advertisers to deliver ads to personas. The median price for $W^L$ is \$0.561 CPM, and the median price for $W^R$ is \$0.667 CPM. Considering the top 25% most expensive ads, $W^R$ got (up to 5×) higher ad-prices than $W^L$. (c) Heat-map with the KS statistic of 2-sample KS test for all pairwise comparisons between distributions of paid prices per persona, and visited $W^L$ and $W^R$. All cells with $p \geq 0.01$ are whited out; only cells with $p < 0.01$ are colored.**

its own cookie on the user side or sync a previously set cookie. It is apparent that the more CS appear in a website, the more third parties can learn about the user.

To explore if there are any discrepancies in this ID synching and merging for the different categories of personas, we leverage existing CS detection methodologies [2, 39]. We apply these detection methodologies in our dataset, and in Figure 10a, we plot for each crawl the detected CS per HPW for the different personas. Although the ratio of CS per website varies per persona, for the vast majority of personas, the websites in $W^R$ not only set more cookies (as shown in Figure 3) but also perform more CSs than $W^L$. By facilitating more CSs, $W^R$ enable more third party trackers to access and track the browsing behavior of the users than $W^L$. This leads to more intense personalization and targeting of ads, which can be measured by the prices paid by advertisers for said ads. We explore this in the next paragraphs.

## 4.5 Which user costs more to advertisers?

Digital advertising is where user tracking and collected user data get translated to money. Users get targeted with ads that match their interests, as extracted from the collected user data. As a next step, we set out to explore the monetary cost advertisers pay for the different personas, with the dimension of hyper-partisanship. To achieve this, we leverage the most prevalent mechanism of digital ad-buying and delivery nowadays, i.e., the real-time programmatic ad-auctions [3]. Under these setting, available ad-slots in a publisher's website can be auctioned as soon as the website is rendered in the user's display. This Vickrey type [51] of auction takes place at real time in ad-markets called ad-exchanges (ADX) and special bidding engines called Demand-Side Platforms (DSPs). These DSPs represent ad-agencies that place bids per auctioned ad-slot. In order for the DSPs to decide if and how much they will bid for an ad-slot, detailed information is needed regarding the interests, geolocation, gender,

age, sexual or political preferences, etc., of the current user viewing the webpage. Such information is being bought from trackers and data warehouses on the background in an asynchronous fashion. To reduce latency, along with the bid, DSPs also send the ad-impression they want to deliver to the users if their bid wins the auction.

In order to assess the winning bid and thus the price paid for the ad-slots each user receives, we leverage the most dominant ad-auction protocol nowadays: the RTB protocol [30]. Specifically, we capture the step where the user's browser (as instructed by the ADX) issues a HTTP request to the winning bidder notifying them about: (i) its win, (ii) the charge price, and (iii) the successful rendering of the ad-impression. Following the methodology of previous works [34, 40], we estimate the monetary values the ad-ecosystem paid for the different category of personas in our dataset.

In Figure 10b, we present the distributions of the ad-prices for $W^L$ and $W^R$. As we can observe, the median price for $W^L$ is \$0.561 CPM (CPM=Cost Per Mille, i.e., cost per thousand impressions), and the median price for $W^R$ is \$0.667 CPM. However, when it comes to the top quartile of the more expensive ads, $W^R$ get (up to 5×) higher prices than the corresponding ones of $W^L$.

Finally, to explore the variability of prices across different personas, in Figure 10c, we plot a heat-map with the statistics of 2-sample KS tests of all pairwise comparisons between the paid ad-prices per persona, for $W^L$ and $W^R$. We not only see a particular variability among the different types of HPWs as demonstrated in Figure 10b, but we also see that there are personas with significantly higher ad-prices than others for the same type of websites.

### 4.6 Summary

In general, our results indicate that right-leaning HPWs have more third parties embedded in their pages, they monitor and track users with more cookies, these third parties synchronize their cookies more often or with more intensity, and all this tracking activity leads to better ad-targeting, and thus higher paid prices for delivered ads. We summarize our findings as follows:

(1) Right-leaning HPWs news have embedded significantly more third parties and store up to 25% more cookies to user browsers than left-leaning HPWs.
(2) User personas exhibiting realistic and representative demographic characteristics tend to receive up to 15% more tracking (cookies) from HPWs than baseline personas (i.e., with no characteristics).
(3) Popular, highly-ranked HPWs track users more intensely than lower-ranked HPWs.
(4) Single-feature personas, e.g., Woman, Man, Youth, are highly tracked by default, no matter what party-leaning they demonstrate through their visits.
(5) It is possible to group personas to left or right-leaning clusters, based on the domains and numbers of cookies they receive while browsing HPWs.
(6) Right-leaning HPWs aid third parties to track more intensely their users than left-leaning HPWs, by facilitating up to 50% more cookie synchronizations between trackers than left-leaning HPWs.

(7) Right-leaning HPWs deliver to users ads costing up to 5x more than in left-leaning HPWs, as a consequence of more intense tracking and ad-targeting performed.

## 5 RELATED WORK

**Politics and Polarization.** Political parties involve user audience often through polarization, as well as by urging them to volunteer and campaign for the party [6]. With most of the political discussion happening in the cyber-space, user experience leads the way of engagement and enchantment [5, 31]. For example, news personalization is a way to send curated news briefing to subscribers [32, 50]. [12] coined dark money which is spent on negative ads. Another selective political initiative is the highly polarized HPWs in US politics, studied by [9]. HPWs declare themselves right- or left-leaning in their web portals. Interestingly, many of these HPWs were deleted after US 2016 elections. However, more than 550 still exist and we focus our study on them. Alternate (or fake) news websites can also be highly engaging. URL sharing of alternate news across the Web was studied in [55]. Studies have tried to understand the political stance of users and news websites by various methods. Examples include the use of NMF clustering ($k = 2$) [27], finding non-binary political ideological scores, sentiment analysis [43], twitter profile data [42], etc.

**User Tracking and Information flow.** Many studies have analyzed the extent of online user tracking. In [25], authors measure the trackers per website and find that, on average, there are 10 different trackers per website; in news websites this average is 15 trackers. Similarly, in [17], authors used Alexa top 500 websites for each of the 16 basic content categories of Alexa. They find that websites providing editorial content embed the largest amount of trackers given their lack of external funding sources, and their need to monetize page views through advertising.

A popular web tracking mechanism is the use of cookies set on the user-side. In [22], researchers measured differences in third party cookies offloaded on a browser by collecting data from real users over a period of 1 year, before and after the GDPR regulation [19], and found that the number of cookies does not change, not least because users tend to pay little attention to the GDPR options. In [34], researchers were the first to analyze the technique of synchronizing different user IDs that various entities set for the same user, thus reducing the user's anonymity on the Web. They also study Cookie Synchronization (CS) in conjunction with the ad-delivery paradigm of RTB auctions. In [37], researchers measure the costs of digital advertising and use CS as a metric for the reduction of user's privacy. Also, [39] conducted a thorough analysis of CS by analyzing a year-long dataset of real mobile users and showed that CS increases the number of domains that track a user's browsing by a factor of 6.75. In [38] authors show that CS can break users' secure SSL sessions and increase their identifiability on the Web, even if they browse over VPN. They probed 12,000 top Alexa websites and found that 1 out of 13 websites expose their visitors to this severe privacy leak. In [2], authors study CS together with re-spawning cookies and investigate how these two can form persistent cookies that users cannot abolish even after clearing their cookie jar. In [18], authors simulate users browsing the web in an attempt to study the possibility of a passive eavesdropper leveraging third party HTTP

tracking cookies for mass monitoring. They find that indeed the eavesdropper can reconstruct 62-73% of a typical user's browsing history.

**Biases in Online Ecosystem.** In an attempt to shed light over today's opaque online ad-ecosystem, [28] proposed Sunlight: a system to detect targeting in large-scale experiments with statistically justifiable results. Their system consists of four components that generate, refine, and validate hypotheses to reveal the possible causes of observed targeting. Similarly, [7] developed a content-agnostic methodology to detect client- and server-side flows of information between ad-exchanges by leveraging retargeted ads. In [56], user browsing history is studied to define audience gender and age demographics. An SVM model is trained on the text in websites, to predict user features and bias in advertising. Evidences of bias in ad-placement is given by [14]. Using a tool named AdFisher, the authors investigate ad-placement with adjustment of Google ads settings of user. Thus, setting gender to female resulted in fewer ads related to high paying jobs. In [26], authors studied 5 million paid ads on Facebook and claim that US users are mostly targeted by foreign groups using loopholes in regulations of Facebook.

## 6 DISCUSSION AND FUTURE WORK

Understanding the dissemination of misinformation over online platforms and fringe news sources (e.g., Breitbart or Infowars) has become a crucial topic of study because of its considerable impact on our online culture. These sources are often found responsible for spreading news and opinion pieces with extreme views along partisan lines. This strategy keeps users engaged with the websites for prolonged periods of time. A major incentive of such Hyper-partisan websites (HPWs) to follow this strategy is monetization by delivering targeted and personalized ad-content to profiled users. In fact, these websites do not need to provide any credible content or news, thus, reducing costs for journalists and other expenses, otherwise incurred on legitimate news websites. However, until now, it remained unclear how such HPWs perform targeted advertising and whether they *differentiate* their online user tracking based on demographics and partisan leaning of their audience.

To shed light to this problem, in this paper, we present a methodology that creates artificial users (personas) with certain demographic features. We find that a stateful browsing of a small number of 10 websites is sufficient to build a persona (cookie state) of hundreds of unique third parties. We store browser states of personas and use them to emulate diverse types of browsing patterns along partisan lines, and in the process, record how these websites track them. The data acquired from such persona-based crawls are an asset for investigating user experience, bias in tracking, and various other privacy metrics. By visiting 556 HPWs, we collected and analyzed a dataset of ~19 million HTTP (requests, responses and redirects) calls and ~3.5 million cookies. In general, we find a higher amount of tracking in the right-leaning HPWs, and also an increased tracking for personas exhibiting specific patterns with respect to a combination of gender and age demographics.

Among the right-leaning HPWs, the top ranked (by Alexa) websites are consistently loaded with more trackers, in comparison to less popular HPWs. By performing a co-clustering on the personas using the NMF model, we observe that we can group personas and

third parties in a low dimensional space using the types of cookie domains as features. We obtain three optimal clusters which can be labelled as (mild) left, (mild) right and neutral. Finally, investigating the third party cookie synchronization (CS) and examining the prices paid for ads as declared in the real-time bidding protocol, we conclude that right-leaning HPWs apply more intense CS and command higher prices for delivered ads.

For the benefit of reproducibility, we reported an in-depth process for the creation and deployment of said personas. This approach is easily scalable across different demographic features and browsing patterns. Our framework can also be repurposed for auditing other ecosystems such as e-commerce, web search engines, public social media pages, and other recommendation services. Our framework can be useful for enforcing new regulations under data protection (e.g., GDPR [19]) which have been introduced to give back control of personal data to their owners, and increase overall Web transparency. As GDPR enforcement is in its initial phase, most of the control on user's data and tracking still remains with the websites, allowing many HPWs to take liberties on how to comply with regulations. For example, in our study, we found cases of right-leaning HPWs (e.g., spectator.org) and left-leaning HPWs (e.g., newshounds.us) setting more than 1000 cookies per user.

Indeed, there are some limitations of our framework, which we plan to overcome in the near future. For example, data protection laws do not allow cross-border tracking, which results in logging 33% fewer trackers when the crawling is performed from websites under EU's jurisdiction, as compared to traffic logged within US. Another area to explore is the variability in tracking of personas, and its relation with the ad-economies and end-user experience. For example, if a website is visited on two consecutive days, tracking metrics will most likely be different. Also, there is a wide-range of different, less or more pronounced and compound personas that someone can emulate with our framework. Clearly, every persona comes with costs of finding appropriate websites to build it, and additional experiments to make sure it triggers network traffic realistic for an online user of that persona. In fact, less pronounced personas may not be able to collect appropriate and conclusive data of differential tracking. In the future, we aim to build better tools to understand these problems. We also plan to further develop our framework to look into the ad-ecosystem on fringe news websites, and replicate our work for multiple countries to discover user differential tracking and personal data leaks.

# REFERENCES

[1] 2020. OpenWPM Framework. https://github.com/mozilla/OpenWPM.

[2] Gunes Acar, 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 SIGSAC Conference on Computer and Communications Security*. ACM, 674–689.

[3] Adext Corp. 2019. How Programmatic Advertising Has Changed The Online Advertising Landscape. https://blog.adext.com/programmatic-ads-changed-online-advertising/.

[4] Seth Schoen Adi Kamdar, Rainey Reitman. 2013. NSA Turns Cookies (And More) Into Surveillance Beacons. https://www.eff.org/deeplinks/2013/12/nsa-turns-cookies-and-more-surveillance-beacons.

[5] Pushkal Agarwal, Nishanth Sastry, and Edward Wood. 2019. Tweeting MPs: Digital Engagement between Citizens and Members of Parliament in the UK. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 13. 26–37.

[6] Andrew O Ballard, D Sunshine Hillygus, and Tobias Konitzer. 2016. Campaigning online: Web display ads in the 2012 presidential campaign. *PS: Political Science & Politics* 49, 3 (2016), 414–419.

[7] Muhammad Ahmad Bashir, Sajjad Arshad, William Robertson, and Christo Wilson. 2016. Tracing Information Flows Between Ad Exchanges Using Retargeted Ads. In *25th USENIX Security Symposium*.

[8] Leonid Bershidsky. 2018. Yes, Russia Abused Facebook. But Did It Work? https://www.bloomberg.com/opinion/articles/2018-12-18/yes-russia-abused-facebook-but-did-it-work.

[9] Shweta Bhatt, Sagar Joglekar, Shehar Bano, and Nishanth Sastry. 2018. Illuminating an ecosystem of partisan websites. In *Companion Of The Web Conference*. 545–554.

[10] Jean-Philippe Brunet, Pablo Tamayo, Todd R Golub, and Jill P Mesirov. 2004. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences* 101, 12 (2004), 4164–4169.

[11] Juan Miguel Carrascosa, Jakub Mikians, Ruben Cuevas, Vijay Erramilli, and Nikolaos Laoutaris. 2015. I Always Feel Like Somebody's Watching Me: Measuring Online Behavioural Advertising. In *Proceedings of the 11th ACM Conference on Emerging Networking Experiments and Technologies (CoNEXT '15)*.

[12] Daniel E Chand. 2017. Dark Money and Dirty Politics: Are anonymous ads more negative? *Business and Politics* 19, 3 (2017), 454–481.

[13] Federal Trade Commission. 1998. Children's Online Privacy Protection Rule (COPPA). https://www.ftc.gov/enforcement/rules/rulemaking-regulatory-reform-proceedings/childrens-online-privacy-protection-rule.

[14] Amit Datta, Michael Carl Tschantz, and Anupam Datta. 2015. Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies* 2015, 1 (2015), 92–112.

[15] Disconnect.me. 2019. Disconnect tracking protection. https://github.com/disconnectme/disconnect-tracking-protection/blob/master/services.json.

[16] Hazem Elmeleegy, Yinan Li, Yan Qi, Peter Wilmot, Mingxi Wu, Santanu Kolay, Ali Dasdan, and Songting Chen. 2013. Overview of turn data management platform for digital advertising. *Proceedings of the VLDB Endowment* 6, 11 (2013), 1138–1149.

[17] Steven Englehardt and Arvind Narayanan. 2016. Online tracking: A 1-million-site measurement and analysis. In *Proceedings of the SIGSAC Conference on Computer and Communications Security*. ACM, 1388–1401.

[18] Steven Englehardt, Dillon Reisman, Christian Eubank, Peter Zimmerman, Jonathan Mayer, Arvind Narayanan, and Edward W Felten. 2015. Cookies that give you away: The surveillance implications of web tracking. In *Proceedings of the 24th International Conference on World Wide Web*. 289–299.

[19] European Commission. 2018. Rules for the protection of personal data inside and outside the EU. https://ec.europa.eu/info/law/law-topic/data-protection_en/.

[20] Peter Geoghegan and Jenna Corderoy. 2018. Revealed: Arron Banks Brexit campaign's secret meetings with Cambridge Analytica. https://www.opendemocracy.net/uk/brexitinc/peter-geoghegan-jenna-corderoy/revealed-arron-banks-brexit-campaign-had-more-meetings-w.

[21] William T Harding, Anita J Reed, and Robert L Gray. 2001. Cookies and Web bugs: What they are and how they work together. *Information Systems Management* 18, 3 (2001), 17–24.

[22] Xuehui Hu and Nishanth Sastry. 2019. Characterising Third Party Cookie Usage in the EU after GDPR. In *Proceedings of the 10th ACM Conference on Web Science*. 137–141.

[23] Lucie N Hutchins, Sean M Murphy, Priyam Singh, and Joel H Graber. 2008. Position-dependent motif characterization using non-negative matrix factorization. *Bioinformatics* 24, 23 (2008), 2684–2690.

[24] Carter Jernigan and Behram FT Mistree. 2009. Gaydar: Facebook friendships expose sexual orientation. *First Monday* 14, 10 (2009).

[25] Arjaldo Karaj, Sam Macbeth, Rémi Berson, and Josep M Pujol. 2018. WhoTracks.Me: Monitoring the online tracking landscape at scale. *arXiv preprint arXiv:1804.08959* (2018).

[26] Young Mie Kim, Jordan Hsu, David Neiman, Colin Kou, Levi Bankston, Soo Yun Kim, Richard Heinrich, Robyn Baragwanath, and Garvesh Raskutti. 2018. The stealth media? Groups and targets behind divisive issue campaigns on Facebook. *Political Communication* (2018), 1–27.

[27] Preethi Lahoti, Kiran Garimella, and Aristides Gionis. 2018. Joint non-negative matrix factorization for learning ideological leaning on Twitter. In *Proceedings of the 11th International Conference on Web Search and Data Mining*. ACM, 351–359.

[28] Mathias Lecuyer, Riley Spahn, Yannis Spiliopolous, Augustin Chaintreau, Roxana Geambasu, and Daniel Hsu. 2015. Sunlight: Fine-grained targeting detection at scale with statistical confidence. In *Proceedings of the 22nd SIGSAC Conference on Computer and Communications Security*. ACM, 554–566.

[29] Daniel D Lee and H Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* 401, 6755 (1999), 788.

[30] Jack Marshall. 2014. What is Real Time Bidding. https://digiday.com/media/what-is-real-time-bidding/.

[31] John McCarthy, Peter Wright, Jayne Wallace, and Andy Dearden. 2006. The experience of enchantment in human–computer interaction. *Personal and ubiquitous computing* 10, 6 (2006), 369–378.

[32] Solomon Messing and Sean J Westwood. 2014. Selective exposure in the age of social media: Endorsements trump partisan source affiliation when selecting news online. *Communication Research* 41, 8 (2014), 1042–1063.

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

[34] Lukasz Olejnik, Tran Minh-Dung, and Claude Castelluccia. 2014. Selling off privacy at auction. In *Network and Distributed System Security Symposium (NDSS)*.

[35] Michalis Pachilakis, Panagiotis Papadopoulos, Evangelos P. Markatos, and Nicolas Kourtellis. 2019. No More Chasing Waterfalls: A Measurement Study of the Header Bidding Ad-Ecosystem. In *Proceedings of the Internet Measurement Conference*. ACM, 280–293.

[36] Elias P. Papadopoulos, Michalis Diamantaris, Panagiotis Papadopoulos, Thanasis Petsas, Sotiris Ioannidis, and Evangelos P. Markatos. 2017. The Long-Standing Privacy Debate: Mobile Websites vs Mobile Apps. In *Proceedings of the 26th International Conference on World Wide Web* (Perth, Australia). 153–162.

[37] Panagiotis Papadopoulos, Nicolas Kourtellis, and Evangelos P. Markatos. 2018. The Cost of Digital Advertisement: Comparing User and Advertiser Views. In *Proceedings of the World Wide Web Conference* (Lyon, France). 1479–1489.

[38] Panagiotis Papadopoulos, Nicolas Kourtellis, and Evangelos P. Markatos. 2018. Exclusive: How the (Synced) Cookie Monster Breached My Encrypted VPN Session. In *Proceedings of the 11th European Workshop on Systems Security* (Porto, Portugal) *(EuroSec'18)*. 6:1–6:6.

[39] Panagiotis Papadopoulos, Nicolas Kourtellis, and Evangelos P. Markatos. 2019. Cookie Synchronization: Everything You Always Wanted to Know But Were Afraid to Ask. In *Proceedings of the 28th International Conference on World Wide Web*.

[40] Panagiotis Papadopoulos, Nicolas Kourtellis, Pablo Rodriguez Rodriguez, and Nikolaos Laoutaris. 2017. If you are not paying for it, you are the product: How much do advertisers pay to reach you?. In *Proceedings of the Internet Measurement Conference*. ACM, 142–156.

[41] Paul Hilder Paul Lewis. 2018. Leaked: Cambridge Analytica's blueprint for Trump victory. https://www.theguardian.com/uk-news/2018/mar/23/leaked-cambridge-analyticas-blueprint-for-trump-victory.

[42] Marco Pennacchiotti and Ana-Maria Popescu. 2011. Democrats, republicans and starbucks afficionados: user classification in twitter. In *Proceedings of the 17th SIGKDD International Conference on Knowledge Discovery and Data mining*. ACM, 430–438.

[43] Ferran Pla and Lluís-F Hurtado. 2014. Political tendency identification in twitter using sentiment analysis techniques. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical Papers*. 183–192.

[44] Criag Silverman, Jane Lytvynenko, Lam Thuy Vo, and Jeremy Singer-Vine. 2017. Data, analytic code, and findings related to the BuzzFeed News article. https://github.com/BuzzFeedNews/2017-08-partisan-sites-and-facebook-pages.

[45] Criag Silverman, Jane Lytvynenko, Lam Thuy Vo, and Jeremy Singer-Vine. 2017. Inside The Partisan Fight For Your News Feed. https://www.buzzfeednews.com/article/craigsilverman/inside-the-partisan-fight-for-your-news-feed.

[46] Konstantinos Solomos, Panagiotis Ilia, Sotiris Ioannidis, and Nicolas Kourtellis. 2019. Talon: An Automated Framework for Cross-Device Tracking Detection. In *International Symposium on Research in Attacks, Intrusions and Defenses (RAID)*. 227–241.

[47] Sarah Spiekermann, Rainer Böhme, Alessandro Acquisti, and Kai-Lung Hui. 2015. Personal data markets. *Electronic Markets* 25, 2 (01 Jun 2015), 91–93.

[48] Patrick Svitek and Haley Samsel. 2018. Ted Cruz says Cambridge Analytica told his presidential campaign its data use was legal. https://www.texastribune.org/2018/03/20/ted-cruz-campaign-cambridge-analytica/.

[49] The Federal Trade Commission. 2000. FTC Sues Failed Website, Toysmart.com, for Deceptively Offering for Sale Personal Information of Website Visitors. https://www.ftc.gov/news-events/press-releases/2000/07/ftc-sues-failed-

website-toysmartcom-deceptively-offering-sale.

[50] Neil Thurman and Steve Schifferes. 2012. The future of personalization at news websites: lessons from a longitudinal study. *Journalism Studies* 13, 5-6 (2012), 775–790.

[51] William Vickrey. 1961. Counterspeculation, auctions, and competitive sealed tenders. *The Journal of finance* (1961).

[52] whotracks.me by Cliqz and Ghostery. 2019. Bringing Transparency to online tracking. https://github.com/cliqz-oss/whotracks.me/tree/master/whotracksme/data/assets.

[53] John Wilander. 2019. Intelligent Tracking Prevention 2.3. https://webkit.org/blog/9521/intelligent-tracking-prevention-2-3/.

[54] World Wide Web Consortium (W3C). 2010. Same Origin Policy.

[55] Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. 2017. The web centipede: understanding how web communities influence each other through the lens of mainstream and alternative news sources. In *Proceedings of the 2017 Internet Measurement Conference.* ACM, 405–417.

[56] Yong Zhang, Hongming Zhou, Nganmeng Tan, Saeed Bagheri, and Meng Joo Er. 2017. Targeted Advertising Based on Browsing History. *arXiv preprint arXiv:1711.04498* (2017).

[57] Shenghuo Zhu, Kai Yu, Yun Chi, and Yihong Gong. 2007. Combining Content and Link for Classification Using Matrix Factorization. In *Proceedings of the 30th SIGIR Conference on Research and Development in Information Retrieval* (Amsterdam, The Netherlands). ACM, 487–494.