Under the Spotlight: Web Tracking in Indian Partisan News Websites

Vibhor Agarwal[⋆α], Yash Vekaria[⋆α], Pushkal Agarwal[β], Sangeeta Mahapatra[γ],
Shounak Set[β], Sakthi Balan Muthiah[α], Nishanth Sastry[δ], Nicolas Kourtellis[ζ]
[α]The LNM Institute of Information Technology, Jaipur, India
[β]King's College London, London, United Kingdom
[γ]German Institute for Global and Area Studies, Hamburg, Germany
[δ]University of Surrey, Surrey, United Kingdom
[ζ]Telefonica Research, Barcelona, Spain
[α]{vibhor.agarwal.y16, yash.vekaria.y16, sakthi.balan}@lnmiit.ac.in, [β]{pushkal.agarwal, shounak.set}@kcl.ac.uk
[γ]sangeeta.mahapatra@giga-hamburg.de, [δ]n.sastry@surrey.ac.uk, [ζ]nicolas.kourtellis@telefonica.com

ABSTRACT

India is experiencing intense political partisanship and sectarian divisions. The paper performs, to the best of our knowledge, the first comprehensive analysis on the Indian online news media with respect to tracking and partisanship. We build a dataset of 103 online, mostly mainstream news websites. With the help of two experts, alongside data from the Media Ownership Monitor of the Reporters without Borders, we label these websites according to their partisanship (Left, Right, or Centre). We study and compare user tracking on these sites with different metrics: numbers of cookies, cookie synchronizations, device fingerprinting, and invisible pixel-based tracking. We find that Left and Centre websites serve more cookies than Right-leaning websites. However, through cookie synchronization, more user IDs are synchronized in Left websites than Right or Centre. Canvas fingerprinting is used similarly by Left and Right, and less by Centre. Invisible pixel-based tracking is 50% more intense in Centre-leaning websites than Right, and 25% more than Left. Desktop versions of news websites deliver more cookies than their mobile counterparts. A handful of third-parties are tracking users in most websites in this study. This paper, by demonstrating intense web tracking, has implications for research on overall privacy of users visiting partisan news websites in India.

1 INTRODUCTION

India represents the largest and the most diverse news media market among democracies, with more than 100,000 registered newspapers and 400 news channels[2] in 22 official languages.[3] The growth of online news has been the fastest in the emerging markets, with India ranking among the top ten globally when it comes to print and online news media [21]. Unfortunately, this growth of online political communications has been accompanied by rising partisanship [9, 25]. The mainstream news media as major agents of information and influence, become important here.

This paper focuses on major news websites in terms of how they track their users. Tracking allows them to obtain rich information about readers, which may serve their business interest in revenue generation through targeted ads, as well as their political interest in setting agendas. There have been US-based studies about partisan media mostly in terms of their polarizing effects [5, 16, 38] and a few on tracking [2, 24]. For India, while there have been a few works on the division in the news media along partisan lines [27], there is a lack of comprehensive, data-driven research on news websites and tracking behavior. Indian news media are a major source of information for the population [34].[4] Their tracking behavior has socio-political implications as they are, by and large, a trusted source of public information [22].

In this work, for the first time, we provide a comprehensive study of the news websites in India with respect to partisanship and tracking of online users. We focus on the online platforms of the largest English, Hindi, and regional language news media (including those with print or broadcast platforms and the digital only ones) that can reach more than 77% of India's population[5,6], making them vulnerable to tracking. We first identify the major Indian news publications based on their circulation figures from the Registrar of Newspapers for India (RNI) supplemented with Indian Readership Survey of Q4 2019. We then create a list of 103 news websites, curated primarily from Alexa [3] and Feedspot [14]. Secondly, with the help of two experts in political science and journalism, alongside data from the Media Ownership Monitor of the Reporters without Borders, which traces associations between the media and political parties and corporate interests [28], we label the 103 websites according to their partisanship as Right-, Left-, Centre-leaning, or Unknown (methodology explained in Section 3).

We address the following questions: RQ1: What is the extent of tracking across partisan news websites? RQ2: What kind of tracking methods are used on users? To answer them, we measure the intensity of user tracking across partisan websites with simple and advanced mechanisms: basic first and third-party cookies, cookie synchronization, device fingerprinting, and invisible pixel-based tracking (Section 4).

We share our Dataset, OpenWPM Crawls, and Codes publicly with the research community for reproducibility and extension of

[⋆] Both the authors contributed equally to this research.

[2] https://www.indiantelevision.com/regulators/ib-ministry/total-of-television-channels-in-india-rises-to-892-with-three-cleared-in-june-160709

[3] Registrar of Newspapers for India: http://rni.nic.in/

[4] https://bestmediainfo.in/mailer/nl/nl/IRS-2019-Q4-Highlights.pdf

[5] Media Research Users Council: https://bestmediainfo.in/mailer/nl/nl/IRS-2019-Q4-Highlights.pdf

[6] Broadcast Audience Research Council, India: https://barcindia.co.in/

our work[7]. From this study, we derive the following key findings (Section 5). The 103 Indian news websites studied have more than 100K cookies, for an average of over 100 cookies per website, but several websites have much higher number of cookies. For example, ~1400 cookies are set on the first-party – *Sandesh.com*, by itself and its third-parties. Left- and Centre-leaning websites serve more (median) cookies than Right-leaning websites. Desktop versions of websites set more cookies than their mobile versions, with interesting exceptions. Third-party domain *doubleclick.net* is present in 86% of news websites; such ubiquitous presence allows the tracking of a huge proportion of users' browsing histories.

In addition to the large numbers of cookies, we also find evidence of practically every known advanced method of user fingerprinting. Around 18% of all distinct third-parties, and 25% of all distinct first-parties in our data are involved in cookie synchronization. Around 50% of unique user IDs are synced across tracking domains through cookie synchronization. Cookie synchronization is higher among Left-leaning websites and their third-parties than for Right- and Centre-leaning websites. Over 25% of news websites use device fingerprinting, which is invisible to the user and invasive to their online privacy. Around 25.7% of Left, 23.7% of Right, and 17.9% of Centre websites employ different fingerprinting scripts to track users. More than 2.5K invisible (1x1 pixel) images (i.e., 23% of all sent images) are detected on news website homepages. Invisible pixel-based tracking is employed more by Centre, followed by Left and then the Right websites.

2 BACKGROUND AND RELATED WORK

We briefly discuss here the partisan nature of Indian news websites as well as online tracking techniques studied in literature.

Partisan nature of Indian news: This paper takes partisanship to mean an adherence to the political beliefs and identification with a political party or cause, manifesting positively as a civic ideal of shared values or negatively as a pathology where loyalty to a party's ideology/values/goals may trump logic and tolerance to other political views [39]. While numerous political parties exist in India, the three broad strands of political worldviews correspond to three principal political formations at the national level of Indian politics: "Left" represented by parties like the Communist Party of India (Marxist), "Right to Right-Centre" represented by the Bharatiya Janata Party, and "Left-Centre" corresponding to the Indian National Congress. As India is a highly diverse country with their political parties and media reflecting this diversity, we take Right-leaning news media to correspond with the Right to Right of the Centre spectrum of ideologies, the Left-leaning news media to correspond with the Left to Left of the Centre spectrum, and the Centre-leaning media to be positioned in between the Right-Centre and the Left-Centre. The growth of heightened political partisanship may have a dramatic impact on media behavior and their influence on public opinion, especially if they intensely track users.

Online tracking ecosystem and measurements: With the rise of online information consumption, online platforms have attracted third parties for online advertising [26, 30]. These advertisements are strategically drafted and placed on websites to get more user attention including pop-ups and banners [26, 35]. These websites track users by injecting cookies at the users' side [7, 10, 20, 37] for content personalization and improving user experience. However, cookies and other data are also shared with other third parties, raising privacy concerns. Users have an option to accept or reject these third-party cookies, but many users are not aware of the consequences if they accept them. These websites also use more sophisticated tracking techniques like cookie synchronization [1, 2, 11, 20, 31, 36], device fingerprinting [11, 29], and invisible (1x1) pixel-based tracking [15]. Since users are often unaware of their presence, such methods pose a greater privacy threat to the websites' visitors. Studies have shown that some popular trackers like Doubleclick and Google Analytics (both Google trackers) can be present in up to 50% and 70%, respectively, of top one million visited websites [11]. Specifically, news websites have seen large volume of trackers and advertisements including political campaigns [2, 11, 32]. Among USA news websites, Right-leaning websites track users more and have high cookie synchronization within the partisan group websites [2]. Having said that, less is known about the tracking ecosystem of Indian news media, which has recently seen exponential growth in online consumption. There are studies in online engagement (including social media) showing polarization and media bias, but none covers the exposure of user data to the tracking world [8, 25, 33]. With our work, we aim to fill this gap by measuring the extent to which users are exposed to a high amount of web tracking, using the aforementioned four tracking techniques. We also explore tracking on desktop and mobile platforms in Indian news media with partisan leanings.

3 DATA COLLECTION AND LABELING

Here, we discuss the methodology followed to curate a list of top news websites in India, including metadata crawled for each using *Feedspot* [14] and *Alexa.com* [3], to label these websites based on their political leanings (Sec. 3.1). Furthermore, in Sec. 3.2, we provide details of our website traffic crawling using *OpenWPM* [11, 12], a tool for desktop browser automation and crawling, and *Cookies.txt* [17], a browser plug-in for mobile browser automation.

3.1 Websites Partisan Labeling

We follow the methodology outlined in Figure 1 (left part) for website list creation and partisanship labeling.

List Creation: We first examined a list of 141 top Indian news websites on the Web (ranked as on 28 April 2020) provided by Feedspot [14]. This website, maintained by over 25 experts, is updated daily and covers a wide range of factors to rank and discover the most prominent online news websites in India. They curate websites whose publishers explicitly publish their content via Feedspot, as well as by monitoring search engines and social media through in-house media tools. The next list of websites we studied is from Alexa (29 April 2020) [4]. Alexa Internet, Inc., is an American Web traffic analysis company, whose toolbar gathers information of around 30 million websites across the globe, based on their internet browsing behavior and traffic patterns. Their website stores the data and provides extensive analysis of the websites. From Alexa, we got a list of 49 top Indian news websites based on their online popularity and traffic. Some of them were common with the Feedspot data. We combined Feedspot and Alexa lists to obtain a list of 153 websites.

[7] Data and code available at http://tiny.cc/india-tracking

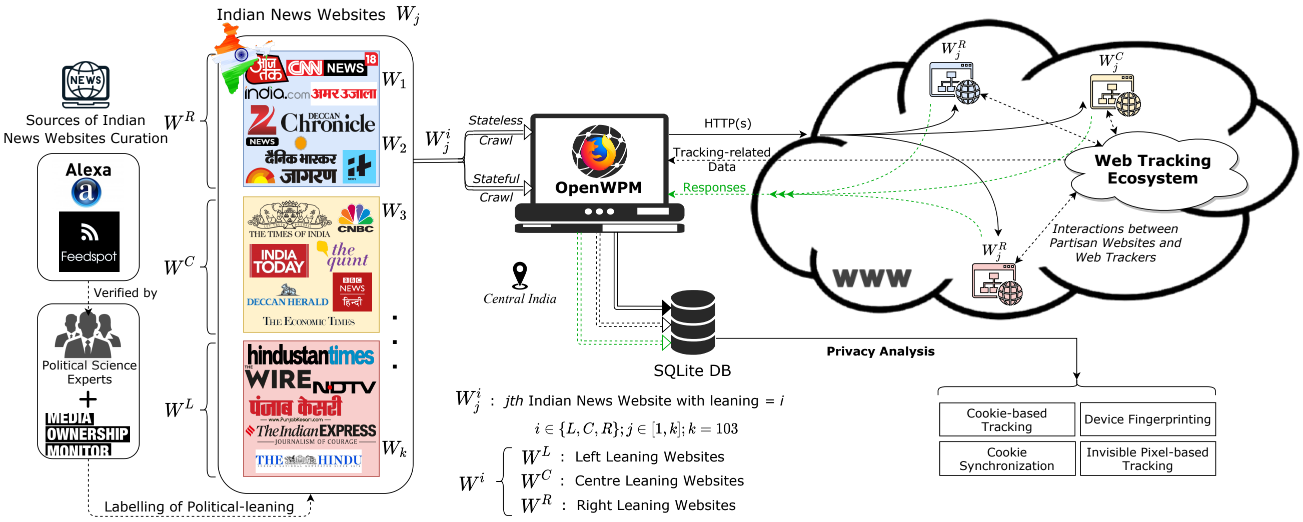
Indian News Websites W_j
Sources of Indian News Websites Curation
Alexa
Feedspot
Verified by
Political Science Experts
+
MEDIA OWNERSHIP MONITOR
Labelling of Political-leaning
W^R
W^C
W^L
W_1
W_2
W_3
W_k
W_j^i
Stateless Crawl
Stateful Crawl
OpenWPM
Central India
HTTP(s)
Tracking-related Data
Responses
W_j^R
W_j^C
Web Tracking Ecosystem
Interactions between Partisan Websites and Web Trackers
WWW
SQLite DB
Privacy Analysis
Cookie-based Tracking
Device Fingerprinting
Cookie Synchronization
Invisible Pixel-based Tracking
W_j^i : jth Indian News Website with leaning = i
$i \in \{L, C, R\}; j \in [1, k]; k = 103$
W^i
W^L : Left Leaning Websites
W^C : Centre Leaning Websites
W^R : Right Leaning Websites

Figure 1: Our framework for labeling Indian news websites along partisan lines and collecting web traffic data for studying web tracking mechanisms. Colors represent party-leaning: Right=Blue, Centre=Yellow, and Left=Red.

A large portion of news consumption in India happens through online platforms (Facebook, Twitter, and Instagram) rather than TV/Radio [34]. Therefore, we further augment our data by visiting each website's Facebook, Twitter, and Instagram pages for metadata collection. After opening a particular website on Facebook, Twitter or Instagram, we performed (in April 2020) a breadth first search on other 'Indian news page recommendations' shown in the right-side panel under the heading of "Related Pages" in Facebook, "You might like" in Twitter, and "Related Accounts" at the bottom in Instagram. We added to our list all Indian news media shown in recommendations (as described above) while visiting the social media pages of initially curated websites. In the second-iteration, we repeated this with newly collected news media from the first-iteration. We repeated this approach up to five times, by which we observed that 90% of recommendations were already in our dataset. Using this approach, we added to our list 65 new Indian news media leading to a total of 218 websites. Then we removed websites with inactive web pages and retained only those which had more than 10K followers on at least one of the three social media platforms investigated (to ensure we only include the popular ones). Our final list has 123 Indian news websites, spanning nine languages and 28 states. All have an online website, which can be freely accessed over the internet. Out of 123 websites, 10.56% are popular as TV channels, 53.66% are print media and remaining 35.78% only have a website (no TV channel or print media). We determine popularity in terms of viewership/readership in TV/print media.

Website Labeling: In order to understand and categorize websites based on their partisan leanings, we undertook a three-step labeling process. First, we approached two political science and journalism experts who manually coded the political leanings of these websites. This approach has been used by media monitors at Buzzfeed News[8] in past studies to review political leaning in the US news ecosystem. Second, we checked for their partisan associations from Media Ownership Monitor [28] including data on parent company. The labeling was then done along a spectrum of Right (Conservative: Right to Right-Centre), Left (Liberal: Left to Left-Centre), and Centre (i.e., less biased or a combination of both Left and Right, that is, when the same parent company has two ideologically different news sites) categories based on ownership and ideological association. 20 websites were discarded due to uncertainty in their leaning. And the remaining 103 websites were labeled with a partisan leaning and considered for our study. The inter-annotator agreement between experts, measured by Cohen's Kappa, is 0.97. Throughout the paper, we use this categorization, with short names: "Left" for "Left to Left-Centre", "Right" for "Right to Right-Centre", and "Centre" for "Centrist or representing view-points of Right and Left". Our dataset consists of *40* Left-, *26* Centre-, and *37* Right-leaning websites.

3.2 Websites Traffic Data

We start our data collection using OpenWPM [11] by performing five stateless crawls, while visiting the websites' homepages from Central India between August 10, 2020 to August 30, 2020. Stateless crawls make each website visit independent. Parallel browser instances were launched to allow multiple, simultaneous crawls of these news websites from a single location. We performed such crawls across different times and days to account for infrequent but unavoidable network errors during each crawl. We recorded more than 100K cookies in total.

We also performed five time-variant and order-variant, stateful crawls of the websites' homepages from September 01, 2020 to September 15, 2020. Stateful crawls are important since we want to study tracking mechanisms such as cookie synchronization (CS). CS requires state information to be maintained across different websites and visits, to detect if user IDs from previous visits are being synced

[8] https://www.buzzfeed.com/craigsilverman/inside-the-partisan-fight-for-your-news-feed

in future visits and with other websites and their third-parties. Time-variance is applied by crawling on different days with days-long time between crawls.

Order-variant means the websites are visited in a shuffled order for each crawl, for the results to be independent of the website ordering. In stateful crawls, no parallel browser instances are launched to detect third-parties that indulge in cross-site tracking of users.

For 23 of the 103 websites, we also find manually that they serve separate mobile versions. Therefore, we perform five additional crawls for these mobile websites to compare tracking behavior in desktop websites and their mobile counterparts. The crawling for mobile websites uses *Cookies.txt*, a Firefox Plug-In [17] to get browser cookies information. We automate this process using Selenium[9]. At first, a Firefox browser is set to not block any type of cookies. Further steps include opening a Firefox Mobile Emulator in an incognito mode, loading the plug-in, visiting the mobile versions of the websites' homepages (e.g., *m.timesofindia.com*), and storing cookies information. In these five crawls, we store 1400 cookies in total.

4 MEASURING TRACKING MECHANISMS

In this section, we detail the methodology to measure various tracking methods used by Indian news websites and the associated ad-ecosystem – Figure 1 (right part).

4.1 First and Third-party Cookie Analysis

To perform the cookie-based analysis, we use the *javascript_cookies* table of SQLite dump from the OpenWPM crawled data. This data provides information on all different types of cookies being set by different domains. In addition, we use the Disconnect List[10], which is extensively used by the research community to report known tracking domains, and categorize them into eight distinct categories: Advertising, Analytics, Content, Social, Fingerprinting, Cryptomining, Disconnect, and Unknown. We use this list to understand the distribution of cookies across these categories.

4.2 Cookie Synchronization Analysis

Cookie synchronization (CS) is a cross-site tracking mechanism that enables two trackers to generate a detailed browsing profile of the user, by sharing unique user IDs with each other. CS circumvents the Same-Origin Policy (SOP)[11]. Past works have studied CS in different contexts [1, 2, 11, 13, 20, 31, 36]). However, CS has never been studied specifically for Indian news websites along partisan lines or with respect to the privacy implications that it has in the context of India. CS can be abstracted as a two-step process. In the first step, a unique user ID is exchanged between two TPs in the form of HTTP(s) requests, responses, or redirects in an effort to learn the identity of the given user on the web. This ID can be used to aggregate user information by a variety of means [19] through step two. In the second step, domains exchange or merge the identified user's data including browsing histories, browsing patterns, and interests through a separate "data sharing channel" to build a complete, consolidated user profile.

Privacy impact: Tracking and targeting based on CS primarily helps advertisers [23], especially in programmatic (real-time bidding) advertising, where data sharing and purchasing involves CS for better targeting [18]. As a result of CS, trackers are able to track a given user over a larger set of websites, where they may not even be embeded as TPs. In fact, repetitive CS across websites can enrich a particular user's profile built by trackers, helping them to precisely track and target a user over time. Also, server-to-server exchanges of user data (CS step 2 above) have become common [11], enabling deeper user profiling.

Methodology: We capture CS for websites in our dataset using similar methodology of past studies [1, 13, 31]. We use the fundamental structure of the open-source python code from [1] (referred to as `CSCode` hereafter) and make modifications to work for our scenario: unlike [1] that crawled data simultaneously on two machines before analyzing them with `CSCode`, we perform time-variant crawls (Sec. 3.2).

For each crawl, we detect CS for each leaning group and a combination of them. For example, while studying CS between Left and Right, we iterate over all distinct pairs of websites `(w1,w2)` where `w1` is any website which is Left only, while `w2` is Right only (with `w1!=w2` and `(w1,w2)` ≡ `(w2,w1)`). Since we have 39 Left and 37 Right websites, there are 39x37=1443 total pairs. For intra-party comparisons like Right-Right for instance, the total unique pairs will be computed as ${}^{37}C_2 = 666$. Next, for each pair, we consider all the HTTP(s) request, response, and cookies data related to `w1` and `w2`, and use `CSCode` to search for IDs synced between FPs and TPs while visiting `w1` and `w2`. We try all possible combinations of website pairs falling into different partisan lines, i.e.:

- $w1 \in W^L$ and $w2 \in W^L$; $w1 \in W^R$ and $w2 \in W^R$
- $w1 \in W^C$ and $w2 \in W^C$; $w1 \in W^L$ and $w2 \in W^R$
- $w1 \in W^L$ and $w2 \in W^C$; $w1 \in W^R$ and $w2 \in W^C$

Since [1] is an older paper on CS, we validated `CSCode`, as well as various parameters used with recent works on CS [2, 20, 31, 36]). We made the following key changes to ensure result correctness. First, for each URL, `CSCode` extracts the top-level-domain (e.g., *com* from *rtb.gumgum.com*) in [1]. However, it is not relevant to study CS across such top-level domains. Instead, we follow [31] and map all domains (from cookies, requests, response URLs, etc.) to the high-level domains returned by the WhoIS tool[12] (e.g., *rtb.gumgum.com* is mapped to *gumgum.com* as obtained from WhoIS). Second, `CSCode` constraints minimum length of an ID to be `6` characters. However, [36] suggests to discard shorter IDs, since they do not contain sufficient entropy to represent a user ID. We follow [31] and use threshold of `11` characters to minimize false positives. Interestingly, the shortest ID detected in our data is `12` characters long. Third, we upgraded `CSCode` to support python3 and dependencies.

Limitations: `CSCode` gives a strict conservative ID detection with fewer false positives [1]. However, false negatives may occur when ID is shared in URL parameters in an encoded or encrypted format [6, 31], or when ID strings are hidden inside the longer strings with non-standard delimiters. According to [1], the adversarial trackers could have short-lived cookies[13] mapped to user IDs at the backend-server to later on track the user. Such cases are not captured by our code.

[9] https://www.selenium.dev/documentation/en/

[10] https://github.com/disconnectme/disconnect-tracking-protection

[11] SOP allows tracking domains to access only cookies set by them.

[12] https://www.whois.com/

[13] As [1], we consider cookies with expiration date ≤ 30 days

Hence, our results represent a *lower bound* on the actual CS taking place in a real-time scenario.

4.3 Device Fingerprinting Analysis

Privacy impact: A device or browser fingerprinting is a powerful technique that websites and TPs use to identify unique users and track their online behavior. This method collects information about the user's browser type and version, operating system, time-zone, language, screen resolution, and other settings. It can lead to serious privacy issues as users are oblivious to this happening, and can have important implications on the way third-parties track users across the Web *without cookies* in the future.

Methodology: Our fingerprinting measurement methodology [11] utilizes data collected by OpenWPM, as described in Sec. 3.2. In particular, we detect different types of fingerprinting such as canvas, WebRTC, and audioContext, by checking webpages and the interfaces they call, such as *HTMLCanvasElement* and *CanvasRenderingContext2D* for canvas, *RTCPeerConnection*, *createDataChannel* and *createOffer* for WebRTC, and *AudioContext* and *OscillatorNode* for audioContext.

4.4 Invisible Pixel-based Tracking Analysis

Privacy impact: Invisible pixels are 1x1 pixel images that do not add any content to the websites hosting them. TPs use these invisible pixels to track user's behavior on a website. Whenever a website loads, it sends subsequent requests to the server to load various assets like images, ads, and other media on the website. To load these invisible (1x1) pixels on the websites, TPs send some information using the requests sent to retrieve the images. Crucially, the users are unaware of the pixels' existence on the websites and that these pixels report user's activity. Therefore, every such pixel represents a threat to the user's privacy.

Methodology: We follow [15], and for every crawl using OpenWPM, we store all HTTP requests, responses, and redirects, along with response headers, to capture the communication between a client and a server. We then filter HTTP requests and responses by checking the *content-type* in the response header. If the *content-type* is an *image*, the corresponding requests and responses are for images. Next, we check for *content-length* in the response headers to filter out only those HTTP requests and responses with *content-length* less than 1KB. This threshold is used to save storage space (i.e., not to store all images but only probable 1x1 pixel images). In [15], they use 100KB threshold, but this is a very large size for such 1x1 pixel images. In fact, we found all detected invisible pixels in our dataset are less than 1KB in size. All such images are downloaded using the image's URL recorded in the filtered HTTP requests and responses and then checked for the image's dimensions. If both height and width of an image are 1 pixel, then the image is labeled as invisible pixel. The corresponding HTTP request/response, image URL, content length, and third-party setting of each invisible pixel are recorded for further analysis.

5 USER TRACKING VS. PARTISANSHIP

In this section, we present our privacy analysis on the partisan websites of our dataset, and how they track users. We start with cookie-based tracking analysis (Sec. 5.1). We then study more complex tracking techniques such as cookie synchronization (Sec. 5.2), device fingerprinting (Sec. 5.3), and invisible pixel-based tracking (Sec. 5.4).

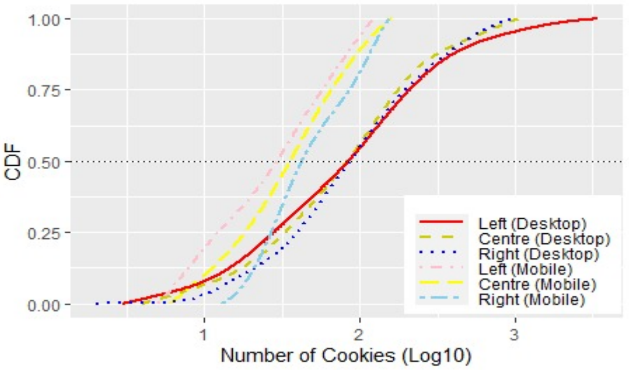

Figure 2: CDF of number of cookies for Left, Centre, and Right-leaning news websites, for their desktop and mobile versions (if available).

5.1 Number of cookies

We analyze 100K cookies placed by FPs and TPs while visiting the 103 Indian news websites. Figure 2 shows the CDF of the number of cookies for all the Left-, Centre-, and Right-leaning news websites available for desktop (103) and mobile (23) versions of the websites. The median number of cookies are 86, 84, and 92 for Left-, Right-, and Centre-leaning desktop websites, and 30, 42, and 36, respectively, for mobile websites. Therefore, in all political leanings, websites for desktop push more cookies to the user's browser than mobile versions (in median). In mobile versions, Centre and Right websites track users more compared to the Left by 1.2 and 1.4 times (KS-value: 0.33, p-value: 0.007), respectively, and Right websites tracks more than Centre websites by 1.2 times (KS-value: 0.28, p-value: 0.054). In desktop versions, median numbers are close for all leanings. The Right websites have fewer cookies than the Left, and the Left has fewer than the Centre. Interestingly, when considering the case of websites for desktop delivering a lot more cookies than the median, Left tracks more than the Right and Centre. For example, *sandesh.com*, which is in the Left to Left-Centre political spectrum, has the highest number of cookies: more than 1400 cookies (median over five crawls). These cookies are set by the FP and TPs on this website. When desktop websites have cookies less than the median, the trend is reversed, i.e., Right tracks more than Left and Centre.

The different versions for desktop and mobile platforms for the same news website imply opportunity for collaboration or data leakage between the two tracking ecosystems across different devices. In Figure 3, we compare the total number of cookies for each of the 23 news websites with mobile and desktop versions. Most websites (20/23) set more cookies in their desktop as compared to their mobile versions. Interesting exceptions are *Times of India*, *Punjab Kesari*, and *Daily Hunt*, which set more cookies in their mobile websites. More cookies indicate higher intensity of tracking as well as network activity (for storing, updating, and synchronizing said cookies) between the browser and server. Therefore, such (mobile) websites

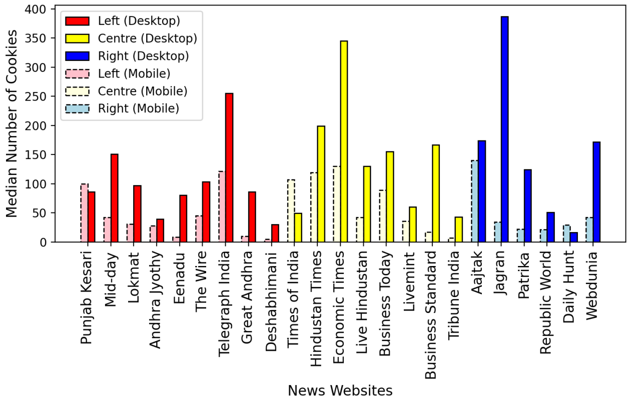

Figure 3: Median number of cookies in mobile vs. desktop versions for 23 news websites, grouped by political leaning in decreasing order of their Facebook followers.

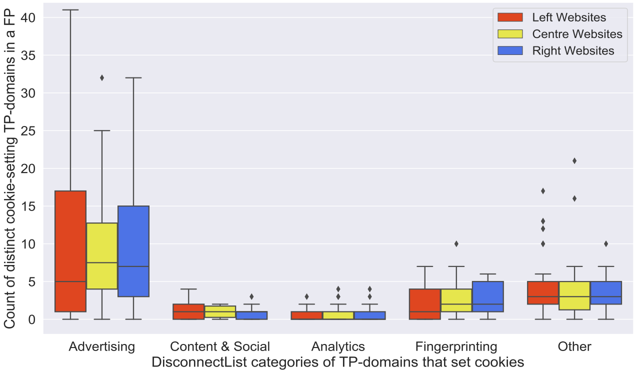

Figure 4: For each first-party (FP), the distribution of count of distinct cookie-setting third-parties (TPs) by DisconnectList categories.

neither respect users' privacy nor consider the mobile device's limited resources regarding power and bandwidth (data) consumption.

We further investigate the difference in tracking between mobile and desktop, and study the unique TP domains that are present in mobile, desktop, or both versions. On one hand, we find 68% of TPs exist in both mobile and desktop versions, allowing them to perform in-depth monitoring of (same) users, and linking them across multiple devices. On the other hand, we find 16% of TPs exist only on mobile versions. For e.g., websites such as *Times of India* and *Punjab Kesari* have more than 50% of their TPs present in their mobile versions and not in their desktop versions.

We also study the type of TPs that set cookies on browsers, using the Disconnect List (DL). Note: we group together "Cryptomining", "Disconnect" & "Unknown" as "Other". Figure 4 shows the box-plot distribution of each category. Statistically, with a KS-value 0.35 (p-value: 0.0195), the largest portion of TP domains is advertising and observed across all partisan websites, with Centre and then, Right being the most frequent. This is unsurprising since most news websites are funded by display ads. Interestingly, the second most frequent category (apart from "Other") is TP domains performing fingerprinting (KS-value: 0.31, p-value: 0.0534). When compared with medians, we again observe Centre and Right websites being more intense with fingerprinting than Left. We investigate such domains further in Sec. 5.3.

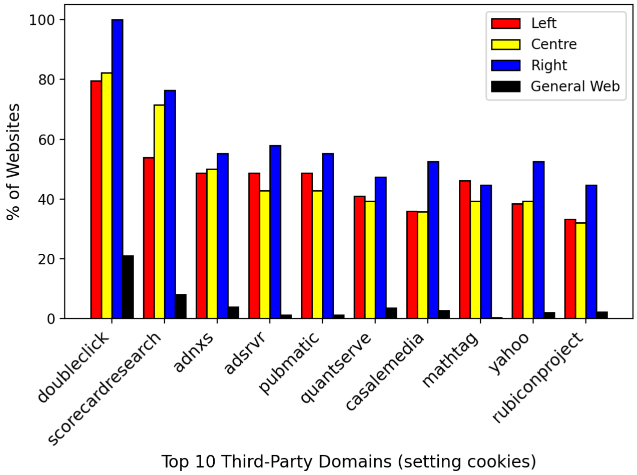

Figure 5: Top 10 TP domains setting cookies in Left, Centre, or Right-leaning news websites. Their presence on general web is also plotted for comparison.

Finally, we look into the top TP domains involved in cookie-based tracking. Figure 5 shows the top 10 TPs, per political leaning of the first-party website embedding them. We also compare the embeddedness of these TPs with their appearance in the "general web". This is to understand how much more or less intensely these TPs track users visiting Indian news websites compared to the general web, following the same strategy as in [2]. For general web, we crawl data from *whotracks.me*, the percentage of websites in which detected third-parties embed their cookies on the Web. We find these TPs are more embedded in the Right-leaning websites than Left or Centre. Unsurprisingly, *doubleclick.net* is present in most websites in our list: 100% of Right, 80% of Left, and 82% of Centre websites, while in general web, it is tracking only 21% of websites. Additionally, we look at the portion of cookies contributed by these TPs. We find *pubmatic.com* sets most cookies, contributing an overall 9% of cookies in our data. Also, the top 10 (2%) TPs set 42% cookies in our dataset.

Takeaways: Desktop versions of websites set more cookies than mobile. Also, Right- and Centre-leaning websites embed more Advertising and Fingerprinting TPs than Left-leaning websites, including the top entity *doubleclick.net*. In general, a handful of TPs provide high coverage of users across all political spectrum of Indian news websites.

5.2 Cookie Synchronization

We compute cookie synchronization (CS) for all stateful crawls as described in Sec. 4.2, and summarize results across different partisan leaning groups, as shown in Table 1.

Table 1: Statistics on cookie synchronizations detected between first party (FP) and third party (TP), or TP-TP domains, for all combinations of FP website pairs crawled, e.g., "Right-Left" means first a visit to a Right-leaning website and then a visit to a Left-leaning website (or vice-versa).

Leaning Group	Avg. ID syncs per unique ID	Avg. ID syncs per TP-TP pair	Avg. ID syncs per FP-TP pair
Right-Right	2.59	3.83	1.65
Left-Left	4.67	4.45	2.23
Centre-Centre	3.37	3.00	1.71
Right-Left	4.75	4.06	1.45
Right-Centre	3.45	3.45	1.63
Left-Centre	**5.92**	**4.81**	**2.46**

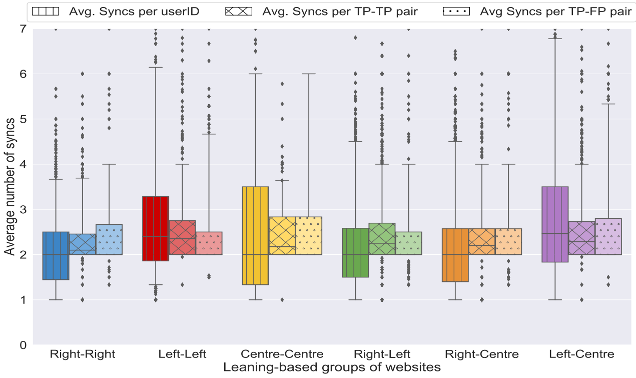

Figure 6: Distributions of average number of CSs per ID, with respect to political leaning groups and combinations.

In general, we see that any user browsing that involves visiting a Left-leaning website (before or after a Left, Right or Centre website) leads to an elevated number of CSs per unique ID, in comparison to only Right- or Centre-leaning websites (first column of Table 1). This is also the case for CSs detected between TP-TP pairs. TPs in Centre-Centre group seem to perform the least amount of such CSs in comparison to other groups. Finally, Left-Left and Left-Centre have the highest CSs in FP-TP pairs in comparison to other groups. Right-related groups perform the least CSs.

In Figure 6 we look at the distribution of CSs performed per pair of websites visited, per combination of partisan website groups. With a KS-value of 0.0748 at 0.0029 significance, the highest number of CS happens when Left-Left (i.e., intra-partisan) group of websites is visited. Similarly, among the inter-partisan groups, Left-Centre website visits involve high CS tracking (KS-test: 0.0431, p-value: 0.0003)

To further investigate the trackers involved in CS, we look at the domains and observe that ~24% of FPs and ~18% of TPs are performing CS. In fact, we observe tracking domains like *pubmatic.com*, which sync with other domains as high as 87 IDs. Additionally, some IDs are synced with multiple domains. For example, ID *c3514a4b-11de-4cce-b428-365a3f6294b1-tuct65bc2e7* was found to be synced across 24 different tracking domains (from approx. 600+ TPs in our data). Moreover, a higher median number of TPs are performing CS in Left and Centre websites than Right. We also plot the top 10 TPs most involved in CS in Figure 7. We observe that the top cookie-setting domains are also present here in CS. In fact, *pubmatic.com* which is setting most cookies, is also performing most CS and in most websites: ~25% Left, ~19% Centre, ~16% Right. Also, *rubiconproject.com* and *doubleclick.net* perform CS in 15-22% of websites.

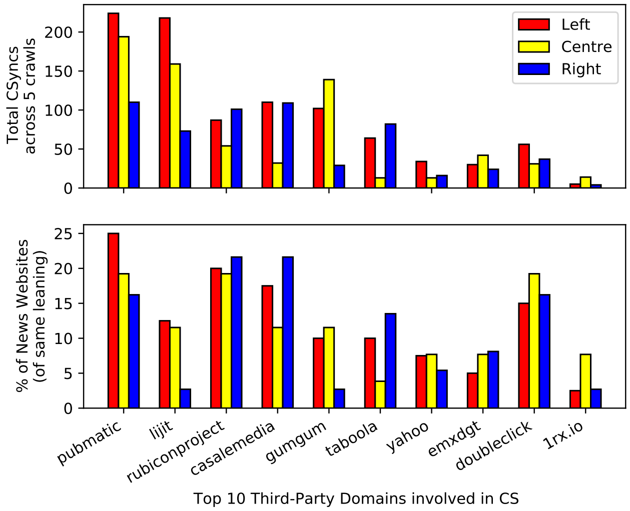

Figure 7: Top 10 TPs involved in CSs, grouped by political leaning. Total CSs (top y-axis) is (TP-TP)+(TP-FP) CSs.

Takeaways: Detected user IDs are synchronized two to six times, on average, between one to five parties, on average, depending on the type of pair entity involved (TP-TP or FP-TP). Same top domains setting cookies, appear to do heavy CS as well, covering up to 25% of websites. Left-leaning websites and their TPs do more CS than Right- or Centre-leaning ones.

5.3 Device Fingerprinting

In this section, we present results of different fingerprinting techniques like Canvas, WebRTC, and AudioContext fingerprinting based on the methodology discussed in Sec. 4.3. Overall, we find 32 distinct fingerprinting scripts set by 18 domains on 25.7% of Left-, 23.7% of Right-, and 17.9% of Centre-leaning news websites. Also, the most dominant type of fingerprinting is Canvas. In particular, 26 canvas scripts are found on 23 (18.7%) websites, from 13 unique domains; top three: *jsc.mgid.com*, *s0.2mdn.net*, and *razorpay.com*. Also, we find one WebRTC script set by *adsafeprotected.com*, and four audioContext scripts in four websites.

Takeaways: Overall, 18-25% of FPs and TPs perform tracking using user device fingerprinting, with Left and Right adopting equally this tracking technology.

5.4 Invisible Pixels

We find 11582 images on the website homepages, out of which 5121 images have less than 1 KB size. Following the process outlined in Sec. 4.4, we identify 2513 invisible (1x1) pixel images, i.e., 21.7% of all images found. Figure 8 shows the CDF of median number of invisible pixels embedded in Left-, Right-, and Centre-leaning

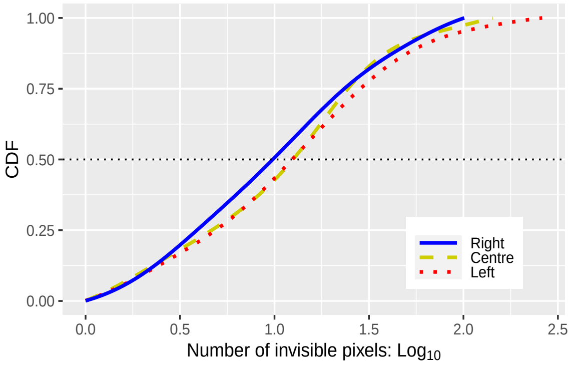

Figure 8: CDF of median number of invisible pixels for Left, Centre, and Right-leaning websites.

websites. These medians are 12, 10, and 15, respectively. The CDF shows more intense pixel tracking by Left and Centre, than Right.

Figure 9 represents the top 20 FP websites having the highest number of invisible pixels, ordered by number of pixels found on their homepages. Out of the top 20, nine are Left, seven are Right, and four are Centre. Again, *Sandesh.com* with its third-parties, earlier found to set most cookies, has the highest number of detected invisible pixels (261). Moreover, 138 distinct TPs are detected setting these 2,513 invisible pixels.

Figure 10 shows the top 10 TPs setting invisible pixels, ordered by total number of pixels set in the news websites. It also shows the total number of pixels set per TP. Google-related properties (*googlesyndication.com*, *google-analytics.com*, and *google.co.in*) dominate the market, as the largest cumulative third-party domain that uses invisible pixels to track users' behavior on these websites. Interesting outliers exist such as *rtb.gumgum.com* that sets 113 invisible pixels on just two Left websites.

Takeaways: Websites embed TPs performing invisible pixel-based tracking, with Centre-leaning websites tracking 50% more intensely than Right, and 25% more than Left. Top TPs in other tracking methods (cookies, CS etc.) also perform heavy pixel-tracking, with *Google* properties covering 60-80% of the websites.

6 DISCUSSION & FUTURE WORK

In this work, and for the first time in literature, we have done an extensive, data-driven study on the Indian online news ecosystem with respect to tracking by websites of mainstream news media with partisan leanings. The sample of news media studied have comparable resources and reach.

Dataset: One of our contributions from this study is the labeled dataset of 103 news websites (reaching 77% of Indian population) with their political leanings (Left, Right, and Centre), which we make publicly available to the research community (along with all crawls and coded methods). The aim of this paper is to show the types and extent of tracking done by mainstream news websites, which sets the essential foundation for future studies on the purpose of such targeting. Further, our findings on tracking in mobile and desktop versions is crucial as more and more Indians have started to consume news on mobile versions.

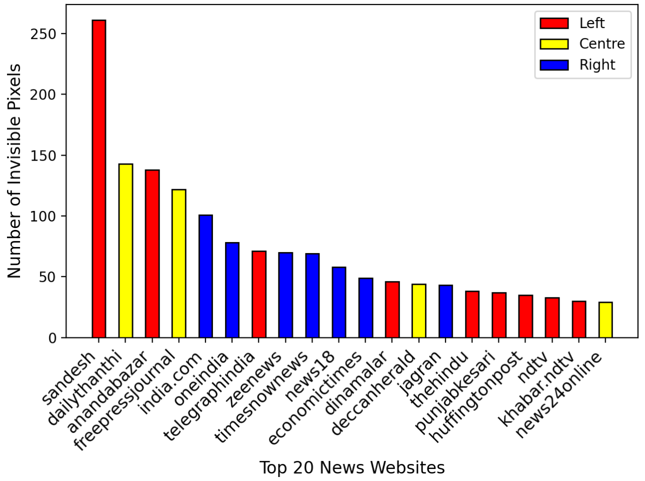

Figure 9: Top 20 news websites having invisible pixels vs. their political leanings.

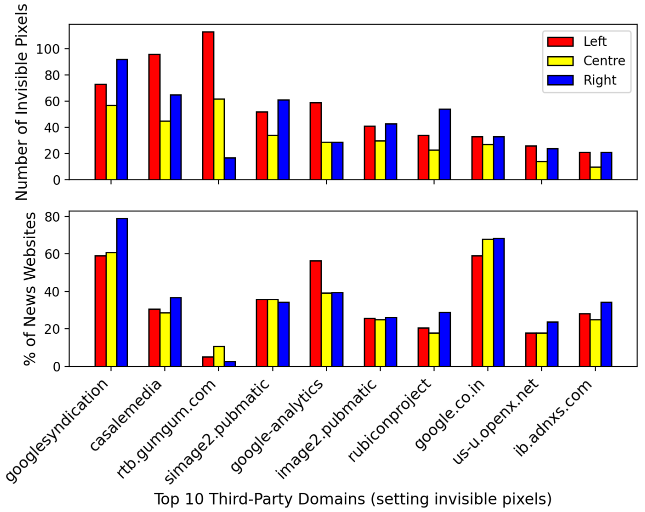

Figure 10: Top 10 Third-Party Domains setting invisible pixels on first-parties. Upper figure: total number of pixels set. Bottom figure: % of websites embedding each third-party.

Findings on user tracking: Our study shows the extensive presence of cookies irrespective of a news website's partisan leanings: on average, over 100 cookies are placed by first (FP) and third parties (TP) when visiting any of the news media websites we studied. In general, more cookies are placed in the desktop than the mobile platforms. Right-leaning websites place 1.2x and 1.4x the number of cookies than Centre- and Left-leaning ones in the mobile platform, whereas in the case of desktop, it is the opposite: Left tracks more than Centre and Right. We also find that 68% of TPs exist in both mobile and desktop versions, allowing them to perform in-depth monitoring by linking users across multiple devices. When analyzing the categories of TPs, we find that Right- and Centre-leaning websites embed more advertising and fingerprinting TPs than Left-leaning ones. Also, the top TP *doubleclick.net* is present in 86% of FP news websites, showing the capability of one TP domain to dominate the tracking culture across all partisan news websites in India. Tracking with cookies goes beyond their mere presence on the browser. About one-fourth of FPs and one-fifth of TPs are involved in cookie synchronization (CS). We detect user IDs being synchronized close to six times (on average) between up to five parties, on average, depending on the type of syncing pair entity (TP-TP or FP-TP). We find that the Left-leaning websites and their TPs do more CS than Right- or Centre-leaning ones. Although around 20% of all websites use canvas fingerprinting for tracking purposes, there is little difference between Right and Left (Centre is somewhat less) here. In terms of invisible pixel-based tracking, TP domains in Centre-leaning websites track more than Left and the Left more than the Right. We note that the same top TPs in other tracking methods (cookies, CS etc.) are also at the top here: *Google* properties cover 60-80% of websites, underlining the domination of the tracking market by one entity.

Absence of Privacy Laws: "The Wild Tracking East". Our results on user tracking demonstrate that in the absence of explicit privacy laws in India, partisan websites employ different, and at times invasive tracking strategies to profile their visitors. Left-leaning websites set more cookies, do more CS, and more pixel-based tracking, and Left and Right are almost equally intense in terms of device fingerprinting. But what is interesting is the domination of just a few TPs that track across the studied news websites irrespective of their partisanship. With a reach of 77% of population from these 103 websites, the data tracked by one or few TP domains across partisan websites means that not only news websites, but even a handful of TP domains can play a very crucial role by serving political and other targeted ads.

Implications for Privacy: In India, if structured privacy laws are to come into effect, online user privacy must be given high importance. Methods of tracking currently in place can not only expose a user's website visits and browsing histories to the tracker, but also help tracking domains to aggregate the user's browsing patterns and interests. These can be used to generate in-depth, detailed profiles via data synchronization through separate channels, which in turn can be exploited in numerous ways beyond just showing targeted ads. In fact, the differential tracking across websites of different political leanings, and the opportunities offered by the above mechanics, can allow propagation of user profiles to a large number of trackers over the time. Therefore, there is scope for these profiles being used by vested groups for targeting a user and invading the user's privacy, with the potential to influence the users visiting news websites.

Future Work: The limitations of our present study along the following main lines can be tackled in future works:

1. Vernacular diversity: Our dataset was primarily focused on websites using English language (76/103 English, with 14/103 in Hindi and 13/103 in regional languages). However, the diversity of languages in this country (apart from Hindi and English, India has 22 scheduled languages and several state-based official languages) raises the question: Do different political leanings perform different type and intensity of tracking across languages and news websites representing them in the regional Indian space?

2. Wide & Complex Political Spectrum: Templates derived from the reference points and cases in Western settings can only partially explain the underlying political dynamics in India. Political parties in India typically defy linear binaries of Left and Right. In such a context, the coverage bias and media effects are variable and are contingent upon subject, personalities, and circumstances. While the categorizations herein of "Left" and "Right" have been used as a heuristic tool, future research should dive into the contextual specifics of Indian political lines, and offer analysis with finer granularity of the political spectrum.

3. Fake News & Hyper-partisanship: Recent rise in misinformation from online, hyper-partisan news websites serving fake news, coupled with tracking of users for better profiling and political ad delivery, erodes user trust in the online news ecosystem. It requires an in-depth study of the hyper-partisan Indian news websites to assess how political websites violate their visitors' privacy.

ACKNOWLEDGMENTS

N. Kourtellis has been partially supported by the European Union's Horizon 2020 Research and Innovation Programme under grant agreements No 830927 (Concordia), No 871793 (Accordion), and No 871370 (Pimcity). S. Set is a Marie-Sklodowska Curie (Global India ETN) Research Fellow at King's College London. Authors also thank Arpan Gupta for helping in crawling. These results reflect only the authors' findings and do not represent the views of their institutes/organisations.

REFERENCES

[1] Gunes Acar et al. 2014. The web never forgets: Persistent tracking mechanisms in the wild. In *Proc. CCS*. 674–689.

[2] Pushkal Agarwal et al. 2020. Stop tracking me bro! Differential tracking of user demographics on hyper-partisan websites. In *Proc. WWW*. 1479–1490.

[3] Alexa. 2018. Alexa Internet. Keyword Research, Competitor Analysis, and Website Ranking. Available at https://www.alexa.com, accessed on 11 May 2020.

[4] Alexa. 2018. Top Indian News Sites. Available at https://www.alexa.com/topsites/category/Top/News/Newspapers/Regional/India, accessed on 29 April 2020.

[5] Shweta Bhatt et al. 2018. Illuminating an ecosystem of partisan websites. In *Proc. WWW*. 545–554.

[6] Nataliia Bielova et al. 2020. Missed by Filter Lists: Detecting Unknown Third-Party Trackers with Invisible Pixels. *Proc. PETs* 2020, 2 (2020), 499–518.

[7] Reuben Binns et al. 2018. Measuring third-party tracker power across web and mobile. *ACM TOIT* 18, 4 (2018), 1–22.

[8] Sunandan Chakraborty et al. 2018. Political Tweets and Mainstream News Impact in India: A Mixed Methods Investigation into Political Outreach. In *Proc. ACM COMPASS* (Menlo Park and San Jose, CA, USA). Article 10, 11 pages. https://doi.org/10.1145/3209811.3209825

[9] Anupam Das and Ralph Schroeder. 2020. Online disinformation in the run-up to the Indian 2019 election. *Information, Communication & Society* (2020), 1–17.

[10] Steven Englehardt et al. 2015. Cookies that give you away: The surveillance implications of web tracking. In *Proc. WWW*. 289–299.

[11] Steven Englehardt and Arvind Narayanan. 2016. Online tracking: A 1-million-site measurement and analysis. In *Proc. CCS*. 1388–1401.

[12] Steven Englehardt and Arvind Narayanan. 2020. OpenWPM Framework. Available at https://github.com/mozilla/OpenWPM, accessed on 08 May 2020.

[13] Marjan Falahrastegar et al. 2016. Tracking personal identifiers across the web. In *International Conference on PAM*. Springer, 30–41.

[14] Feedspot. 2020. Top 100 Indian News Websites on the Web. Available at https://blog.feedspot.com/indian_news_websites/, accessed on 28 April 2020.

[15] Imane Fouad et al. 2018. Missed by Filter Lists: Detecting Unknown Third-Party Trackers with Invisible Pixels. *arXiv* (2018), arXiv–1812.

[16] R. Kelly Garrett et al. 2019. From Partisan Media to Misperception: Affective Polarization as Mediator. *Journal of Communication* (2019), 490–512.

[17] Genuinous. 2017. Cookies.txt Chrome Extension. Available at https://chrome.google.com/webstore/detail/cookiestxt/njabckikapfpffapmjgojcnbfjonfjfg?hl=en, accessed on 17 May 2020.

[18] Arpita Ghosh et al. 2015. To match or not to match: Economics of cookie matching in online advertising. *ACM TEAC* 3, 2 (2015), 1–18.

[19] Roberto Gonzalez et al. 2017. The cookie recipe: Untangling the use of cookies in the wild. In *TMA*. IEEE, 1–9.

[20] Xuehui Hu and Nishanth Sastry. 2020. What a Tangled Web We Weave: Understanding the Interconnectedness of the Third Party Cookie Ecosystem. In *Proc.*

WebScience.

[21] Global Web Index. 2019. Digital versus Traditional Media Consumption. Available at https://www.amic.media/media/files/file_352_2142.pdf, accessed on 01 June 2020.

[22] Sandhya Keelery. 2020. Trust in Media-2020. Available at https://www.statista.com/statistics/684946/media-trust-india/, accessed on 19 Oct 2020.

[23] Adam Lerner et al. 2016. Internet jones and the raiders of the lost trackers: An archaeological study of web tracking from 1996 to 2016. In *USENIX Security*.

[24] Tim Libert and Victor Pickard. 2015. Think you're reading the news for free? New research shows you're likely paying with your privacy. Available at https://theconversation.com/think-youre-reading-the-news-for-free-new-research-shows-youre-likely-paying-with-your-privacy-49694, accessed on 19 Oct 2020.

[25] Sangeeta Mahapatra and Johannes Plagemann. 2019. Polarisation and politicisation: the social media strategies of Indian political parties. (2019).

[26] Scott McCoy, Andrea Everard, Peter Polak, and Dennis F Galletta. 2007. The effects of online advertising. *Commun. ACM* 50, 3 (2007), 84–88.

[27] Dibyendu Mishra and Joyojeet Pal. 2020. Freedom of press and social media partisanship in India: A visualization of tweets around the 2020 FIR against The Wire. Available at http://joyojeet.people.si.umich.edu/freedom-of-press-and-social-media-partisanship-in-india, accessed on 17 Oct 2020.

[28] Media Ownership Monitor. 2020. Media Ownership Matters. Available at https://www.mom-rsf.org/, accessed on 01 June 2020.

[29] Keaton Mowery and Hovav Shacham. 2012. Pixel perfect: Fingerprinting canvas in HTML5. *Proceedings of W2SP* (2012), 1–12.

[30] Panagiotis Papadopoulos et al. 2017. If you are not paying for it, you are the product: How much do advertisers pay to reach you?. In *Proc. IMC*. ACM, 142–156.

[31] Panagiotis Papadopoulos et al. 2019. Cookie synchronization: Everything you always wanted to know but were afraid to ask. In *WWW*. 1432–1442.

[32] Stylianos Papathanassopoulos et al. 2013. Online threat, but television is still dominant: A comparative study of 11 nations' news consumption. *Journalism Practice* 7, 6 (2013), 690–704.

[33] Adnan Qayyum et al. 2018. Exploring media bias and toxicity in south asian political discourse. In *ICOSST*. IEEE, 01–08.

[34] Reuters India Digital News Report. 2019. Reuters India Digital News Report. Available at https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2019-03/India_DNR_FINAL.pdf, accessed on 17 Oct 2020.

[35] Till Speicher et al. 2018. Potential for Discrimination in Online Targeted Advertising. In *Proc. FAT*, Vol. 81. 1–15.

[36] Tobias Urban et al. 2020. Measuring the Impact of the GDPR on Data Sharing in Ad Networks. In *Proc. ACM ASIA CCS*.

[37] Narseo Vallina-Rodriguez et al. 2016. Tracking the trackers: Towards understanding the mobile advertising and tracking ecosystem. *arXiv preprint arXiv:1609.07190* (2016).

[38] Chris J. Vargo and Lei Guo. 2017. Networks, Big Data, and Intermedia Agenda Setting: An Analysis of Traditional, Partisan, and Emerging Online U.S. News. *Journalism & Mass Communication Quarterly* (2017), 1031–1055.

[39] Jonathan White and Lea Ypi. 2016. *The Meaning of Partisanship*. Oxford University Press.