Under the Spotlight: Web Tracking in Indian Partisan News Websites

Vibhor Agarwal,[*1] **Yash Vekaria,**[*1]
Pushkal Agarwal,[2] **Sangeeta Mahapatra,**[3] **Shounak Set,**[2]
Sakthi Balan Muthiah,[1] **Nishanth Sastry,**[4] **Nicolas Kourtellis**[5]
[1]The LNM Institute of Information Technology, Jaipur, India
[2]King's College London, London, United Kingdom
[3]German Institute for Global and Area Studies, Hamburg, Germany
[4]University of Surrey, Surrey, United Kingdom
[5]Telefonica Research, Barcelona, Spain
{vibhor.agarwal.y16, yash.vekaria.y16, sakthi.balan}@lnmiit.ac.in, {pushkal.agarwal, shounak.set}@kcl.ac.uk,
sangeeta.mahapatra@giga-hamburg.de, n.sastry@surrey.ac.uk, 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, 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 synchronization, 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 demonstrates the intensity of Web tracking happening in Indian news websites and discusses implications for research on overall privacy of users visiting partisan news websites in India.

1 Introduction

India represents the largest and most diverse news media market among democracies, with more than 100,000 registered newspapers and 400 news channels (Indian Television 2020) in English, Hindi, and several regional languages (RNI 2020). The growth of online news in India has been the fastest in emerging markets, with India ranking among the top ten globally when it comes to print and online news media (Index 2019). Unfortunately, this growth of online political communications has been accompanied by rising partisanship (Das and Schroeder 2020; Mahapatra and Plagemann 2019; Verma and Sardesai 2014).

Unlike fringe media, with their influence limited to specific segments of readers, the mainstream media is a major determiner of the quality of information being generated, disseminated, and internalized by citizens across the country. Partisan media generate biased information (Jamieson and Cappella 2008; Karamshuk et al. 2016), which in turn, may harden the biases of readers, dividing the society and limiting the scope for open discussions and policy consensus (Levendusky 2013). A common identifier of partisan media is their ideological slant as Left, Right, or Centre. This determines their priorities on generating 'news' for public consumption (Shultziner and Stukalin 2020).

Another concern is that currently, there is a lack of privacy laws in India, which allows websites to track readers. India's Personal Data Protection Bill (PDPB) is yet to be enacted and in its current form, is considered problematic by some, as it allows the government to access personal data under vaguely described circumstances (PRS India 2019; Carnegie India 2020; Wired 2020). India's privacy regime, at present, is broadly guided by the Constitution's Articles 14, 19, and 21 on the freedom of speech, expression, and liberty, a 2017 Supreme Court judgement on privacy as a fundamental right, and the Information Technology Act (ITA), 2000; 2008 (CIS India 2018). The ITA has critical gaps as it provides scope for government surveillance and does not address user's rights to be notified of the presence of cookies and do-not track options, and allows the use of electronic personal identifiers across databases (DW.COM 2020). Therefore, unregulated tracking can easily take place (e.g., as has been done before the 2019 elections in India (Singh 2019)), for user profiling and micro-targeting particular population segments.

The simultaneous growth of partisanship and the digital presence of mainstream media in a privacy-poor environment raises the possibility of readers being tracked for not just commercial reasons but also for political targeting. Our aim is to study the extent to which readers are tracked along

[*]*Equal Contribution*

partisan lines and provide a starting point for more specific studies on the nature and purposes of targeting.

There have been US-based studies about partisan media mostly in terms of their polarizing effects (Garrett, Long, and Jeong 2019; Vargo and Guo 2017; Bhatt et al. 2018) and a few on tracking (Libert and Pickard 2015; Agarwal et al. 2020b). For India, while there have been a few works on the division in the news media along partisan lines (Mishra and Pal 2020), there is a lack of comprehensive, data-driven research on news websites and tracking behavior. To the best of our knowledge, this is the first such study in India, one of the largest media markets and the world's largest democracy.

In this work, 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). India has a fragmented media space with some publications having an out-sized reach compared to others. This is true for print newspapers, especially those in local languages like the Hindi language press, which can reach 40% of the population covering rural and urban readers (Neyazi 2018). The digital-only websites in the study sample have largely grown in the past few years, targeting a growing demographic of readers who use mobile phones for getting news (Reuters Institute 2019). This study, therefore, includes both traditional news publications and emerging news publications to provide a comprehensive coverage. Collectively, the websites we study can reach more than 77% of India's population (Media Research Users Council 2019; BARC India 2020), 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 (Alexa 2018a) and Feedspot (Feedspot 2020). 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 (Monitor 2020), we label the 103 websites according to their partisanship as Right-, Left-, Centre-leaning, or Unknown (Methodology explained in Section 3).

With this data, we address the following questions:

RQ1: What is the extent of tracking across partisan news websites in India?

RQ2: What kind of tracking methods are commonly employed on such websites?

To answer these questions, we measure the intensity of user tracking across partisan websites with simple and advanced mechanisms: basic first-party (FP) and third-party (TP) cookies, cookie synchronization (CS), device fingerprinting, and invisible pixel-based tracking (Section 4). We publicly share our website list, crawls and code with the research community for reproducibility and extensions[1].

From our measurements, 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 higher number of cookies than average. For example, $\sim$1400 cookies are set on the FP *Sandesh.com*, by itself and its TPs. Left- and Centre-leaning websites tend to serve more cookies than Right. Desktop versions of websites set more cookies than their mobile versions, with interesting exceptions. The TP 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 TPs, and 25% of all distinct FPs 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 TPs 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 the partisan nature of Indian news websites and 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 (White and Ypi 2016). 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 (McCoy et al. 2007; Papadopoulos et al. 2017). These ads are strategically drafted and placed on websites to get more user at-

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

tention including pop-ups and banners (McCoy et al. 2007; Speicher et al. 2018).

News and other websites inject cookies in the users' browsers for content personalization and improving user experience. However, third parties commonly inject cookies to track users, raising privacy concerns (Englehardt et al. 2015; Binns et al. 2018; Vallina-Rodriguez et al. 2016; Hu and Sastry 2020; Hu, Suarez-Tangil, and Sastry 2020). Among Alexa top 20K sites, it has been found that necessary and functional cookies constitute less than 13% and 10% of the cookies respectively, with the remaining being advertising and analytics cookies (Hu, Sastry, and Mondal 2021).

Many websites also use more sophisticated tracking techniques like cookie synchronization (Acar et al. 2014; Englehardt and Narayanan 2016; Papadopoulos, Kourtellis, and Markatos 2019; Agarwal et al. 2020b; Urban et al. 2020; Hu and Sastry 2020), device fingerprinting (Mowery and Shacham 2012; Englehardt and Narayanan 2016), and invisible (1x1) pixel-based tracking (Fouad et al. 2020). 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-owned) can be present in up to 50% and 70%, respectively, of top one million visited websites (Englehardt and Narayanan 2016). Specifically, news websites have seen large volume of trackers and advertisements including political campaigns (Englehardt and Narayanan 2016; Agarwal et al. 2020b; Papathanassopoulos et al. 2013). Among USA news websites, Right-leaning websites track users more and have high cookie synchronization within the partisan group websites (Agarwal et al. 2020b). Having said that, less is known about the tracking ecosystem of Indian news media homepages, or their topical subpages (Vekaria et al. 2021). 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 (Mahapatra and Plagemann 2019; Starbird 2017; Qayyum et al. 2018). 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* (Feedspot 2020) and *Alexa.com* (Alexa 2018a), to label these websites based on their political leaning (Sec. 3.1). In Sec. 3.2, we provide details of our website data crawling using *OpenWPM* (Englehardt and Narayanan 2016, 2020), a tool for desktop browser automation and crawling, and *Cookies.txt* (Genuinous 2017), 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 (Feedspot 2020). 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) (Alexa 2018b). 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.

A large portion of news consumption in India happens through online platforms (Facebook, Twitter, and Instagram) rather than TV/Radio (Reuters Institute 2019). 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. The readership data are derived from the circulation figures of the Registrar of Newspapers for India, Government of India, and the data of the Indian Readership Survey (RNI 2020) conducted by the Media Research Users Council(Media Research Users Council 2019) and the Readership Studies Council of India. Viewership data are obtained from the social media pages of the studied news media alongside figures of the Broadcast Audience Research Council of India (BARC India 2020) for those that have television channels.

Website Labeling: In order to understand and categorize

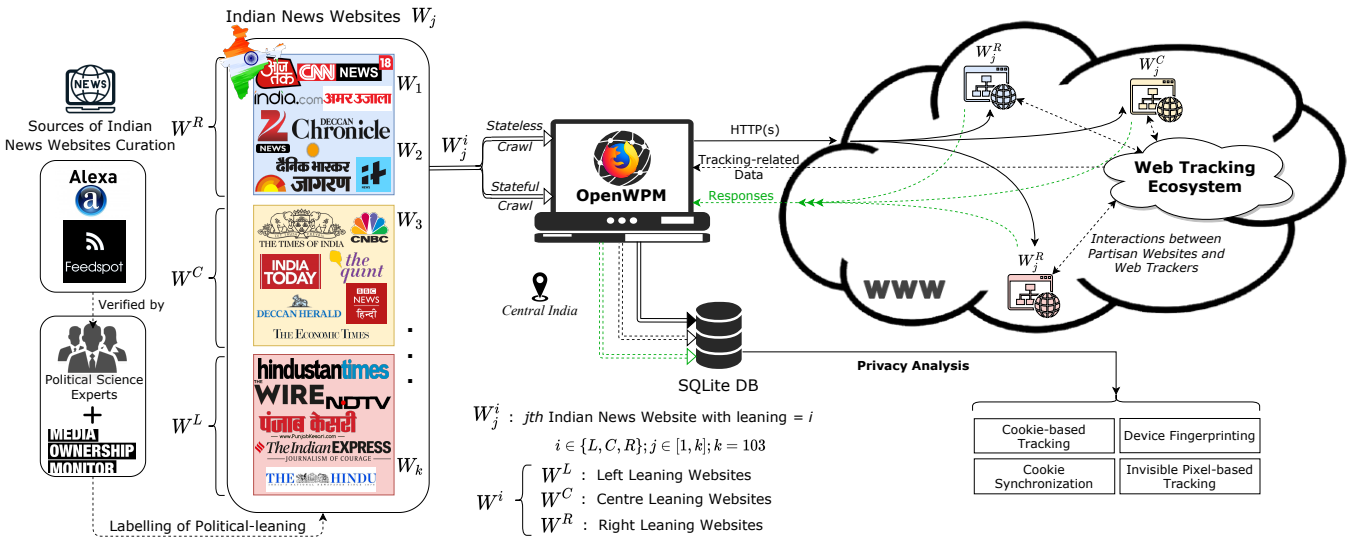

Figure 1: Our framework for labeling Indian news websites along partisan lines and collecting web traffic data for studying web tracking mechanisms. Colors represent party-leaning: Right = Blue, Centre = Yellow, and Left = Red.

websites based on their partisan leaning, we undertook a three-step labeling process. First, we approached two political science and journalism experts who manually coded the political leaning of each website. This approach has been used by media monitors at Buzzfeed News in past studies to review political leaning in the US news ecosystem (Bhatt et al. 2018). Second, the manual partisan associations of these websites labeled by the experts was further validated by looking at sources like the Media Ownership Monitor (Monitor 2020) in India, initiated by Reporters without Borders and conducted by the Delhi-based company, DataLEADS.

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, 20 websites were discarded because there was no clear partisanship indicator based on the publication's ownership and content and the experts could not label them clearly. The remaining 103 websites were labeled with a partisan leaning and considered in 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 *39* Left-, *27* Centre-, and *37* Right-leaning websites.

3.2 Websites OpenWPM Crawling

Our data collection used OpenWPM (Englehardt and Narayanan 2016), 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 in future visits and with other websites and their TPs. 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 TPs 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 (Genuinous 2017) to get browser cookies information. We automate this process using Selenium (Selenium 2013). First, a Firefox browser is set to not block any type of cookies. Further steps include opening a Firefox Mobile Emulator in 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 (Disconnect, Inc. 2013), 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)[2]. Past works have studied CS in different contexts (Acar et al. 2014; Falahrastegar et al. 2016; Englehardt and Narayanan 2016; Papadopoulos, Kourtellis, and Markatos 2019; Agarwal et al. 2020b; Urban et al. 2020; Hu and Sastry 2020; Hu, Suarez-Tangil, and Sastry 2020)). 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 (Gonzalez et al. 2017) through the second step. In this 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 (Lerner et al. 2016), especially in programmatic (real-time bidding) advertising, where data sharing and purchasing involves CS for better targeting (Ghosh et al. 2015). 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 (Englehardt and Narayanan 2016; Papadopoulos, Kourtellis, and Markatos 2019), enabling deeper user profiling.

Methodology: We capture CS for websites in our dataset using similar methodology of past studies (Acar et al. 2014; Falahrastegar et al. 2016; Papadopoulos, Kourtellis, and Markatos 2019). We use the fundamental structure of the open-source python code from (Acar et al. 2014) (referred to as `CSCode` hereafter) and make modifications to work for our scenario: unlike (Acar et al. 2014) that crawled data simultaneously on two machines before analyzing them with `CSCode`, we perform time-variant crawls (Section 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)` $\equiv$ `(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 (Acar et al. 2014) is an older paper on CS, we validated `CSCode`, as well as various parameters used with recent works on CS (Papadopoulos, Kourtellis, and Markatos 2019; Agarwal et al. 2020b; Urban et al. 2020; Hu and Sastry 2020)). 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 (Acar et al. 2014). However, it is not relevant to study CS across such top-level domains. Instead, we follow (Papadopoulos, Kourtellis, and Markatos 2019) and map all domains (from cookies, requests, response URLs, etc.) to the high-level domains returned by the WhoIS tool (Lookup 2020) (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, (Urban et al. 2020) suggests to discard shorter IDs, since they do not contain sufficient entropy to represent a user ID. We follow (Papadopoulos, Kourtellis, and Markatos 2019) 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 related dependencies.

Limitations: `CSCode` gives a strict conservative ID detection with fewer false positives (Acar et al. 2014). However, false negatives may occur when an ID is shared in URL parameters in an encoded or encrypted format (Papadopoulos, Kourtellis, and Markatos 2019; Fouad et al. 2020), or when ID strings are hidden inside the longer strings with non-standard delimiters. According to (Acar et al. 2014), the adversarial trackers could have short-lived cookies[3] mapped to user IDs at the backend-server to later on track the user. Such cases are not captured by our code. 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

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

[3] Like (Acar et al. 2014), we consider cookies with expiration date $\leq$ 30 days

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, and can have important implications on the way TPs track users across the Web *without cookies* in the future (Papadogiannakis et al. 2021).

Methodology: Our fingerprinting measurement methodology (Englehardt and Narayanan 2016) utilizes data collected by OpenWPM, as described in Section 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 (Fouad et al. 2020), 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 (Fouad et al. 2020), 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 TP 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 (Section 5.1). We then study more complex tracking techniques such as cookie synchronization (Section 5.2), device fingerprinting (Section 5.3), and invisible pixel-based tracking (Section 5.4).

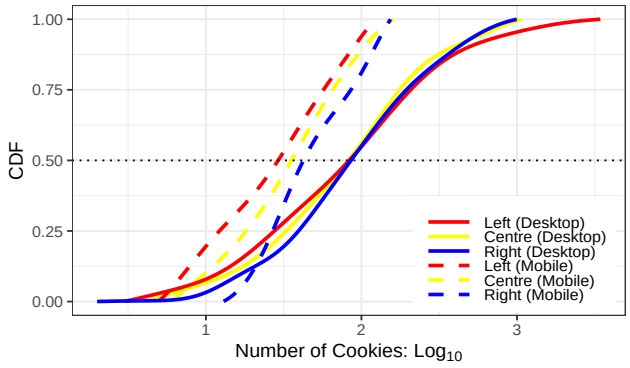

Figure 2: CDF of number of cookies for Left, Centre, and Right-leaning news websites, for their desktop and mobile versions (where available).

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 fewer cookies than the median, the trend is reversed, i.e., Right-leaning websites track more than Left and Centre-leaning.

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. We speculate higher cookies on desktop counterparts due to several reasons such as: 1) the desktop-based tracking ecosystem is more evolved, since there was traditionally more news consumption on desktop than mo-

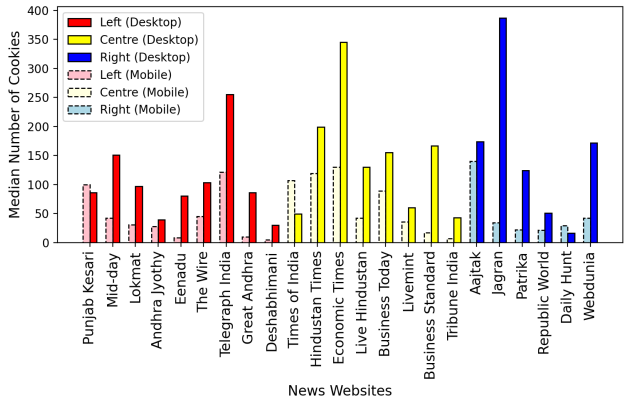

Figure 3: Median number of cookies in mobile vs. desktop versions for 23 news websites, grouped by political leaning in decreasing order of their Facebook followers.

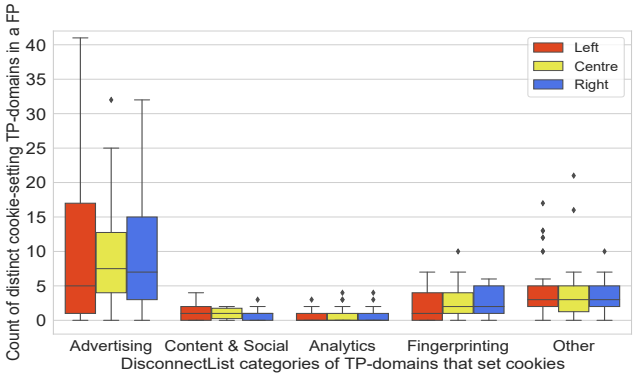

Figure 4: For each FP, the distribution of count of distinct cookie-setting TPs by DisconnectList categories.

bile; however this is changing over the last years, 2) mobiles have fewer resources (including storage, battery and bandwidth) and mobile-based websites are more careful and wary of using these resources intensely and making their webpages "heavy"; instead they try to make them mobile-friendly. Therefore, such (mobile) websites 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 example, 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

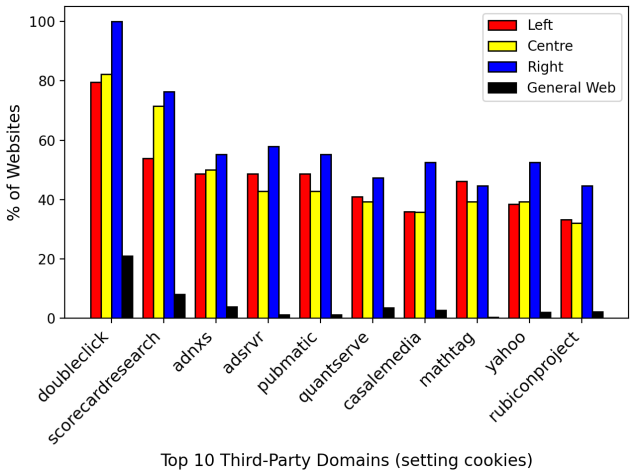

Figure 5: Top 10 TP domains setting cookies in Left, Centre, or Right-leaning news websites. Their presence on general web is also plotted for comparison.

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-leaning websites performing more intense fingerprinting than Left-leaning. We investigate such domains further in Sec. 5.3.

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 FP 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 (Agarwal et al. 2020b). For general Web, we crawl data from *whotracks.me*, the percentage of websites in which detected TPs 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.

Table 1: Statistics on cookie synchronizations detected between FP and 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**

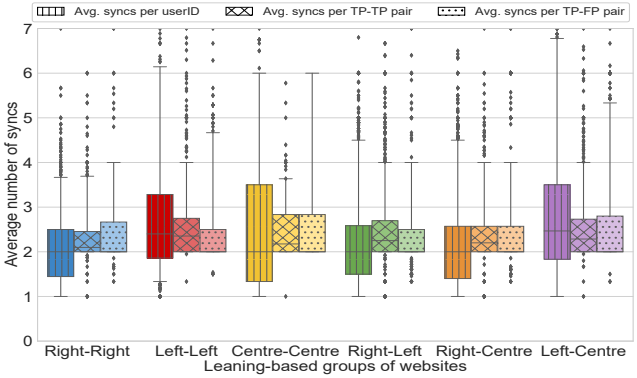

Figure 6: Distributions of average number of CSs per ID, with respect to political leaning groups and combinations.

5.2 Cookie Synchronization

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

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-test: 0.0748 (p-value: 0.0029), 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 synced across 24 different tracking domains (from ∼600+ TPs in our data). Moreover, a higher median number of TPs is 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.

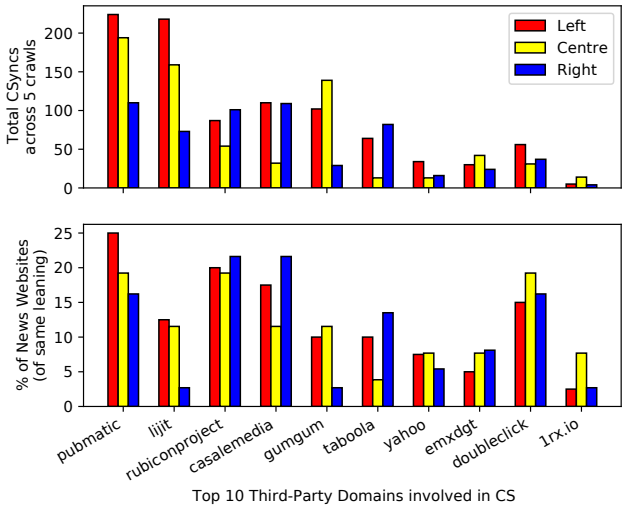

Figure 7: Top 10 TPs involved in CSs, grouped by political leaning. Total CSs (top y-axis) is (TP-TP)+(TP-FP) CSs.

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 websites.

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 Section 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 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 equally adopting this tracking technology.

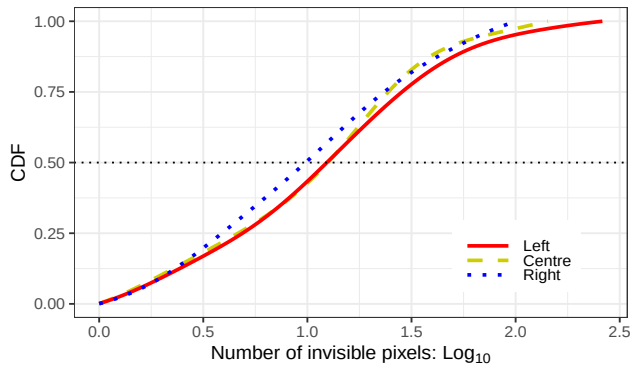

Figure 8: CDF of median number of invisible pixels for Left, Centre, and Right-leaning websites.

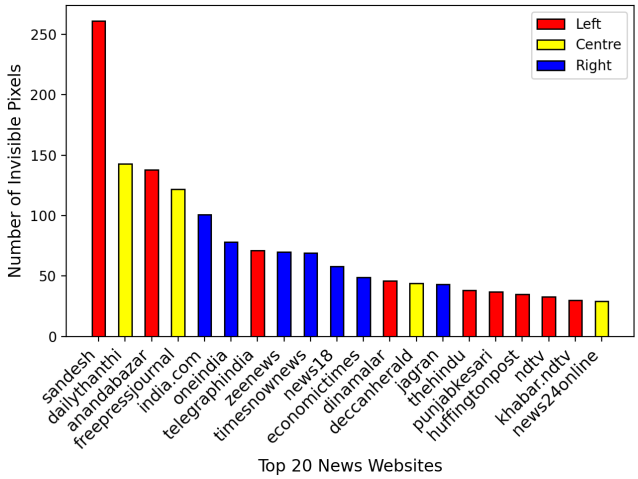

Figure 9: Top 20 news websites having invisible pixels vs. their political leanings.

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 Section 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 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 TPs, earlier found to set most cookies, has the highest number of detected invisible pixels (261). In general, 138 TPs are detected setting these 2,513 1x1 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

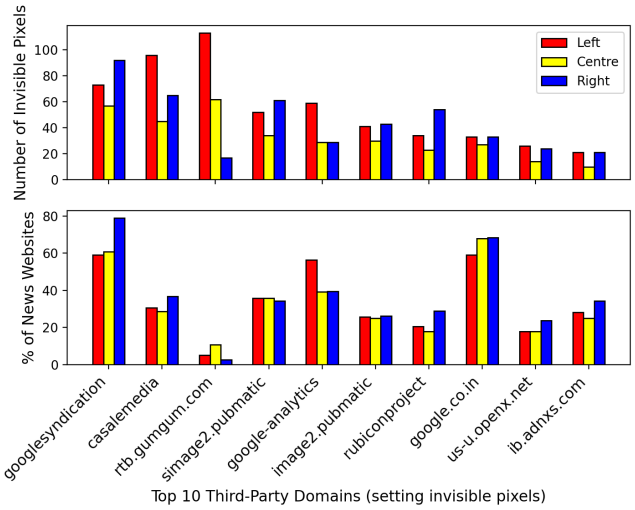

Figure 10: Top 10 TP Domains setting invisible pixels on FPs. Upper figure: total number of pixels set. Bottom figure: % of websites embedding each TP.

largest cumulative TP 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 of 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.

Findings on user tracking: Our study shows the extensive presence of cookies irrespective of a news website's partisan leaning: 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 top TPs in other tracking methods (cookies, CS, etc.) are also 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.

Tracking in Wild East (India) vs. West (USA). When comparing India's ad-tracking ecosystem with USA's, we find interesting differences. India tracks with cookies similarly in both Left- and Right-leaning websites (median: 86 vs. 84, respectively), whereas in USA, Right-leaning websites clearly track with more cookies than Left (median: 21 vs. 14) (Agarwal et al. 2020b). Interestingly, the median cookies in India (Left or Right) are ∼4-5x more than USA, revealing an aggressive effort from Indian websites and their TPs to track users. On the other hand, considering CS, we find an opposite trend. Indian Left websites perform more (∼1.8x) CS than Right (average CSs: 4.7 vs. 2.6, respectively). However, in USA, Left and Right websites perform similar amount of CS (average CSs: 12 vs. 13, respectively). When comparing the number of CS in India (Left or Right) with USA, we find that it is ∼2-5x less. Given that CS is a more advanced method of tracking than simple cookies, we explain these counter-intuitive findings as USA's ad-tracking being more efficient (needs fewer cookies), but more intense and effective in advanced methods of tracking (more CS).

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). Multilingual online users consist of a large portion in India (Agarwal et al. 2020a). 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: Are the patterns of tracking similar or different among regional Indian News websites?

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.

7 Acknowledgments

N. Sastry acknowledges support via EPSRC Grant Ref: EP/T001569/1 for "Artificial Intelligence for Science, Engineering, Health and Government", and particularly the "Tools, Practices and Systems" theme for "Detecting and Understanding Harmful Content Online: A Metatool Approach". 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). P. Agarwal is supported by a Sir Rick Trainor Scholarship at King's College London. S. Set is a Marie-Sklodowska Curie

(Global India ETN) Research Fellow at King's College London. Authors also thank Arpan Gupta (The LNMIIT, Jaipur) for helping in crawling. These results reflect only the authors' findings and do not represent the views of their institutes/organisations.

References

Acar, G.; Eubank, C.; Englehardt, S.; Juarez, M.; Narayanan, A.; and Diaz, C. 2014. The web never forgets: Persistent tracking mechanisms in the wild. In *Proceedings of the ACM SIGSAC CCS*, 674–689.

Agarwal, P.; Garimella, K.; Joglekar, S.; Sastry, N.; and Tyson, G. 2020a. Characterising user content on a multilingual social network. In *Proceedings of the AAAI ICWSM 2020*, volume 14, 2–11.

Agarwal, P.; Joglekar, S.; Papadopoulos, P.; Sastry, N.; and Kourtellis, N. 2020b. Stop tracking me bro! differential tracking of user demographics on hyper-partisan websites. In *Proceedings of ACM WWW*, 1479–1490.

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

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

BARC India. 2020. Broadcast Audience Research Council. Available at barcindia.co.in/, accessed on 17 Oct 2020.

Bhatt, S.; Joglekar, S.; Bano, S.; and Sastry, N. 2018. Illuminating an ecosystem of partisan websites. In *Companion Proceedings of WWW 2018*, 545–554.

Binns, R.; Zhao, J.; Kleek, M. V.; and Shadbolt, N. 2018. Measuring third-party tracker power across web and mobile. *ACM TOIT* 18(4): 1–22.

Carnegie India. 2020. Will India's proposed data protection law protect privacy and promote growth? Available at https://carnegieindia.org/2020/03/09/will-india-s-proposed-data-protection-law-protect-privacy-and-promote-growth-pub-81217, accessed on 17 Oct 2020.

CIS India. 2018. Internet privacy in India. Available at https://cis-india.org/telecom/knowledge-repository-on-internet-access/internet-privacy-in-india, accessed on 17 Oct 2020.

Das, A.; and Schroeder, R. 2020. Online disinformation in the run-up to the Indian 2019 election. *Information, Communication & Society* 1–17.

Disconnect, Inc. 2013. DisconnectList - Disconnect Tracking Protection Project. Available at https://github.com/disconnectme/disconnect-tracking-protection, accessed on 24 June 2020.

DW.COM. 2020. How India's loose data privacy laws open the door to hackers? Available at https://www.dw.com/en/how-indias-loose-data-privacy-laws-open-the-door-to-hackers/a-53120972, accessed on 17 Oct 2020.

Englehardt, S.; and Narayanan, A. 2016. Online tracking: A 1-million-site measurement and analysis. In *Proceedings of the ACM SIGSAC CCS*, 1388–1401.

Englehardt, S.; and Narayanan, A. 2020. OpenWPM Framework. Available at https://github.com/mozilla/OpenWPM, accessed on 08 May 2020.

Englehardt, S.; Reisman, D.; Eubank, C.; Zimmerman, P.; Mayer, J.; Narayanan, A.; and Felten, E. W. 2015. Cookies that give you away: The surveillance implications of web tracking. In *Proceedings of the WWW*, 289–299.

Falahrastegar, M.; Haddadi, H.; Uhlig, S.; and Mortier, R. 2016. Tracking personal identifiers across the web. In *ACM PAM*, 30–41. Springer.

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

Fouad, I.; Bielova, N.; Legout, A.; and Sarafijanovic-Djukic, N. 2020. Missed by Filter Lists: Detecting Unknown Third-Party Trackers with Invisible Pixels. In *PETS*.

Garrett, R. K.; Long, J. A.; and Jeong, M. S. 2019. From partisan media to misperception: Affective polarization as Mediator. *Journal of Communication* 69(5): 490–512.

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

Ghosh, A.; Mahdian, M.; McAfee, R. P.; and Vassilvitskii, S. 2015. To match or not to match: Economics of cookie matching in online advertising. *ACM TEAC* 3(2): 1–18.

Gonzalez, R.; Jiang, L.; Ahmed, M.; Marciel, M.; Cuevas, R.; Metwalley, H.; and Niccolini, S. 2017. The cookie recipe: Untangling the use of cookies in the wild. In *IFIP TMA*, 1–9. IEEE.

Hu, X.; and Sastry, N. 2020. What a Tangled Web We Weave: Understanding the Interconnectedness of the Third Party Cookie Ecosystem. In *ACM Web Science*.

Hu, X.; Sastry, N.; and Mondal, M. 2021. CCCC: Corralling Cookies into Categories with CookieMonster. In *ACM Web Science*.

Hu, X.; Suarez-Tangil, G.; and Sastry, N. 2020. Multi-country Study of Third Party Trackers from Real Browser Histories. In *IEEE EuroS&P*, 70–86. IEEE.

Index, G. W. 2019. Digital versus Traditional Media Consumption. Available at https://www.amic.media/media/files/file_352_2142.pdf, accessed on 01 June 2020.

Indian Television. 2020. Total television channels in India. Available at https://www.indiantelevision.com/regulators/ib-ministry/total-of-television-channels-in-india-rises-to-892-with-three-cleared-in-june-160709, accessed on 01 June 2020.

Jamieson, K. H.; and Cappella, J. N. 2008. *Echo chamber: Rush Limbaugh and the conservative media establishment*. Oxford University Press.

Karamshuk, D.; Lokot, T.; Pryymak, O.; and Sastry, N. 2016. Identifying partisan slant in news articles and twitter during political crises. In *SocInfo*, 257–272. Springer.

Lerner, A.; Simpson, A. K.; Kohno, T.; and Roesner, F. 2016. Internet jones and the raiders of the lost trackers: An archaeological study of web tracking from 1996 to 2016. In *25th USENIX Security Symposium*.

Levendusky, M. S. 2013. Why do partisan media polarize viewers? *American Journal of Political Science* .

Libert, T.; and Pickard, V. 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.

Lookup, W. D. 2020. Top 100 Indian News Websites on the Web. Available at https://www.whois.com/whois/, accessed on 06 May 2020.

Mahapatra, S.; and Plagemann, J. 2019. Polarisation and politicisation: the social media strategies of Indian political parties. *DEU* .

McCoy, S.; Everard, A.; Polak, P.; and Galletta, D. F. 2007. The effects of online advertising. *Communications of the ACM* 50(3): 84–88.

Media Research Users Council. 2019. Media Research Users Council, Indian Readership Survey. Available at https://bestmediainfo.in/mailer/nl/nl/IRS-2019-Q4-Highlights.pdf, accessed on 17 Oct 2020.

Mishra, D.; and Pal, J. 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.

Monitor, M. O. 2020. Media Ownership Matters. Available at https://india.mom-rsf.org/, accessed on 01 June 2020.

Mowery, K.; and Shacham, H. 2012. Pixel perfect: Fingerprinting canvas in HTML5. *Proceedings of W2SP* 1–12.

Neyazi, T. A. 2018. *Political communication and mobilisation: The Hindi media in India*. Cambridge University Press.

Papadogiannakis, E.; Papadopoulos, P.; Kourtellis, N.; and Markatos, E. P. 2021. User Tracking in the Post-cookie Era: How Websites Bypass GDPR Consent to Track Users. In *Proceedings of the WWW*.

Papadopoulos, P.; Kourtellis, N.; and Markatos, E. 2019. Cookie synchronization: Everything you always wanted to know but were afraid to ask. In *Proceedings of the ACM WWW*, 1432–1442.

Papadopoulos, P.; Kourtellis, N.; Rodriguez, P. R.; and Laoutaris, N. 2017. If you are not paying for it, you are the product: How much do advertisers pay to reach you? In *Proceedings of the ACM IMC*, 142–156.

Papathanassopoulos, S.; Coen, S.; Curran, J.; Aalberg, T.; Rowe, D.; Jones, P.; Rojas, H.; and Tiffen, R. 2013. Online threat, but television is still dominant: A comparative study of 11 nations' news consumption. *Journalism Practice* 7(6): 690–704.

PRS India. 2019. PDP Bill. Available at https://www.prsindia.org/billtrack/personal-data-protection-bill-2019/, accessed on 02 June 2020.

Qayyum, A.; Gilani, Z.; Latif, S.; and Qadir, J. 2018. Exploring media bias and toxicity in south asian political discourse. In *12th ICOSST*, 01–08. IEEE.

Reuters Institute. 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.

RNI. 2020. Registrar of Newspaper for India. Available at http://rni.nic.in/, accessed on 19 Oct 2020.

Selenium. 2013. The Selenium Browser Automation Project. Available at https://www.selenium.dev/documentation/en/, accessed on 17 May 2020.

Shultziner, D.; and Stukalin, Y. 2020. Politicizing What's News: How Partisan Media Bias Occurs in News Production. *Mass Communication and Society* 1–22.

Singh, S. S. 2019. *How to Win an Indian Election: What Political Parties Don't Want You to Know*. Penguin Random House India Private Limited.

Speicher, T.; Ali, M.; Venkatadri, G.; Ribeiro, F. N.; Arvanitakis, G.; Benevenuto, F.; Gummadi, K. P.; Loiseau, P.; and Mislove, A. 2018. Potential for discrimination in online targeted advertising. In *FAccT*, 5–19. PMLR.

Starbird, K. 2017. Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on Twitter. In *AAAI ICWSM*.

Urban, T.; Tatang, D.; Degeling, M.; Holz, T.; and Pohlmann, N. 2020. Measuring the impact of the GDPR on data sharing in ad networks. In *15th ACM ASIACCS*.

Vallina-Rodriguez, N.; Sundaresan, S.; Razaghpanah, A.; Nithyanand, R.; Allman, M.; Kreibich, C.; and Gill, P. 2016. Tracking the Trackers: Towards Understanding the Mobile Advertising and Tracking Ecosystem. In *Proceedings of Workshop on Data and Algorithmic Transparency*.

Vargo, C. J.; and Guo, L. 2017. Networks, big data, and intermedia agenda setting: An analysis of traditional, partisan, and emerging online US news. *Journalism & Mass Communication Quarterly* 94(4): 1031–1055.

Vekaria, Y.; Agarwal, V.; Agarwal, P.; Mahapatra, S.; Muthiah, S. B.; Sastry, N.; and Kourtellis, N. 2021. Differential Tracking Across Topical Webpages of Indian News Media. In *Proceedings of the ACM WebSci*.

Verma, R.; and Sardesai, S. 2014. Does media exposure affect voting behaviour and political preferences in India? *Economic and Political Weekly* 82–88.

White, J.; and Ypi, L. 2016. *The meaning of partisanship*. Oxford University Press.

Wired. 2020. India's data protection bill threatens global cybersecurity. Available at www.wired.com/story/opinion-indias-data-protection-bill-threatens-global-cybersecurity/, accessed on 17 Oct 2020.