Good News for People Who Love Bad News: Centralization, Privacy, and Transparency on US News Sites

Timothy Libert
Carnegie Mellon University
timlibert@cmu.edu

Reuben Binns
University of Oxford
reuben.binns@cs.ox.ac.uk

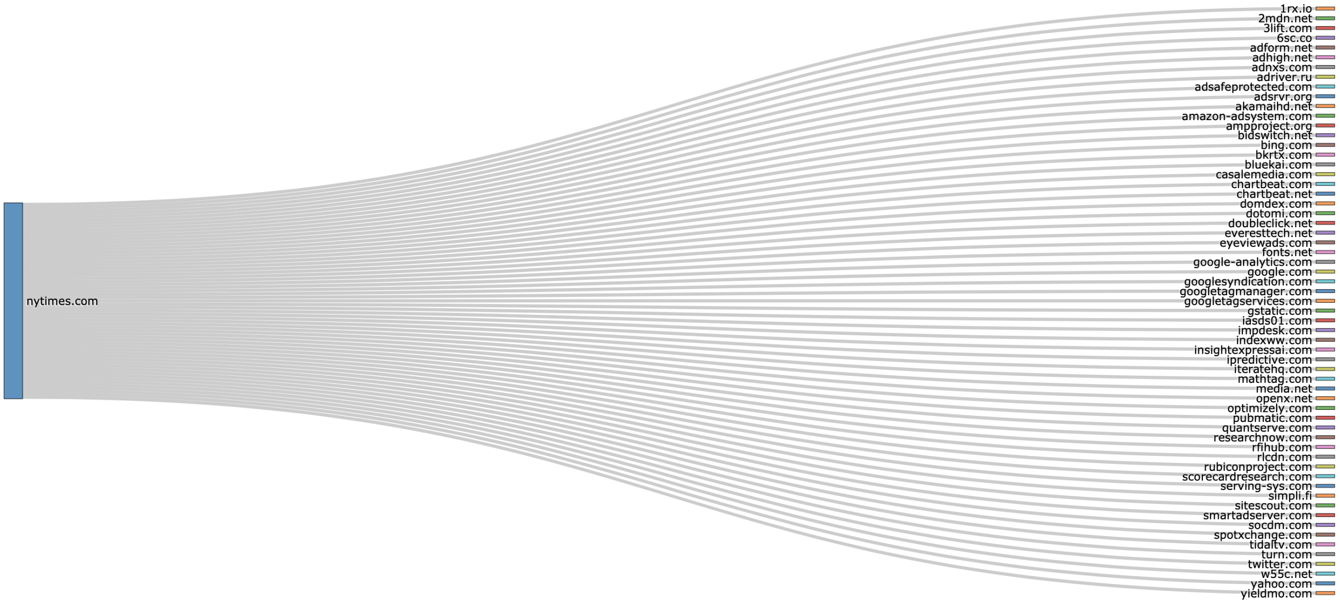

Figure 1: The New York Times homepage exposes visitors to 61 third-party domains.

ABSTRACT

The democratic role of the press relies on maintaining independence, ensuring citizens can access controversial materials without fear of persecution, and promoting transparency. However, as news has moved to the web, reliance on third-parties has centralized revenue and hosting infrastructure, fostered an environment of pervasive surveillance, and lead to widespread adoption of opaque and poorly-disclosed tracking practices.

In this study, 4,000 US-based news sites, 4,000 non-news sites, and privacy policies for 1,892 news sites and 2,194 non-news sites are examined. We find news sites are more reliant on third-parties than non-news sites, user privacy is compromised to a greater degree on news sites, and privacy policies lack transparency in regards to observed tracking behaviors. Overall, findings indicate the democratic role of the press is being undermined by reliance on the "surveillance capitalism" funding model.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
WebSci '19, June 30-July 3, 2019, Boston, MA, USA
© 2019 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-6202-3/19/06.
https://doi.org/10.1145/3292522.3326019

CCS CONCEPTS

• **Security and privacy → Human and societal aspects of security and privacy**; **Usability in security and privacy**;

KEYWORDS

Web; Privacy; Security; Tracking; News Media

ACM Reference format:
Timothy Libert and Reuben Binns. 2019. Good News for People Who Love Bad News: Centralization, Privacy, and Transparency on US News Sites. In *Proceedings of 11th ACM Conference on Web Science, Boston, MA, USA, June 30-July 3, 2019 (WebSci '19),* 10 pages.
https://doi.org/10.1145/3292522.3326019

1 INTRODUCTION

News media in the United States has historically been decentralized and reliant upon a mixture of subscription and advertising revenue [31].[1][2] In legacy media such as print, radio, and television, advertisements are targeted at specific audiences only to the degree that given publications or programs are known to be popular with certain groups, such as young women, sports fans, or retirees. The

[1] Publicly-funded news media have a larger role in other Western democracies and the findings of this study are limited to the US market.
[2] The degree of centralization has increased over time due to mergers.

best means of determining the impact of advertisements are indirect measures of sales volume and brand awareness.

As news consumption has shifted to the web, subscription revenue has declined and advertisements are now primarily brokered by specialized advertising technology ("adtech") companies [26]. In contrast to legacy media, the web facilitates monitoring the actions of specific users, allowing advertisers to target messages based on inferences gleaned from "tracking" users as they browse the web, a process known as "online behavioral advertising" (OBA). The technological systems facilitating OBA are highly centralized, allowing a handful of companies to monitor the web browsing behaviors of billions of people and broker the flow of advertising revenue to millions of sites.

The most common way user behavior is monitored is via the inclusion of third-party services on web pages which initiate network connections between a user and a given third-party. Such connections often occur without user interaction and may expose users to persistent tracking carried out by cookies, browser fingerprints, and other identifiers. Prior research has determined that news websites contain significantly more behavioral tracking mechanisms than other types of sites [4, 7] and the news industry is reliant on a handful of adtech firms for revenue [26].

Beyond advertising, news sites may expose users to a range of third-parties that provide services for measuring the number of visitors to a page, recommending related articles, facilitating the sharing of articles on social media, and hosting content. From the perspective of the publisher, being able to target advertisements and offload the development of core site functions to outside parties makes economic sense: limited space on a given page may be used to display the most relevant advertisements, developer time may be spent on adding custom features rather than duplicating third-party services, and the complexities of hosting web pages may be delegated to cloud hosting companies.

While the centralization of advertising and hosting has a well-documented impact across the web [7, 19], the news sector represents a specific case for concern because the press serves an important democratic role in holding powerful actors to public account. There are three primary aspects of this role pertinent to today's adtech-driven web. First, as an independent social institution, the press should be free from outside influence and control [3, 5, 16]. Second, the press functions best when citizens are free to access information without fear of persecution: freedom to listen and read is as important as freedom to speak [33]. Third, the press must be transparent and honest so that citizens can have well-placed trust in the information they receive [16].

Reliance on third-parties compromises the above functions in several ways. First, while press outlets require independence to operate without influence, today's web fosters a centralization of both revenue and content-delivery infrastructure, which gives a handful of advertising and hosting firms massive unseen leverage over the press. This leverage has manifested itself in at least one known effort by Google to coerce a news outlet to include additional tracking code on their pages by asserting that not using the code would cause "search results [to] suffer" [13]. Second, citizens rely on *privacy* to enable them to safely seek out potentially controversial content [33] and web tracking directly undermines the privacy and security of readers. Research demonstrates that awareness of surveillance reduces citizens' comfort in seeking out information [22] and commenting on controversial topics [39]. Last, the essential nature of online advertising is premised on extracting user data in covert ways which run directly counter to the goal of transparency, potentially eroding the most essential resource of any news organization: trust.

To examine the impacts of third-parties on news sites, 4,000 US-based news sites are analyzed to determine how often users are exposed to third-party services, the privacy impacts of such exposure, and the nature of third-party services. To understand how news sites differ from other popular sites, an additional 4,000 popular non-news sites in the US are analyzed to provide a comparative benchmark. 12.5 million requests for third-party content and 3.4 million third-party cookies are examined to measure privacy impacts of several types of third-party services. Finally, 1,892 news and 2,194 non-news privacy policies are examined to determine if policies are clearly written and if third-parties are transparently disclosed.

We find news sites are highly dependent on third-parties for advertising revenue, core page functionality, and web hosting. 97% of news pages include content from Google, with 84% using the DoubleClick advertising service. A range of services from audience measurement to social media are hosted by third-parties, and just three web hosting companies are responsible for 43% of all news pages examined. The privacy impacts of centralization are profound: 99% of news pages examined load third-party content from an average of 41 distinct domains. 91% of sites include a third-party cookies, of those that have such cookies, we find 63 on average. This tracking is designed to be invisible to users and privacy policies are difficult to understand, time consuming to read, and only disclose 10% of observed third-party tracking. The majority of these measures are significantly worse for news than non-news pages.

2 BACKGROUND & RESEARCH QUESTIONS

While there are general risks associated with tracking on any category of site, there are particular concerns associated with tracking on news sites which may be organized by three themes: independence, privacy, and transparency. The following sections outline each of these concerns and their attendant research questions.

2.1 Independence

The Internet has been characterized as a decentralized network which distributes media power away from legacy intermediaries and into the hands of the public writ large [23]. However, the rise of a corporate giants in search (Google) and social media (Facebook, Twitter), shows that instead of removing intermediaries, the web has centralized even more power into fewer hands [40]. Pew's 2015 State of News Report revealed that Google, Facebook, Microsoft, Yahoo and Aol were responsible for "61% of total domestic digital ad revenue in 2014", with Google accounting for 38% of digital revenue [26]. Thus, a move to the web does not necessarily equate with increased independence, rather the dominance of behavioral advertising and centralized hosting services may reduce the underlying independence publishers have enjoyed for centuries.

The concept of press independence is well-defined and scholars have noted that press independence "has come to mean working with freedom: from state control or interference, from monopoly, from market forces, as well as freedom to report, comment, create and document without fear of persecution" [3]. Likewise, independence is a value held closely by "reporters across the globe [who] feel that their work can only thrive and flourish in a society that protects its media from censorship; in a company that saves its journalists from the marketers" [5]. Freedom from commercial influence is additionally put at risk by "native advertising and other practices online that blur the line between journalism and sponsored content" thereby threatening "the fundamentals of journalistic independence" [15].

Press independence may be undermined if a small group of organizations controls the underlying revenue generation function of the press or if a small group controls the publishing infrastructure which is now composed of servers and data centers rather than printing presses. If such centralization exists, the press may find themselves less able to challenge powerful entities, resist privacy-invasive business practices, and may be exposed to censorship if intermediaries are coerced into removing content. We pursue the following questions related to independence:

- How centralized, or distributed, are revenue generating mechanisms on news websites?
- How centralized, or distributed, is the use of third-party content on news websites?
- How centralized, or distributed, is the hosting of news websites?

2.2 Privacy

In the same way the free press depends on free speech to be able to write controversial content without interference, citizens rely on *privacy* to enable them to seek out content without being watched. Richards notes that there is little value in being free to write what you want if surveillance makes citizens too afraid to read it [33]. A 2015 study of search trends before and after revelations of NSA surveillance revealed that "there is a chilling effect on search behavior from government surveillance on the Internet" [22]. Likewise, users primed to be cognizant of government surveillance were significantly less likely to comment on a fictional news story describing US military action [39]. If news consumers feel they are being monitored they may be less likely to visit news websites which offer an adversarial take on the actions of the government, or discuss controversial matters with other citizens.

Web tracking techniques are designed to centralize the collection of reader habits into corporate-controlled databases as part of a economic model referred to as "panoptic" [11], "platform"[30, 38], "cognitive"[27], or "surveillance"[10, 44] capitalism. Regardless of the name, the underlying concept is that data gleaned from monitoring users may be used to generate profit, leading to an unending search for new sources of data.

These trends also make it easier for governments to leverage commercial surveillance for political and security needs as corporations may be exploited or coerced into giving access to data to government intelligence agencies such as the NSA [2]. Even without coercion, so-called "data brokers" may sell personal information to military and law enforcement organizations. A 2009 report revealed that the FBI's National Security Branch Analysis Center (NSAC) possessed "nearly 200 million records transferred from private data brokers such Accurint, Acxiom and Choicepoint" [36]. Likewise, according to an internal email regarding the now-defunct US Department of Defense "Total Information Awareness" project, a military official discussed obtaining Acxiom's data with the company's Chief Privacy Officer in 2002 [14].

Prior research has noted that news websites tend to have more tracking mechanisms than other websites [4, 7], but to date there have been few large-scale studies of tracking on news sites specifically (the Trackography project is one notable exception[3]). To add to existing knowledge on the topic, we pursue the following research questions:

- How is user privacy impacted by different types of third-party content?
- Does third-party content expose users to state surveillance?

2.3 Transparency

More than ink, paper, or advertising revenue, the press has always relied on the trust of readers to thrive. Reader trust is first and foremost grounded in the degree to which news organizations provide transparent accounting of relevant events. However, the technical underpinnings of web tracking rely on covert surveillance of users' web browsing habits, which is fundamentally antithetical to principals of transparency. One way this situation could be partially remedied is if privacy policies on news websites disclose the tracking taking place. Thus, a final question is asked:

- Do the privacy policies of news websites transparently disclose data flows to third-parties?

Pursuing the above questions provides insights into how third-party services could negatively impact the democratic role of the press, and require a multifaceted methodological approach.

3 METHODOLOGY

To answer our research questions, we collect and analyze a set of news and non-news web pages across several dimensions. Considerations regarding the design of the set of pages examined, methods for capturing and categorizing third-party content, and locating privacy policies are described below.

3.1 Data sampling and page collection

To determine if the risks associated with news sites are comparable to other types of popular sites we assemble lists of popular news and non-news websites. News sites are drawn from the US Newspaper List (USNPL.com), a well-organized and up-to-date list of newspapers, news-related magazines, television, and radio stations. From this list we scan over 7,000 pages to identify those that do not redirect to another domain and have at least 50 internal links, indicating the site has a variety of content and is not a placeholder.[4] We find 4,000 pages that meet our criteria. To build the non-news set of pages we draw 4,000 pages from the Alexa top 7,000 US sites which also do not redirect and have at least 50 internal links. The Alexa list is commonly used in web measurement research [7, 19, 34].

[3] https://myshadow.org/trackography

[4] We judged redirection based on the pubsuffix, thus "example.com" and "www.example.com" are not counted as a redirect whereas "example.com" and "example.net" are. We use the same criteria to define "internal link".

Given the dynamic nature of modern websites, we load the homepages from each set ten times to capture requests which may not have been found on a single page load. This yields a total of 80,000 page loads, 12.5 million third-party HTTP requests, and 3.4 million third-party cookies inclusive of news and non-news data sets. The computer used for this study is located at an academic institution in the United States, and data collection is performed in April, 2019.

3.2 Detecting third-party services

Once the sets of pages are established the open-source software tool `webXray` is used to detect third-party HTTP requests and cookies. `webXray` is given a list of URLs and loads each page in the Chrome web browser, closely reflecting real user behavior. During page loading the browser waits 45 seconds to give an opportunity for page scripts to download and execute. For each page load, `webXray` creates a fresh Chrome user profile which is free of prior browsing history and cookie data. During page loading no interaction takes place, meaning that notifications to accept cookies are not acted on, and all cookies are set without express user consent. `webXray` is an established tool used in prior web privacy measurement studies [12, 19–21].

The main benefit of `webXray` for this study is it provides fine-grained attribution library of the entities which operate third-party web services. While requests to third-party services are made to a specified domain, it is not always clear who owns a domain. For example, third-party content hosted on the the domain "1e100.net" comes from Google and content from "fbcdn.net" is hosted by Facebook. The `webXray` domain owner library is organized in a hierarchical fashion so that a single domain may be traced to its parent companies. For example, the domain "doubleclick.net" is owned by the DoubleClick service, which is a subsidiary of Google, which is a subsidiary of Alphabet. The `webXray` domain ownership library has been used to augment findings using the OpenWPM platform as well as studies of Android applications [7, 32].

3.3 Categorization of third-party content

There are a variety of reasons why a first-party site may include third-party services, and the `webXray` domain ownership library is extended with a service categorization. For over 200 services and companies, the homepage is visited to manually evaluate why a first-party would include content for the given service. It is important to note that our categorization is from the perspective of the *first*-party as the *third*-party may have different objectives. For example, while a site may utilize Google Analytics to gain insights into site traffic, Google may use that data for marketing purposes. This process yields several types of content, details of which are as follows:

- **Advertising** services are used to identify consumers, track their browsing behavior, predict their purchasing interests, and show them advertisements reflective of such predictions.
- **Audience measurement** systems allow site operators to learn about the people who visit a site and the actions they perform.
- **Compliance** tools allow sites to manage their privacy policies and consent notifications in order to comply with data protection laws.
- **Content recommendation** systems are often found at the bottom of articles and provide links to related articles on the same site and partner sites, as well as sponsored advertising content.
- **Design optimization** tools allow site designers to experiment with different designs (a process often called "A/B Testing").
- **Hosting services** run the physical infrastructure which delivers site content. Specialized types of content such as code libraries, fonts, and videos may be hosted from third-party domains. Likewise, generic hosting domains may serve first-party content under a third-party address.
- **Security** services exist to help site operators cope with threats such as distributed denial of service (DDoS) attacks and to prevent criminals using automated means to commit ad fraud and scrape content.
- **Social media** services have two main purposes: embedding user-generated content in a given page and facilitating users sharing a given URL on their social network of choice.
- **Tag managers** are a type of hosted code library with a specific function: helping sites to cope with large volumes of third-party tracking scripts ("tags"). Instead of reducing the number of tags, these services assist web developers with adding even more.

3.4 Identifying web hosting providers

To investigate the hosting of websites, we determine the parties which own a site's IP address using *whois* data. Such owners could be the entity which owns the site, as well as cloud-hosting providers such as Amazon Web Services. We calculate the average number of unique sites hosted by a given provider, revealing how centralized hosting is across the pages examined.

3.5 Collecting and analyzing privacy policies

In addition to monitoring content and cookies, `webXray` searches for and extracts links to privacy policies on a given page. The text of all links is evaluated to find matches in a list of terms associated with privacy policies. Once policy links are discovered, a second tool, `policyXray`, is used to harvest and analyze privacy policies.

`policyXray` has been used in prior research for auditing privacy policies [21]. `policyXray` uses the open-source Javascript library "Readability.js" to isolate and extract policy text [28]. The use of Readability.js is an essential step as it removes sections of the page which are not part of the policy. For sites with sidebar or footer links to Facebook or Twitter, removing non-policy content ensures that such text is not interpreted as part of the policy.

Once policy text is extracted, mentions of third-party services identified by `webXray` are searched for. If the names of companies are found, they are interpreted as disclosed in the policy. To give the most opportunities for disclosure, both the owner of the domain, variations on its spelling, and its parent companies are searched for. For example, if the domain "doubleclick.net" is found, the policy is searched to find matches for the strings "DoubleClick", "Double Click" (with a space), "Google", and "Alphabet". Additionally, `policyXray` analyzes the difficulty of reading a given policy using the English-language Flesch Reading Ease and Flesch-Kinkaid

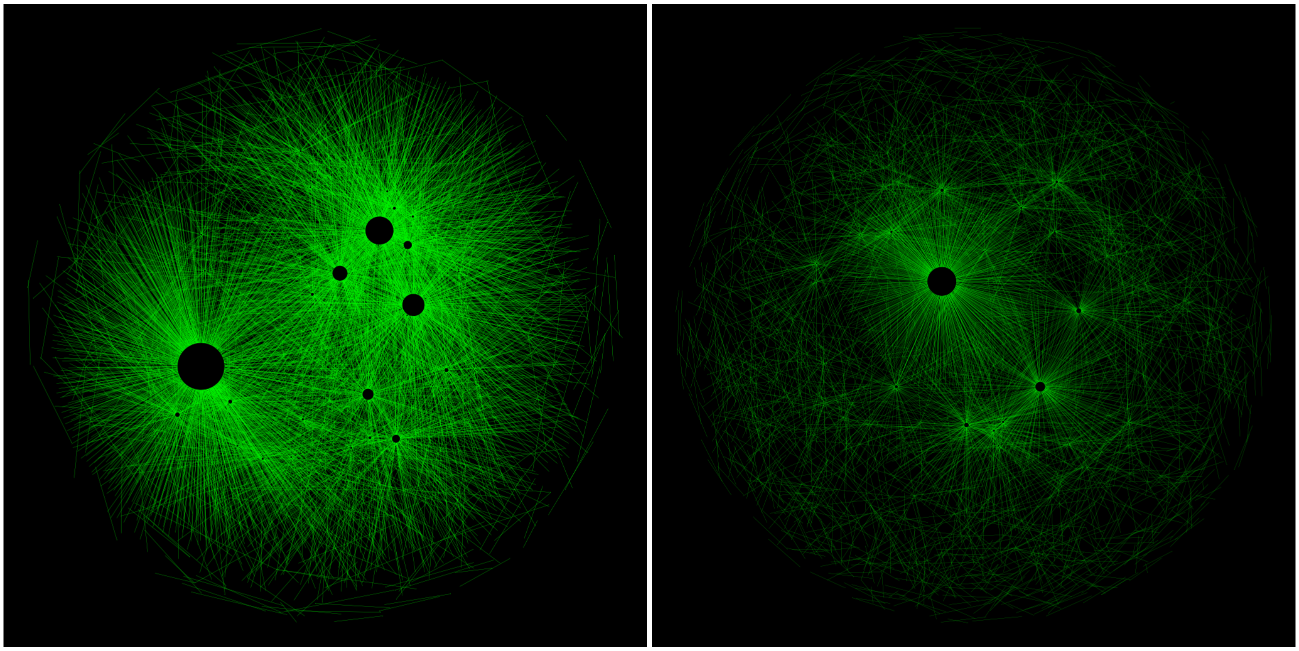

Figure 2: News sites (left) exhibit greater hosting centralization than non-news (right).

Grade Level metrics. We follow MacDonald and Cranor's prior work in this regard [24].

3.6 Limitations

There are several potential limitations to the approaches detailed above. First, the set of pages may not be fully comprehensive and thus not representative of larger trends. Second, webXray may potentially miss some tracking mechanisms, or be flagged as a "bot", resulting in an under-count of exposure. Third, webXray may miss some links to privacy policies if they do not match expected policy text. Finally, policyXray is not always able to parse the text found in a policy and sections of a policy may be erroneously discarded, thereby impacting the accuracy of disclosure measurements.

4 FINDINGS

Across all dimensions examined, use of third-party content by news websites has a negative impact on the democratic utility of the press. News websites rely on highly centralized revenue and hosting infrastructure, placing user privacy at risk, and such risks are not revealed in privacy policies. Furthermore, when compared to non-news sites, news website exhibit more centralization, worse privacy, and less transparency.

4.1 Centralization of revenue, third-party services, and hosting

To explore the independence of news sites we examine revenue generation, reliance on third-party services, and site hosting. We position the possibilities between two extremes: on the first, sites may broker their own advertisements, develop their own code, and host their own sites. Traditionally, news publishers have done many equivalent tasks in-house. For example, one of the authors delivered newspapers in his youth. On the other extreme, a small number of companies could control the purse strings for an entire industry, unilaterally make essential decisions on digital infrastructure, and own the physical apparatus which delivers the news. We find news on the web tracks closer to the second extreme.

Figure 3 shows the top ten third-party service providers found on news pages along with their equivalent reach on non-news pages. Of the top ten companies, only Amazon is not primarily an advertiser (though that is quickly changing as Amazon's ad services expand). The most remarkable finding is one company, Google, is found on 98% of news and 97% of non-news sites. Likewise, Facebook is able to track users on 53% of news and 51% of non-news sites. While these companies are dominant on both sets of sites, an additional nine companies are found on over 40% of news sites. In contrast, on non-news sites, only Google and Facebook cross the 40% threshold. Thus, while there is a diversity of third-parties, each party has a significantly more central role in the news ecosystem and the overwhelming majority of the most prevalent parties broker advertising.

These findings suggest two main threats to revenue independence. First, the scale of major advertising networks obviates the need for advertisers to engage with publishers directly, making it harder for news outlets to operate independently. Second, Pew found that digital advertising on news websites is dominated by "display ads such as banners or video" as opposed to "search ads" [26]. These types of ads rely on behavioral data for targeting, which is only possible when data is collected from a large range of sites and users. Although a news outlet may want to take control of their advertising, the inventory they offer advertisers will be more cumbersome to buy and less targeted to specific users.

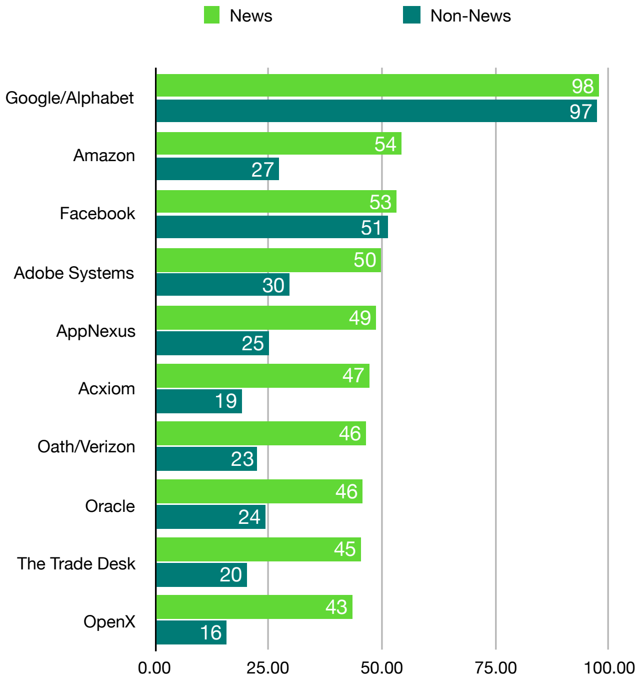

Figure 3: A large percentage of pages include content from a relatively small number of third-parties, a trend more pronounced in news.

Beyond third-party advertising revenue, websites may enhance the bottom line by utilizing third-party services for essential page functions, in turn reducing personnel expenses related to hiring software engineers and maintaining code. This is a less direct form of centralization than revenue, yet in some ways is more powerful. While a given company may part ways with an advertising network, replacing core page functionality can be enormously expensive, thereby establishing a hidden dependence.

Third-parties provide a number of services, and Table 1 shows the prevalence across sites examined. 85% of news sites contain third-party advertising content whereas only 76% of non-news sites do. Likewise, we find much higher reliance on marketing-driven content recommendation systems in news (19%) than non-news (10%). Due to the nature of news funding in the United States this is not a surprising finding. However, what is more surprising is that among sites which do have advertising, news sites use 21 distinct third-parties on average, whereas non-news use half as many (9).

Table 1 details the utilization of third-party services across several categories. In only two categories, design optimization and fonts, do non-news sites exhibit higher dependence on third-parties. For two categories, audience measurement and hosting, there is general parity between news and non-news. For the remaining eight categories, advertising, code, compliance, content recommendation, security, social media, tag management, and video, news sites exhibit higher dependence on third-parties. News shows much higher use of social media (72% vs 64%) - an issue we revisit in the privacy analysis.

The above measures only account for the number of network connections made, rather than the volume of data transferred. We find a high level of centralization when looking at data transfer: 64% of all data on news websites comes from a third-party domain, compared to 41% for non-news. Google is responsible for delivering the most third-party data (15% for news, 8% for non-news). The cost of hosting a website is often directly related to the volume of data transferred and using third-party services may keep costs down while simultaneously fostering dependence.

News websites also rely on third-parties to deliver *first-party* website content. Transferring this data takes physical material resources (e.g. electricity, computer hardware, air conditioning) which are comparable to what was formerly needed to deliver newspapers (e.g. paper, printing presses, delivery persons). We find most publishers outsource the hosting of their sites. Centralization of first-party hosting is in some ways the most troublesome issue related to press freedoms as it opens the door for authorities to conduct censorship for a large number of news outlets by putting pressure on a much smaller number of hosting providers.

We find a total of 268 unique web hosts for news sites compared to 1,084 for non-news, a nearly four-fold difference. Not only are the number of hosts for news sites far smaller, but only three companies (Lee Enterprises, Incapsula, and Amazon) host 44% of all US news sites examined. In contrast, for non-news, only one company hosts more than 7% of sites. Figure 2 illustrates site hosting networks, with several prominent nodes shown hosting news sites.

A final risk to publisher independence is large volumes of third-party advertising and tracking content may make news pages more expensive to view and slower to load than non-news websites. News websites take an average of 28 seconds to complete downloading compared to 17 seconds for non-news.[5] This may lead users to get their news from social media or aggregator websites which control which articles get presented to users, sidestepping news editors, and reducing the likelihood users will develop long-term relationships with a news outlet. The two most frequently detected third-parties on news sites, Google and Facebook, aggregate, select, and optimize delivery of news articles in centralized systems (*AMP* and *Instant Articles* respectively), making them direct competitors to news outlets. As with advertising, giving these companies access to user browsing histories allows them to provide targeting and selection of news content which may be superior to that provided by decentralized news outlets. For many users, download time will be the least of their worries when it comes to visiting news websites.

4.2 Privacy

While users may turn to the news to learn of the ways in which corporations compromise their privacy, it is news sites where we find the greatest risks to privacy. While nearly all sites expose user browsing behavior to third-parties (99% for news, 98% for non-news), on a per-page basis, news sites expose users to an average of 41 third-parties simultaneously compared to 21 for non-news. News sites expose users to third-party cookies on 91% of pages, compared to 84% for non-news. On a per-page basis, the number

[5] Measures of time are useful as relative rather than absolute measures given variations in network latency across locations and times. However, in this case pages are loaded from two computers sitting next to each other.

of third-party cookies is nearly three times greater: 63 on average for news compared to 23 for non-news. News sites also exhibit poorer security: only 70% of news pages use transport encryption, compared to 85% of non-news pages. While top-level analysis indicates that news sites fare more poorly in their stewardship of user security and privacy, further examination reveals different third-party services impact user privacy and security in distinct ways.

While all third-party requests expose users to potential tracking via analysis of HTTP log data [43], the presence of third-party cookies is a strong indicator that a given third-party is making a *purposeful attempt* to compromise users’ privacy by tracking them as they navigate between sites.[6] As Table 1 details, different types of services set cookies at considerably different rates, and for news sites, cookies are often set at greater rates. Advertising content sets cookies at high rates: 76% of news and non-news pages with advertising contain an advertising cookie. Content recommendation, which is arguably a sub-type of advertising, sets cookies at even greater rates (82% for news, 81% for non-news).

	% of Pages		*% w Cookie*	
Service Type	News	Non-News	News	Non-News
Advertising	85	76	76	76
- Content Rec	19	10	82	81
Audience Measure	93	93	45	23
Compliance	14	4	0	2
Design Optimize	14	28	45	44
Hosting	95	94	7	5
- Code	88	74	0	0
- Font	3	16	0	0
- Video	23	22	22	29
Security	28	16	4	6
Social Media	72	64	40	47
Tag Manager	77	61	0	0

Table 1: Most types of content are more prevalent on news sites, with the presence of cookies varying considerably.

Furthermore, each page with advertising content has an average of 21 distinct third-party domains compared with nine for non-news. There is a finite amount of room for advertising on a given page and even the most insufferable designs cannot accommodate 21 banner advertisements. Thus, the presence of seemingly redundant advertising content is best explained by the fact that even in cases where ads are not shown, an advertising network may track user behavior to display targeted advertisements on *other* pages.

Other types of services which utilize cookies at fairly high rates are audience measurement (44% news, 23% non-news), design optimization (45% news, 44% non-news), and social media (40% news, 47% non-news). In the case of audience measurement, privacy concerns are high as these systems are specifically used to record browsing behaviors. Cookies utilized by design optimization tools may not be designed for tracking users across sites, but may nevertheless represent a privacy risk. Social media content comes with the added privacy concern that cookies may be linked with specific off-line identities (Facebook for example requires the use of “real” names). Cookies are rarely, if ever, set for services related to compliance, fonts, code libraries, and tag managers, suggesting that these types of services may not be intentionally tracking users.

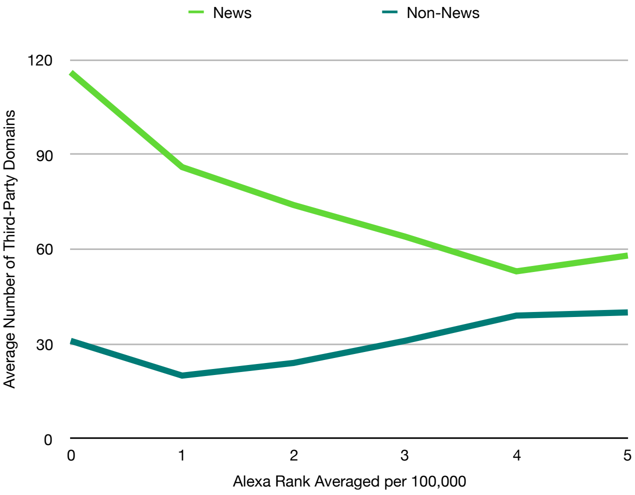

Figure 4: Of sites ranked in the top 500,000 in the United States, higher-ranked news sites have the most third-parties and news sites have more third-parties than non-news sites.

As noted above, under “surveillance capitalism” we might hypothesise that non-profit news outlets would likely have fewer privacy-compromising features than commercial news outlets. Although an imperfect proxy, the use of the “org" and “com" top-level domains are reliable indicators if a given site is a non-profit or commercial organization and thus provide a rough means to test this hypothesis. Of news sites examined, 3% are “org” and 92% are “com”, a breakdown which may reflect the commercial nature of US news media. As expected, we find non-profit sites exhibit a much lower percentage of sites with third-party cookies (78% for “org” vs 92% for “com”), and among sites with third-party cookies, commercial sites have over four times more (15 for “org” vs 66 for “com”). While the percentage of “org” and “com’ sites with any third-party content is similar (98% for “org”, 99% for “com”), the average number of third-party domains is nearly three times greater (15 for “org” vs 43 for “com”). Furthermore, sites which are ranked higher by Alexa, and which are likely more profitable, also have the greatest number of third-parties, as illustrated in Figure 4. The highest-ranked news site, The New York Times, leaks user data to 61 third-party domains (see Figure 1).

4.3 State surveillance

Some of the biggest risks deriving from poor privacy are related to state surveillance. Third-parties potentially expose users to two forms of state surveillance. In the first, third-parties may either be compromised or forced to disclose users’ web browsing information to authorities. In the second, companies which sell or share

[6] Although this analysis focuses on third-party cookies, it is also possible to track users with first-party cookies.

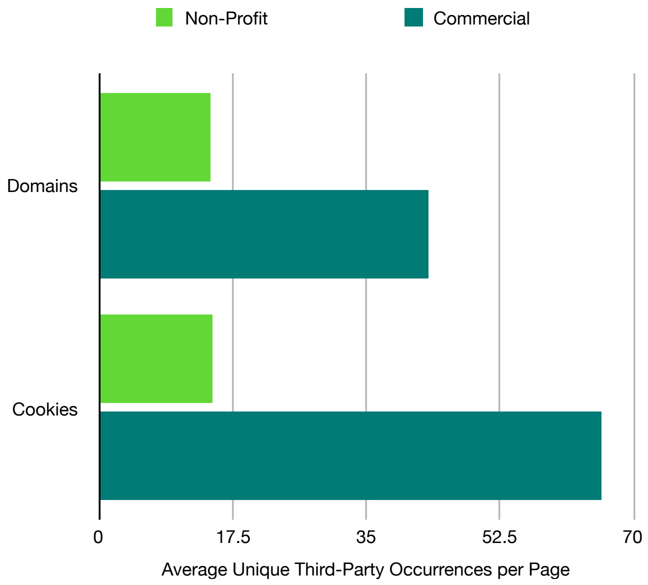

Figure 5: Non-profit news sites have fewer third-party domains and cookies than commercial sites.

personal data directly to the government may include web browsing information. Given the role of news media in exposing state surveillance, it is particularly relevant if news websites facilitate such outcomes themselves.

In 2013, former US National Security Agency (NSA) contractor Edward Snowden leaked details on how the NSA has used web browsing data for spying purposes. According to former Deputy US Chief Technology Officer Ed Felten, Snowden's disclosures revealed "a link between the sort of tracking that's done by Web sites for analytics and advertising and NSA exploitation activities" [37]. Likewise, *The Guardian* revealed a specific Google cookie, "Doubleclick ID", was used in efforts to spy on users of the Tor anonymity service [29]. Englehardt et al studied third-party cookies in 2015 and found users are "vulnerable to the NSA's dragnet surveillance".[8] This problem has not gone away: 74% of news and 50% of non-news pages include DoubleClick cookies.[7]

As noted above, the data broker Acxiom has discussed or sold user data to both the US Department of Defense and the FBI. We find Acxiom on 47% of news and 19% of non-news pages. Another company receiving large volumes of user data from news websites is Oracle, which has content on 46% of news and 24% of non-news pages. Of particular importance to the subject of state surveillance, Oracle has deep ties to US military and law enforcement: among Oracle's divisions are "Immigration and Border Control" which assist with "managing the tracking of individuals within national boundaries".[8]

It is not possible to determine if data Acxiom or Oracle collects from visitors to news websites is made available to government clients, but as noted above, the possibility *alone* is enough to dissuade users from accessing politically sensitive material [22, 39]. The blurring of boundaries between commercial and state surveillance is well-documented, but it is especially concerning given the special role of news media in the democratic process.

One reason to discount the chilling effects of surveillance is users who do not know about surveillance may not be dissuaded from seeking out politically controversial news. However, lack of awareness is another way in which reliance on third-parties may undermine the press.

4.4 Transparency: policy readability and third-party disclosure

To determine if a given web page's privacy policies are both clear and transparent, we harvest the privacy policies of 1,892 news sites (47% of the total), and 2,194 privacy policies of non-news sites (55% of the total). Next, three sets of measures are taken. First, we evaluate how difficult it is to read privacy policies. Second, we estimate how long it would take to read an average policy. Third, each page with a privacy policy is evaluated with `policyxray` to determine if the entities receiving data on the page are disclosed.

A given privacy policy may contain information which is valuable to users and informs them of privacy risks, yet for such information to be useful, it would need to be stated in terms a user could comprehend. The Flesch-Reading Ease (FRE) score is a 0-100 scale of reading difficulty, with 100 being easiest to read. The Flesch-Kinkaid Grade-Level score pegs a text against the US K-12 education system. Scores significantly above grade 12 become increasingly meaningless as the Flesch-Kinkaid formula may generate an infinitely high score. A grade-level score above 20 does not mean a PhD is needed to read the text, it means the text exceeds the utility of the grade-level scale.

Both news and non-news policies have poor FRE scores, 21 and 28 respectively. Likewise, their grade-level scores, 28 and 34 respectively, demonstrate a breakdown of the very applicability of grade-level metrics. While no enforceable standards exist for the readability of online privacy policies, insurance policies written below an FRE level of 45 are not enforceable in Florida [21]. Thus, if these were insurance, rather than privacy, polices they would not meet minimum legal requirements for clarity. Standards of transparency expected of the press should exceed what the state of Florida expects from insurers, yet the failure here is clear.

Privacy polices are also time consuming to read, raising an additional burden to users. On average, privacy policies for news sites are 2,263 words in length and require 9 minutes to read. In comparison, privacy policies for non-news sites average 2,033 words and require slightly over 8 minutes to read. While this amount of time may not initially appear onerous, MacDonald and Cranor have previously calculated the cost in time to read all such policies for an average user is considerable, and Libert has determined the time to read *both* first- and third-party policies exceeds 80 minutes for an average site [21, 24]. It is possible that disclosing a larger number of third-parties makes news policies lengthier, yet this explanation may be ruled out.

Although vague statements about sharing user data with "affiliates" or "partners" may be viewed as a *type* of disclosure, this

[7] Note the "ID" cookie appears to have been deprecated and "DSID" and "IDE" are now used.

[8] See http://www.oracle.com/us/industries/public-sector/046927.html.

Company	*% of Pages Tracked*	*% Disclosed in Policies*
Alphabet (Google)	95	38
Facebook	52	16
Amazon	46	2
Oracle	45	<1
Acxiom	41	0
Verizon (Oath)	40	<1
comScore	38	0
Twitter	34	4
AppNexus	32	1
OpenX	30	<1

Table 2: News privacy policies lack transparency, especially in regards to companies with no consumer-facing products such as Axiom and AppNexus.

study takes the approach that disclosure entails mentioning the *specific* third-parties present on a site. [9] Policies for both news and non-news sites fail to disclose the vast majority of third-parties. Only 10% of third-parties are disclosed in news privacy policies and only 14% in non-news policies. Both news sites and non-news sites share problematic features, with news sites once again faring worse.

Low rates of disclosure are not uniform and there is significant variability in the degree to which different third-parties are disclosed (see Table 2). Companies with services users may already be aware of such as search (Google) and social media (Facebook and Twitter) are more likely to be disclosed than those which users may not directly interact with such as AppNexus and Acxiom. Many parties are not mentioned in any privacy policies despite appearing on large numbers of pages. On news pages, 241 third-parties are found, of these, 169 (70%) are never mentioned in a privacy policy. On non-news pages, 266 third-parties are found, of these 202 (76%) are never mentioned. Thus, even users who make an effort to read privacy policies will likely never learn of many third-party services which may observe their browsing behavior. Given the very essence of journalism is transparency and disclosure, the state of privacy policies on news sites makes a poor case for citizens to place trust in these institutions.

5 RELATED WORK

Work related to the value of the press is detailed in the background section, and it is helpful to provide more context on related technical work here. There is a large literature devoted to the privacy and security impacts of web tracking. One early study correctly predicted that privacy would be at risk if an "advertising agency could add measurement code to the banner ads it distributes" [9]. A number of studies have investigated the presence of browser "fingerprinting" techniques which are used to track users without cookies [1, 6, 43]. Web measurement literature has been documenting the spread of tracking mechanisms at large scale since at least 2006 [17]. Recent studies have expanded the scale and scope of such investigations to document tracking across millions, or even billions, of sites [7, 19, 35] and examined how practices have changed over time [18, 41].

Researchers have also investigated the "notice and choice" privacy regulation framework while relies on users being notified of tracking by reading privacy policies. A large body of research has demonstrated this approach is ineffective as privacy policies are difficult to understand for most users [21, 24, 25], and they rarely disclose the third-party services [21]. Approaches such as crowdsourcing the interpretation of privacy policies may help alleviate this [42], but it remains difficult for users to learn of web tracking practices in general.

6 CONCLUSION

It is important to acknowledge that the ability of news outlets to stay in business is currently tied to a highly-centralized revenue model which is largely out of their control and fundamentally hostile to user privacy. However, difficult financial considerations do not absolve news media of responsibility for minimizing the amount of third-party content or increasing transparency of practices which impact user privacy. Likewise, news outlets have performed admirably when reporting on state surveillance and corporate privacy scandals, but rarely disclose how they also benefit from tracking users, raising the troubling question of what institutions users may rely upon to inform them of web tracking. Nevertheless, many would agree an imperfect press is better than no press at all.

Despite its democratic role, respect for the press is not universal. US President Donald Trump has stated that the news media "is the enemy of the American People".[10] While many around the world have grown weary of, and even inured to, this rhetoric, it is unwise to assume the significant powers of the US government could not be turned against the news media, especially if justified under the pretence of national emergency.

The ability of the press to withstand attempts at censorship and discovery of citizens' reading habits are deeply undermined by centralization of revenue and hosting. The fact that only three companies host 43% of news pages (Lee Enterprises, Incapsula, and Amazon) combined with duopolies in search (Google, Microsoft) and social media (Facebook, Twitter), means pressure on only seven companies could result in rapid and widespread information suppression and identification of political dissidents. While these scenarios may seem far-fetched, sober evaluation of the findings presented herein demands we consider the current state of centralization, privacy, and transparency on the web as a threat to democracy.

7 ACKNOWLEDGEMENTS

The authors thank reviewers for their insightful comments and Prof. Jonathan Smith for use of computing resources.

REFERENCES

[1] Gunes Acar, Marc Juarez, Nick Nikiforakis, Claudia Diaz, Seda Gürses, Frank Piessens, and Bart Preneel. 2013. FPDetective: dusting the web for fingerprinters. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security*. ACM, 1129–1140.

[9] While this study is based on a sample of US sites, it is instructive to note that EU data protection guidelines recommend that data controllers should typically disclose specific recipients rather than broad categories; see Article 29 Working Party Guidelines on Transparency WP260 rev.01, p37.

[10] https://twitter.com/realDonaldTrump/status/832708293516632065

[2] James Ball. 2013. NSA's PRISM surveillance program: how it works and what it can do. *The Guardian* 8 (2013).

[3] James Bennett, James Bennett, and Niki Strange. 2015. Introduction: the utopia of independent media: independence, working with freedom and working for free. *Media Independence: Working with Freedom or Working for Free* (2015), 1–28.

[4] Ceren Budak, Sharad Goel, Justin Rao, and Georgios Zervas. 2016. Understanding emerging threats to online advertising. In *Proceedings of the 2016 ACM Conference on Economics and Computation*. ACM, 561–578.

[5] Mark Deuze. 2005. What is journalism? Professional identity and ideology of journalists reconsidered. *Journalism* 6, 4 (2005), 442–464.

[6] Peter Eckersley. 2010. How unique is your web browser?. In *Privacy Enhancing Technologies Symposium*. Springer, 1–18.

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

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

[9] Edward W Felten and Michael A Schneider. 2000. Timing attacks on web privacy. In *Proceedings of the 7th ACM Conference on Computer and Communications Security*. ACM, 25–32.

[10] John Bellamy Foster and Robert W McChesney. 2014. Surveillance capitalism: Monopoly-finance capital, the military-industrial complex, and the digital age. *Monthly Review* 66, 3 (2014), 1.

[11] Oscar Gandy. 2005. If it weren't for bad luck. *14th Annual Walter and Lee Annenberg Distinguished Lecture* (2005).

[12] Christian Hauschke. 2016. Third-Party-Elemente in deutschen Bibliothekswebseiten. *Informationspraxis* 2, 2 (2016).

[13] Kashmir Hill. 2017. Yes, Google Uses Its Power to Quash Ideas It Doesn't Like—I Know Because It Happened to Me. *Gizmodo* (2017).

[14] Chris Jay Hoofnagle. 2003. Big Brother's Little Helpers: How ChoicePoint and Other Commercial Data Brokers Collect and Package Your Data for Law Enforcement. *North Carolina Journal of International Law and Commercial Regulation* 29 (2003), 595.

[15] Kari Karppinen and Hallvard Moe. 2016. What We Talk About When Talk About "Media Independence". *Javnost-The Public* 23, 2 (2016), 105–119.

[16] Bill Kovach and Tom Rosenstiel. 2007. *The elements of journalism: What newspeople should know and the public should expect*. Three Rivers Press (CA).

[17] Balachander Krishnamurthy and Craig E Wills. 2006. Generating a privacy footprint on the internet. In *Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement*. ACM, 65–70.

[18] Adam Lerner, Anna Kornfeld Simpson, Tadayoshi Kohno, and Franziska Roesner. 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 (USENIX Security 16)*. USENIX Association, Austin, TX. https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/lerner

[19] Timothy Libert. 2015. Exposing the Hidden Web: Third-Party HTTP Requests On One Million Websites. *International Journal of Communication* (2015).

[20] Timothy Libert. 2015. Privacy Implications of Health Information Seeking on the Web. *Commun. ACM* (2015).

[21] Timothy Libert. 2018. An Automated Approach to Auditing Disclosure of Third-Party Data Collection in site Privacy Policies. *Proceedings of the 2018 World Wide Web Conference* (2018), 207–216.

[22] Alex Marthews and Catherine E Tucker. 2015. Government surveillance and internet search behavior. *SSRN* (2015).

[23] Robert W McChesney. 2013. *Digital disconnect: How capitalism is turning the Internet against democracy*. New Press, The.

[24] Aleecia M McDonald and Lorrie Faith Cranor. 2008. The Cost of reading privacy policies. *I/S: A Journal Of Law And Policy For The Information Society* 4 (2008), 543.

[25] Aleecia M McDonald, Robert W Reeder, Patrick Gage Kelley, and Lorrie Faith Cranor. 2009. A comparative study of online privacy policies and formats. In *Privacy Enhancing Technologies*. Springer, 37–55.

[26] Amy Mitchell and Tom Rosenstiel. 2015. State of the news media 2015. *Pew Research Center. Journalism & Media* (2015).

[27] Yann Moulier-Boutang. 2011. *Cognitive capitalism*. Polity.

[28] Mozilla. 2017. Readability.js. https://github.com/mozilla/readability. (07 2017).

[29] National Security Agency. 2013. Tor Stinks Presentation. *The Guardian* http://www.theguardian.com/world/interactive/2013/oct/04/tor-stinks-nsa-presentation-document (2013).

[30] Frank Pasquale. 2016. Two narratives of platform capitalism. *Yale L. & Pol'y Rev.* 35 (2016), 309.

[31] Victor Pickard. 2014. *America's Battle for Media Democracy*. Cambridge University Press.

[32] Max Van Kleek Jun Zhao Timothy Libert Nigel Shadbolt. Reuben Binns, Ulrik Lyngs. 2018. Third Party Tracking in the Mobile Ecosystem. *WebSci âĂŽ18: 10th ACM Conference on Web Science* (2018).

[33] Neil M Richards. 2008. Intellectual privacy. *Texas Law Review* 87 (2008), 387.

[34] Franziska Roesner, Tadayoshi Kohno, and David Wetherall. 2012. Detecting and defending against third-party tracking on the web. In *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation*. USENIX Association, 12–12.

[35] Sebastian Schelter and Jérôme Kunegis. 2016. On the Ubiquity of Web Tracking: Insights from a Billion-Page Web Crawl. *arXiv preprint arXiv:1607.07403* (2016).

[36] Ryan Singel. 2008. Newly declassified files detail massive FBI data-mining project. (2008).

[37] Ashkan Soltani, Andrea Peterson, and Barton Gellman. 2013. NSA uses Google cookies to pinpoint targets for hacking. *The Washington Post* https://www.washingtonpost.com/blogs/the-switch/wp/2013/12/10/nsa-uses-google-cookies-to-pinpoint-targets-for-hacking (2013).

[38] Nick Srnicek. 2017. *Platform capitalism*. John Wiley & Sons.

[39] Elizabeth Stoycheff. 2016. Under surveillance examining Facebook's spiral of silence effects in the wake of NSA internet monitoring. *Journalism & Mass Communication Quarterly* (2016), 1077699016630255.

[40] Siva Vaidhyanathan. 2012. *The Googlization of everything:(and why we should worry)*. University of California Press.

[41] Tim Wambach and Katharina Bräunlich. 2016. Retrospective Study of Third-party Web Tracking. In *Proceedings of the 2nd International Conference on Information Systems Security and Privacy*. 138–145. https://doi.org/10.5220/0005741301380145

[42] Shomir Wilson, Florian Schaub, Rohan Ramanath, Norman Sadeh, Fei Liu, Noah A Smith, and Frederick Liu. 2016. Crowdsourcing Annotations for Websites' Privacy Policies: Can It Really Work?. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 133–143.

[43] Ting-Fang Yen, Yinglian Xie, Fang Yu, Roger Peng Yu, and Martın Abadi. 2012. Host fingerprinting and tracking on the web: Privacy and security implications. In *Proceedings of the Network and Distributed System Security Symposium*.

[44] Shoshana Zuboff. 2015. Big other: surveillance capitalism and the prospects of an information civilization. *Journal of Information Technology* 30, 1 (2015), 75–89.