Original Manuscript

Social Science Computer Review
2023, Vol. 0(0) 1–16

DOI: 10.1177/08944393231218214
journals.sagepub.com/home/ssc
Sage

Quantifying the Systematic Bias in the Accessibility and Inaccessibility of Web Scraping Content From URL-Logged Web-Browsing Digital Trace Data

Ross Dahlke[1], Deepak Kumar[2], Zakir Durumeric[2], and Jeffrey T. Hancock[1]

Abstract

Social scientists and computer scientists are increasingly using observational digital trace data and analyzing these data post hoc to understand the content people are exposed to online. However, these content collection efforts may be systematically biased when the entirety of the data cannot be captured retroactively. We call this often unstated assumption the problematic assumption of accessibility. To examine the extent to which this assumption may be problematic, we identify 107k hard news and misinformation web pages visited by a representative panel of 1,238 American adults and record the degree to which the web pages individuals visited were accessible via successful web scrapes or inaccessible via unsuccessful scrapes. While we find that the URLs collected are largely accessible and with unrestricted content, we find there are systematic biases in which URLs are restricted, return an error, or are inaccessible. For example, conservative misinformation URLs are more likely to be inaccessible than other types of misinformation. We suggest how social scientists should capture and report digital trace and web scraping data.

Keywords

digital trace data, internet measurement, misinformation, web-log data, web scraping, news, news consumption

[1]Department of Communication, Stanford University, Stanford, CA, USA
[2]Department of Computer Science, Stanford University, Stanford, CA, USA

Corresponding Author:
Ross Dahlke, Department of Communication, Stanford University, 450 Jane Stanford Way, Stanford, CA 943056104, USA.
Email: rdahlke@stanford.edu

Data Availability Statement included at the end of the article

Introduction

Social science researchers increasingly use observational web-tracking and digital trace data to understand digital content exposure and effects patterns. However, most social scientists do not collect digital trace data in real-time (Lukito et al., 2023) but instead retroactively try to access them, often through an API (Application Programming Interface, Jünger, 2021; Praet et al., 2022), data vendor (e.g., Lyons, 2022), or scraping the content of web pages (Freelon, 2018). In the present work, we focus on this post hoc scraping of the content of web pages, a common practice among researchers (e.g., Ben-David, 2016; Guess, 2021; Guess et al., 2021; Li et al., 2021; Reiss, 2022; von Hohenberg et al., 2023; Wojcieszak et al., 2021). However, these post-facto content collection efforts may be systematically biased by the inability to capture the content of many of these websites after the fact. For example, a website may have been deleted or behind a paywall.

In this paper, we seek to quantify the systematic bias that may result from scraping web page content from web log data. Specifically, we are concerned that some websites individuals consume may be more difficult to collect content from than others. To examine this bias, we leverage a dataset of misinformation and hard news websites visited by a panel of 1,238 American adults over three months. We scraped each hard news and misinformation URL a participant visited via a fully-fledged web browser (e.g., Google Chrome) to capture the content loaded on the page.

In our paper, we make three core contributions: First, we categorize the output of web scrapes into two main categories: accessible data in which the scrapes are successful and inaccessible data in which the scrapes are unsuccessful. We then further subcategorize accessible data from successful scrapes into unrestricted content, restricted content, or errors. Second, we investigate systematic differences in the distribution of content in these categories and show discrepancies related to the ideology of the source. Third, we provide recommendations for future researchers on how to collect web scraping data and call for adopting a standardized set of reporting metrics and a reporting format that researchers using web scraping can take to standardize reporting of potential systematic biases in their data.

The proliferation of digital trace data (Baumgartner et al., 2022; Choi, 2020; Jungherr et al., 2017; Kreuter et al., 2020; Revilla et al., 2017) has led to a "Big Data" revolution (Chen & Quan-Haase, 2020; Christ et al., 2021; Eck et al., 2021; Gil de Zuniga & Diehl, 2017; Wells & Thorson, 2017). Today, social scientists can explore new questions in human behavior that were difficult or impossible to study in the past. For example, recent research has examined the relationship between political interest and the actual sharing of political information on social media (Haenschen, 2020), gendered differences in civic engagement (Brandtzaeg, 2017), the connection between digital behaviors and vote choice (Bach et al., 2021), and observed digital news consumption (Möller et al., 2020).

These data are collected post hoc and, therefore, can be studied because they are successfully accessed after the fact. However, most past work does not consider that there may be systematic biases in the data stemming from inaccessible data. Furthermore, even if data are accessible in a technical sense, they may be of limited usefulness if the content is restricted or returns errors. We call the reliance on digital trace data in computational social science the *problematic assumption of accessibility*. This assumption is often unstated but assumes that the digital traces available to a researcher are representative and complete. We argue that while a great deal of digital trace data are accessible and reasonably captures social behavior or experiences, some digital trace data are inaccessible or unusable to answer social scientific questions. Below, we lay out these forms of trace data that may undermine assumptions that trace data are representative and complete. We connect these data types to social scientific ideas of persistent and ephemeral communication.

Background on the Web

This paper uses web behavior data collected from a representative panel of 1,238 American adults. To provide more clarity, we detail the necessary background on how web behavior is defined in this section.

Understanding Web Requests

In order to access a website, a web client (e.g., a web browser) must issue a *web request* for the contents of that website from a remote server. Requests are sent using the Hypertext Transfer Protocol (HTTP), a stateless protocol designed for web clients and servers to communicate with one another when delivering content easily. A web request contains several important pieces of information: the URL of the remote server, headers (which can contain information about the client itself or state set onto the browser), and a body, which contains data to send to the web server. In this paper, we log all web *requests* made by our representative panel.

Understanding Web Responses

When a web server responds to a web request, it does so by sending back a *web response*. Responses are also sent via HTTP and chiefly contain the requested content (e.g., the data for a web page) and a *status code*, which ranges from 100 to 599, describing *how* the web server handled the request. For example, a returned status code of 200 indicates the web server handled the request correctly and with no errors, whereas a status code of 404 indicates that the web server could not find the page embedded in the web request. In our paper, we leverage status codes >= 400 to identify if a web server encountered an error when processing our requests.

Accessible Data

From a technical perspective, accessible data can be accessed or retrieved through web scraping, which is the process of automatically extracting data from a webpage. Early internet scholars documented the extent to which web pages were accessible or not. For example, early estimates found that websites are generally accessible, with about 83.8% of web pages accessible (Koehler, 1999). This line of inquiry has also been extended to academic publications. "Citation rot" or "link rot" is when digital academic article reference material becomes unretrievable (Tyler & McNeil, 2003) and potentially disrupts scholarly progress because researchers cannot find relevant reference material. This concern persists today (D Kumar et al., 2015; e.g., Klein et al., 2014; Perkel, 2015) and is shared across social science disciplines (e.g., Dimitrova & Bugeja, 2007; Gertler & Bullock, 2017; Spence & Burns, 2020). Accessibility is important to scholars because it allows for recreating and revisiting the original content they seek to study.

These technical ideas are closely related to the social scientific principle of *persistence*. In the field of communication, persistent communication is permanent, static, and atemporal (Linell, 2004, p. 8). Often, this idea is used to consider the conceptual differences between forms of communication, such as books and spoken language. Books, as long as they are properly maintained, remain persistent.

Accessible but Restricted Data

Just because data are accessible from a technical perspective does not mean they are necessarily usable for answering specific social scientific questions. One may scrape a website without an error, but the desired content may be restricted. For example, paywall journalism creates restricted communication without the proper credentials. Paywalls are barriers between internet users and online content from news organizations (Pickard & Williams, 2014). The news publishing industry quickly adopted (Franklin, 2014) this "retro-innovation" (Arrese, 2016) to find new revenue streams (Pavlik, 2013; Sjøvaag, 2016) with mixed success (Myllylahti, 2014). Journalistic stories behind paywalls continue to exist and are visitable, so they are not inaccessible in a technical sense. However, one must possess proper credentials to access the content—not just anyone can visit the content in the first place. In other words, this content is accessible but restricted.

These accessible-yet-restricted data are often under-considered. News organizations do not randomly construct paywalls; thus, content is not randomly restricted to people, including researchers. For example, even on the same website, hard news and opinion pieces are more likely to be behind paywalls than other web pages (Myllylahti, 2017)—the sort of content most likely to interest scholars. In addition, news organizations will occasionally temporarily drop their paywall for public emergencies, planned special events, and broader access to civically valuable content (Ananny & Bighash, 2016).

Of course, restricted data are not new. For example, one may have had to pay for print newspapers. However, what is new is how researchers attempt to interact with the data. While researchers in the past may have accessed the totality of articles that appeared in The New York Times via a first- or third-party archive, researchers are increasingly collecting their own data, often through web scraping (Krotov & Silva, 2018; Landers et al., 2016; Olmedilla et al., 2016). Thus, restricted data pose additional problems for researchers above and beyond inaccessibility because scholars must also consider how to access the content in addition to simply recording their existence. For example, internet scholars may record a webpage snapshot before the page gets taken down and becomes restricted. Researchers must also decide how to get past the restrictions that may otherwise render a web page's contents unusable for the social science question they are asking.

Inaccessible Data

Technically, inaccessible data cannot be accessed or retrieved through web scraping. In the computer science security community, significant prior work has studied how adversarial actors cloak or hide malicious activity using Fast Flux Domains (Holz et al., 2008). These inaccessible domains are brought online for a short time, typically to conduct some kind of internet abuse (e.g., distributed denial-of-service attacks or DDoS), and quickly taken offline to avoid discovery. Studying the structure of these domains is key to understanding how botnets propagate (Bilge et al., 2011; Stone-Gross et al., 2009) and can inform defenses against abusive Internet behaviors (Perdisci & Lee, 2018).

The technical categorization of some web data as inaccessible is similar to the social scientific idea of ephemerality (e.g., Clark, 1996; Linell, 2004). In contrast to "atemporal" persistent communication, ephemeral communication is fleeting and ceases to exist; it is "distributed in time" (Linell, 2004, p. 5). For example, spoken word, if unrecorded, leaves no tangible evidence of its prior existence and contents. Modern media technology complicates the relationship between persistence and ephemerality. Instagram stories (Bainotti et al., 2021; Carah & Shaul, 2016; Vázquez-Herrero et al., 2019) and Snapchat (Bayer et al., 2016;

Cavalcanti et al., 2017; Chowdhury et al., 2021; McRoberts et al., 2019; Villaespesa & Wowkowych, 2020) are two prominent contemporary media platforms that feature ephemeral content. These platforms are designed to disappear after a specific amount of time, generally 24 hours. Given the fleeting nature of these communications, these ephemeral media model the oral paradigm of communication and storytelling (Soffer, 2016), but they introduce a new dynamic of easy capture where they are designed to be ephemeral but can be captured, for example, through screenshots on personal devices.

The distinction between accessible-yet-restricted and inaccessible data is important because the implications for researchers and their analysis are unequal. While both data types may be missing from previous analyses, how researchers can access and use these types differ significantly. Accessible-yet-restricted data pose additional challenges for researchers, as they must identify the existence of restricted content and find ways to gain access to it. In other words, accessible-yet-restricted data may appear, at first glance, to be the intended content that one desires to study when actually additional precautions are needed to avoid it tainting an analysis. On the other hand, inaccessible data cannot be retrieved through web scraping, making it potentially impossible for researchers to access and use the data. Therefore, understanding the distinction between these two categories is crucial for researchers to determine the feasibility of answering specific social scientific questions and to develop appropriate research methods and strategies.

Accessibility, Inaccessibility, and the Study of Misinformation

In the present paper, we examine accessibility and inaccessibility in the context of misinformation. The study of misinformation on the internet has become an important area of research that relies on digital trace data. Many studies examine how often and in what ways people are exposed to misinformation online (Dahlke et al., 2022; Guess et al., 2020; Moore et al., 2023b) and to what effect (Dahlke & Hancock, 2022). One concern in misinformation research is that it has not accounted for restricted and inaccessible web-based misinformation. Many popular misinformation studies leverage lists of curated misinformation websites, but these websites are often inaccessible or offline by the time studies are conducted (Han et al., 2022; Hanley et al., 2022; Hounsel et al., 2020). Internet measurement studies on misinformation often have to discard up to 50% of domains in these human-curated lists, creating the possibility of significant bias in collected results. For example, past research (Hounsel et al., 2020) found that in a curated set of 758 disinformation websites, 575 (76%) were no longer available and had to be manually reconstructed using historical snapshots. While it is clear that accessibility is a problematic assumption, we do not know to what extent this is an issue, nor do we know whether inaccessibility and unusability are systematic in the actual web pages that people visit.

Quantifying Accessibility and Inaccessibility on the Internet

Quantifying the accessibility of digital trace data is vital to social scientists studying human behavior on the internet because this content may not be randomly accessible or inaccessible. If the distribution is random, there would be less concern. However, a biased distribution would skew findings from internet researchers towards only the information they could collect. This bias is even likely given the examples above of Fast Flux Domain Networks and Paywall Journalism. Linguistics already grapples with this systematic concern by acknowledging a bias toward studying written, persistent language over spoken, ephemeral communication (Linell, 2004). We seek to examine these potential sources of error for

scholars studying content exposure on the internet and document the extent of these possible biases. We consider this bias on two of the most common objects of study on the internet: exposure to hard news and misinformation websites.

Specifically, we ask two research questions:

RQ1. To what extent are hard news and misinformation website visits accessible and inaccessible? Of accessible data, to what extent is the content returned unrestricted, restricted, or an error?

RQ2. Are there systematic biases in the websites and types of websites that are accessible and inaccessible with respect to ideology?

Data, Measures, and Methods

Data

The data for this project come from a two-wave online survey administered via YouGov during the 2020 election to 1,238 American adults. We passively gathered web-browsing data (i.e., URLs) from those participants using YouGov's Pulse browser plugin from August 24, 2020, to December 7, 2020. In total, we collected approximately 21M web visits from these participants. All participants consented to the terms of the research, and YouGov compensated the participants.

Measures

We narrowed our list of 21 million visited URLs to hard news websites, as defined by Baksy et al. (2015) and NewsGuard,[1] and misinformation websites, as categorized by Moore et al. (2023b). We assigned ideological labels to websites using NewsGuard's ratings and classifications from Baksy et al. (2015). In addition, we only examined URLs that were to content webpages, that is, we removed URL visits to pages such as home pages that are not specific pieces of content in an attempt not to consider dynamic web pages and removed the query parameters (i.e., site-specific data embedded in the URL) from the URLs. Some commonly visited domains that were generally home pages, contained mostly sports content, or were labeled as partisan but ostensibly are not (e.g., websites that report the weather) were not included in the calculations.[2] These steps left us with 106,685 unique URLs from 1,238 participants.

Method

One year after collecting the URL logs, we visited each URL using a headless Google Chrome web browser one year after collecting URL logs. We did this to most closely simulate the real-world browsing experience of end-users using an Internet browser.[3] We also stress that significant web content is loaded dynamically. Therefore, a browser-based technique is necessary to retrieve the content of each URL (Kumar et al., 2017). In some cases, the browser *crashed* when visiting the URL. This crash can happen for several reasons, ranging from poorly administered web servers to missing DNS entries. If the browser crashed when visiting a URL, we denoted that scrape as *unsuccessful*. If the browser was able to retrieve some page content, we denoted that scrape as *successful.* One potential limitation of this approach, that future scholars using web-browsing data should consider, is that it does not consider

personalized content. Future work should develop a method to capture this personalized content in real-time.

In investigating the successfully retrieved web content, we noticed that many successful scrapes either returned an HTTP Error (i.e., the status code was >= 400) or were behind a paywall. To better characterize this, we subcategorized each successful scrape into three buckets: *restricted content, unrestricted content, and errors.* We define each below:

1. *Error content* is web content where the web server returns an HTTP status code greater than 400.
2. *Restricted content* sits behind a paywall, login page, or some other error message on the web page itself.
3. *Unrestricted content* is any content that is not restricted or returns an error.

We identify error content simply by observing the HTTP status code returned for each URL we requested. To identify restricted content, we built a simple machine-learning classifier that could discern between content that sits behind a paywall and non-paywalled content (for more details, see Supplemental Materials A). For our training data, two members of the research team hand-coded a random subset of 9,636 webpages (IRR, Cohen's Kappa = .85) for whether the page contained a message restricting access (e.g., "This page is not available right now.") We then leveraged this hand-coded dataset to fine-tune a publicly available Huggingface BERT classifier to identify restricted content. Of the 9,636 hand-coded web pages, we used 7,724 for the training set, 1,405 for the test set, and 507 for the validation set.

The model achieved an F1 score of .92 on the validation set. After applying this model to the entire set, we categorized 97,395 (91.3%) as successfully scraped with unrestricted content, 753 (.7%) as successfully scraped with restricted content, 8,385 (7.9%) as successfully scraped with an error, and 152 (.1%) as unsuccessfully scraped.

We employ a standard chi-squared test on the distributions of accessibility categories of various websites (e.g., liberal misinformation websites). The top-line results for RQ1 are in Table 1, and the heterogeneous results for RQ2 are in Table 2. We also examined alternative specifications to see if the distributions remain significantly different under different categorical groups, finding that the results are robust to other potential groupings (Supplemental Materials B).

To investigate how stable our results are over time, we also scraped each web page at two additional time points: once after one-and-a-half years post data collection and once after two full years (see Figure 1).

Table 1. Percentage of Hard News and Misinformation URLs in Each Category.

	Accessible			Inaccessible
URL Category	Successful - Unrestricted	Successful - Restricted	Successful - Error	Unsuccessful
hard news	91.1%	0.7%	8.1%	0.1%
misinformation	95.9%	0.4%	2.8%	0.9%

Note: $\chi^2(3) = 372.3$, $p < .001$. Distribution of websites in each accessibility category. Accessible websites are those in which the web scrape is successful. Inaccessible websites are those in which the web scrape is unsuccessful. Accessible websites are further categorized into unrestricted content in which the content is not restricted or returns an error, restricted websites in which the content sits behind a paywall, login page, or some other error message on the web page itself, and errors in which the web server returns an HTTP status code greater than 400.

Table 2. Percentage of Hard News and Misinformation URLs in Each Category by Ideological Slant of Website.

		Accessible			Inaccessible
URL Category	Ideology	Successful - Unrestricted	Successful - Restricted	Successful - Error	Unsuccessful
Hard news					
Hard news	conservative	96.3%	0.1%	3.4%	0.1%
Hard news	liberal	89.6%	0.5%	9.9%	0.1%
Hard news	other	90.5%	1.2%	8.1%	0.1%
Misinformation					
Misinformation	conservative	95.4%	0.5%	3.0%	1.1%
Misinformation	liberal	98.4%	0.5%	1.1%	0.0%
Misinformation	other	96.4%	0.0%	3.0%	0.6%

Note: Hard news webpages: $\chi^2(3) = 745.6$, $p < .001$; Misinformation webpages: $\chi^2(3) = 13.1$, $p = .005$. Distribution of websites in each accessibility category. Accessible websites are those in which the web scrape is successful. Inaccessible websites are those in which the web scrape is unsuccessful. Accessible websites are further categorized into unrestricted content in which the content is not restricted or returns an error, restricted websites in which the content sits behind a paywall, login page, or some other error message on the web page itself, and errors in which the web server returns an HTTP status code greater than 400.

Results

To answer RQ1, we quantified the rates of our accessibility categories for hard news and misinformation websites in our data set (Table 1). Most hard news and misinformation web pages were successfully scraped and contained unrestricted content (91.1% of hard news pages and 95.9% of misinformation pages). However, compared to misinformation web pages, hard news sites were almost twice as likely to be successfully scraped but with restricted content and nearly three times as likely to be successfully scraped but returned an error. In contrast, misinformation web pages were nine times more likely to return an unsuccessful scrape than hard news pages.

Some of these findings are aligned with the conventional wisdom. For example, misinformation websites were more likely to be unsuccessfully scraped and, thus, inaccessible. However, some findings are surprising. One that stands out is that hard news is more likely to be successfully scraped but return an error. Speculatively, this result may be due to active maintenance from hard news publishes. For example, some outlets may be archiving old stories. Future work should more deeply investigate the source of this result.

We also analyzed how these results change over three snapshots taken approximately one year, one-and-a-half years, and two years after data collection (Figure 1). The percentage of web pages from hard news and misinformation successfully scraped with unrestricted content was relatively stable, with hard news slightly decreasing from the first snapshot to the third. However, the percentage of hard news websites successfully scraped but with restricted content triples from the first snapshot to the third. Both hard news and misinformation websites showed a jump in the percentage of web pages that were unsuccessfully scraped from the first to the second snapshot. In the case of hard news, this percentage dropped slightly in the third snapshot. However, the main result of a significantly different distribution remains the case over all the snapshots (see Supplemental Materials B for more details). In addition, we detail the rates at which web pages' categorizations change across the snapshots in Supplemental Materials C. We discuss the implications of these results below.

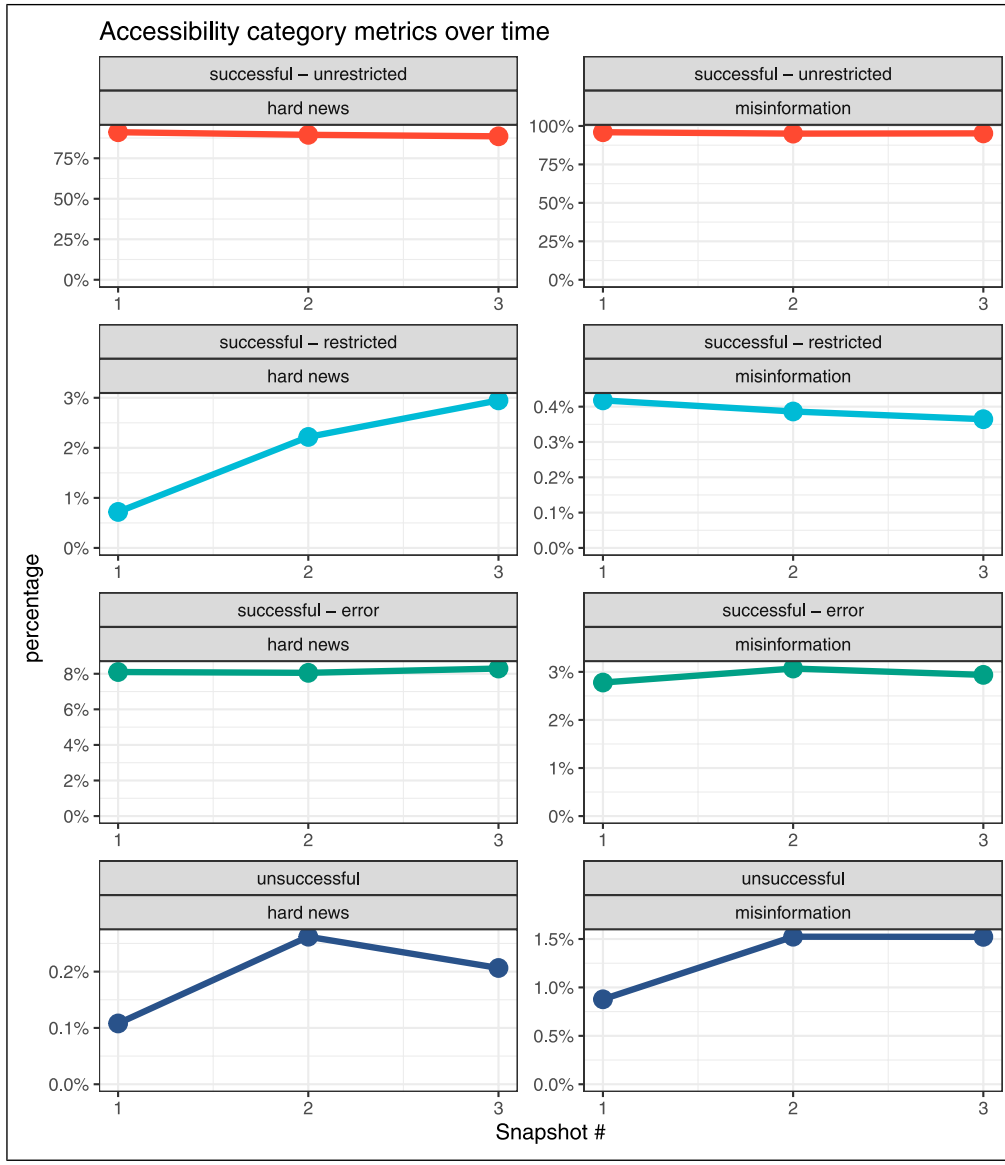

Figure 1. Accessibility category metrics over time.

For RQ2, which asks about the potential biases in accessibility categories, we find significant differences in conservative versus liberal web pages (Table 2). Liberal hard news web pages are more likely to be successfully scraped but return an error than conservative hard news web pages. However, conservative misinformation web pages, compared to liberal web pages, were more likely to be successfully scraped but returned an error and to be unsuccessfully scraped. In other words, there are systematic biases in the ideological bent of the types of web pages that can be recovered for post hoc analysis.

Specific domains are more likely to have URLs that fall into specific buckets. As seen in Figure 2, some websites almost entirely returned unsuccessful or successful yet restricted

content or error messages. For example, over 75% of scrapes to *The New York Times*, a liberal hard news website, were successful but returned an error. Or, scrapes to theredelephants.com, a conservative misinformation website, were entirely unsuccessful. Said another way, there are hard news and misinformation websites that are systematically difficult for researchers to record the content of, which may bias studies including these websites.

Discussion

The present study examined the accessibility and usability of scraped websites in a nationally representative sample of American adults' web browsing during the 2020 U.S. Presidential Election. We find that hard news web pages are more likely than misinformation websites to be successfully scraped but with restricted content or errors. However, misinformation web pages were much more likely to be unsuccessfully scraped. Looking at the ideological slant of the web pages, liberal hard news web pages are more likely to be successfully scraped but with an error than conservative hard news web pages. However, conservative misinformation web pages were more likely to be scraped successfully but with an error or unsuccessfully.

Furthermore, we see that the accessibility status of websites shifts over time. The primary reason for this is a significant increase in restricted content over time—for hard news websites,

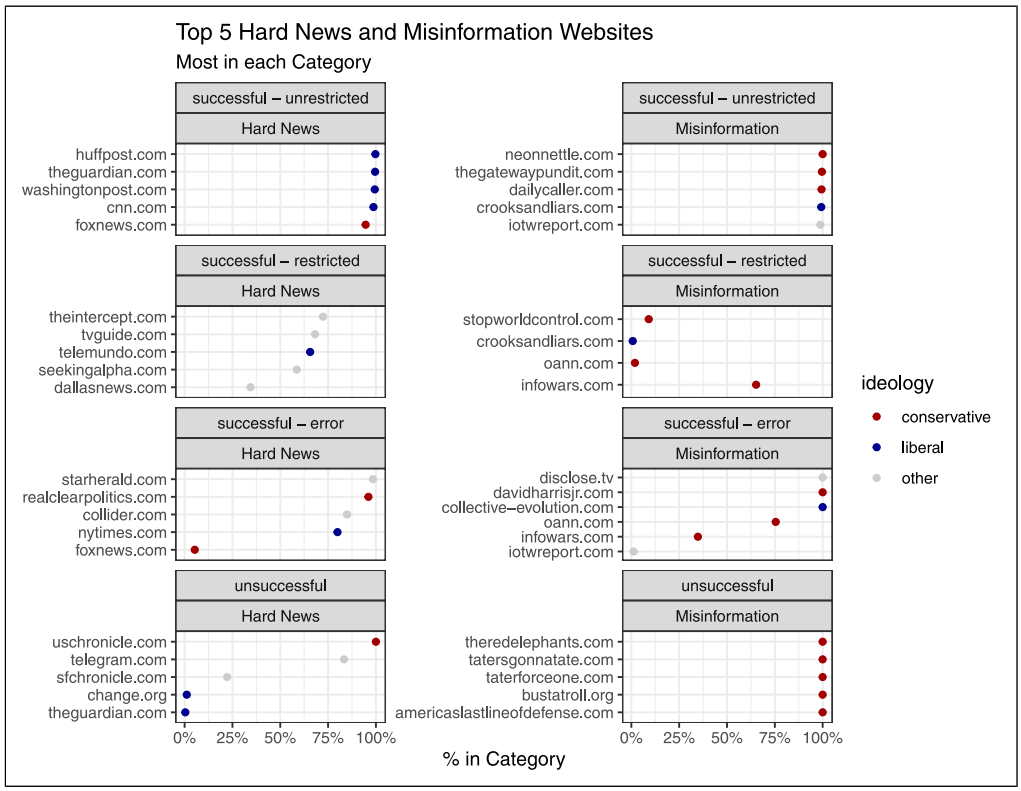

Figure 2. Top 5 Hard News and Misinformation Websites. Graph of the top five hard news and misinformation websites in each category. On the *x*-axis the percentage of the URLs from the given domain that fell into the category.

restricted content makes up just .7% of total requests in the first snapshot but makes up 2.9% in the third snapshot.

These results have implications for misinformation research. Given that misinformation is relatively rare (Dahlke et al., 2022; Guess et al., 2020; Moore et al., 2023a, 2023b), each piece of misinformation exposure is important. Misinformation researchers should work to document the content of misinformation as quickly as possible after its creation or exposure to preserve and study its contents. In particular, researchers should consider that some types of misinformation may be systematically more difficult to capture and either make special efforts to collect that content or consider the implications of potentially missing it.

The present research, however, has much broader implications for any researcher conducting web scraping. Based on these results, we have suggestions for how researchers should capture web scraping data and how to report such data in a manuscript.

First, we recommend leveraging a browser-based scraping infrastructure when collecting web data from URL traces. This infrastructure is so that URL content captured can more closely mirror the end user's behavior when they visit the page (e.g., through a web browser).

Second, we have identified key metrics that quantify potential errors associated with a web page's accessibility status. We encourage future research using scraped web data to report the percentage of web pages that fall into each category: successful and unrestricted content, successful and restricted content, restricted and an error returned, or unsuccessful. After calculating these rates, they should be reported consistently through a table, as modeled in Table 3, where there are at least two categories of websites (Type A, Type B, Type C, etc.). This sort of test has the flexibility to handle granular levels of data, even down to the web-domain level. For data with nested subgroups, we recommend a table such as Table 2. Crucially, we recommend a chi-squared test of the distributions to determine if the content distribution significantly differs across subgroups. If the distributions are significantly different, that suggests a systematic bias in one's data.

Third, when the chi-squared test is significant—and thus the data show systematic bias—we recommend that authors should do three things: 1) authors should consider whether this bias compromises their results or requires other methods to overcome the bias (e.g., recover inaccessible sites via an online archive), 2) conduct an error analysis to examine why some categories' metrics are different, and 3) note in the limitations of the study that there is potential bias that could influence inferences from the analysis. If critical to one's research questions, scholars may consider conducting scrapes that incorporate login credentials to sites that are inaccessible without them. In this study, we show that these additional steps may be necessary to increase coverage of successfully scraped sites. We note that there is no perfect sampling of websites, in the same way that sampling of human participants in studies is never perfectly representative of the underlying

Table 3. Reporting Table Template.

	Accessible			Inaccessible
URL Category	Successful - Unrestricted	Successful - Restricted	Successful - Error	Unsuccessful
Type A	__%	__%	__%	__%
Type B	__%	__%	__%	__%

Note: $\chi^2(\_\_) = \_\_$, $p = \_\_$. Distribution of websites in each accessibility category. Accessible websites are those in which the web scrape is successful. Inaccessible websites are those in which the web scrape is unsuccessful. Accessible websites are further categorized into unrestricted content in which the content is not restricted or returns an error, restricted websites in which the content sits behind a paywall, login page, or some other error message on the web page itself, and errors in which the web server returns an HTTP status code greater than 400.

population. Therefore, just as sampling metrics are always reported in human participant studies, we argue here that the metrics should always be reported for web scraping studies to give readers an understanding of the potential biases in a study's data. Hopefully, future meta-analytic work can use these standardized metrics to gain a more holistic understanding of the distribution of the metrics across the internet and websites of interest to scholars.

Conclusion

We examine the accessibility and inaccessibility of web scraping data from web-browsing logs of all hard news and misinformation websites that 1,238 individuals visited across 107k visits to hard news and misinformation websites. We find significant amounts of systematic bias in the scraped data. Misinformation web pages, particularly conservative ones, are more likely to be inaccessible. Hard news web pages, specifically liberal hard news web pages, are more likely to be accessible to restricted or returned an error. We suggest that future researchers should take care to consider and report the systematic biases in their own data by reporting on the accessibility statuses of their URLs in a standard way that makes clear the potential biases in one's data and allows for easy interpretation across studies.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Army Research Office Multidisciplinary University Research Initiative Award; W911NF-20-1-0252. R.D. is supported by graduate fellowship awards from Knight-Hennessy Scholars and Stanford Data Science Scholars at Stanford University. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

ORCID iD

Ross Dahlke https://orcid.org/0000-0002-5179-2525

Data Availability Statement

Data availability: Code and materials to replicate this analysis are available at osf.io/7beuv/.

Supplemental Material

Supplemental material for this article is available online.

Notes

1. newsguardtech.com.
2. These sites included: msn.com, news.yahoo.com, en.wikipedia.org, finance.yahoo.com, sports.yahoo.com, m.youtube.com, profootballtalk.nbcsports.com, bleacherreport.com, theringer.com, espn.com, weather.com, accuweather.com, vimeo.com, soccer.nbcsports.com, whitehouse.gov.
3. The "location," including IP address, of the headless browsers we used is at a large research institution on the West Coast of the United States.

References

Ananny, M., & Bighash, L. (2016). Why drop a paywall? Mapping industry accounts of online news decommodification. *International Journal of Communication*, *10*(2016), 3359–3380.

Arrese, Á. (2016). From gratis to paywalls: A brief history of a retro-innovation in the press's business. *Journalism Studies*, *17*(8), 1051–1067. https://doi.org/10.1080/1461670x.2015.1027788

Bach, R. L., Kern, C., Amaya, A., Keusch, F., Kreuter, F., Hecht, J., & Heinemann, J. (2021). Predicting voting behavior using digital trace data. *Social Science Computer Review*, *39*(5), 862–883. https://doi.org/10.1177/0894439319882896

Bainotti, L., Caliandro, A., & Gandini, A. (2021). From archive cultures to ephemeral content, and back: Studying instagram stories with digital methods. *New Media & Society*, *23*(12), 3656–3676. https://doi.org/10.1177/1461444820960071

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Political science. Exposure to ideologically diverse news and opinion on Facebook. *Science*, *348*(6239), 1130–1132. https://doi.org/10.1126/science.aaa1160

Baumgartner, S. E., Sumter, S. R., Petkevič, V., & Wiradhany, W. (2022). A novel iOS data donation approach: Automatic processing, compliance, and reactivity in a longitudinal study. *Social Science Computer Review*, *41*(4), 1456–1472. https://doi.org/10.1177/08944393211071068

Bayer, J. B., Ellison, N. B., Schoenebeck, S. Y., & Falk, E. B. (2016). Sharing the small moments: Ephemeral social interaction on snapchat. *Information, Communication & Society*, *19*(7), 956–977. https://doi.org/10.1080/1369118x.2015.1084349

Ben-David, A. (2016). What does the web remember of its deleted past? An archival reconstruction of the former yugoslav top-level domain. *New Media & Society*, *18*(7), 1103–1119. https://doi.org/10.1177/1461444816643790

Bilge, L., Kirda, E., Kruegel, C., & Balduzzi, M. (2011). *Exposure: Finding malicious domains using passive DNS analysis*. Ndss, 1–17.

Brandtzaeg, P. B. (2017). Facebook is no "great equalizer" a big data approach to gender differences in civic engagement across countries. *Social Science Computer Review*, *35*(1), 103–125. https://doi.org/10.1177/0894439315605806

Carah, N., & Shaul, M. (2016). Brands and instagram: Point, tap, swipe, glance. *Mobile Media & Communication*, *4*(1), 69–84. https://doi.org/10.1177/2050157915598180

Cavalcanti, L. H. C., Pinto, A., Brubaker, J. R., & Dombrowski, L. S. (2017). Media, meaning, and context loss in ephemeral communication platforms: A qualitative investigation on snapchat. Proceedings of the 2017 ACM Conference on computer Supported Cooperative work and social computing (pp. 1934–1945). https://doi.org/10.1145/2998181.2998266

Chen, W., & Quan-Haase, A. (2020). Big data ethics and politics: Toward new understandings. *Social Science Computer Review*, *38*(1), 3–9. https://doi.org/10.1177/0894439318810734

Choi, S. (2020). When digital trace data meet traditional communication theory: Theoretical/methodological directions. *Social Science Computer Review*, *38*(1), 91–107. https://doi.org/10.1177/0894439318788618

Chowdhury, F. A., Liu, Y., Saha, K., Vincent, N., Neves, L., Shah, N., & Bos, M. W. (2021). Ceam: The effectiveness of cyclic and ephemeral attention models of user behavior on social platforms. *Proceedings of the International AAAI Conference on Web and Social Media*, *15*(1), 117–128. https://doi.org/10.1609/icwsm.v15i1.18046

Christ, A., Penthin, M., & Kröner, S. (2021). Big data and digital aesthetic, arts, and cultural education: Hot spots of current quantitative research. *Social Science Computer Review*, *39*(5), 821–843. https://doi.org/10.1177/0894439319888455

Clark, H. H. (1996). *Using language*. Cambridge University Press.

Dahlke, R., & Hancock, J. (2022). *The effect of online misinformation exposure on false election beliefs*. OSF Preprints. https://doi.org/10.31219/osf.io/325tn

Dahlke, R., Moore, R., Forberg, P., & Hancock, J. (2022). *A mixed methods analysis of americans' QAnon website consumption*. OSF Preprints. https://doi.org/10.31219/osf.io/u6vgz

Dimitrova, D. V., & Bugeja, M. (2007). The half-life of internet references cited in communication journals. *New Media & Society*, *9*(5), 811–826. https://doi.org/10.1177/1461444807081226

Eck, A., Cazar, A. L. C., Callegaro, M., & Biemer, P. (2021). Big data meets survey science. In *Social Science Computer Review* (No. 4; Vol. 39, pp. 484–488). Sage Publications Sage CA.

Franklin, B. (2014). The future of journalism: In an age of digital media and economic uncertainty. In *Journalism Studies* (No. 5; Vol. *15*, pp. 481–499). Taylor & Francis.

Freelon, D. (2018). Computational research in the post-API age. *Political Communication*, *35*(4), 665–668. https://doi.org/10.1080/10584609.2018.1477506

Gertler, A. L., & Bullock, J. G. (2017). Reference rot: An emerging threat to transparency in political science. *PS: Political Science & Politics*, *50*(01), 166–171. https://doi.org/10.1017/s1049096516002353

Gil de Zuniga, H., & Diehl, T. (2017). Citizenship, social media, and big data: Current and future research in the social sciences. *Social Science Computer Review*, *35*(1), 3–9. https://doi.org/10.1177/0894439315619589

Guess, A. M. (2021). (Almost) everything in moderation: New evidence on americans' online media diets. *American Journal of Political Science*, *65*(4), 1007–1022. https://doi.org/10.1111/ajps.12589

Guess, A. M., Barberá, P., Munzert, S., & Yang, J. (2021). The consequences of online partisan media. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(14), Article e2013464118. https://doi.org/10.1073/pnas.2013464118

Guess, A. M., Nyhan, B., & Reifler, J. (2020). Exposure to untrustworthy websites in the 2016 US election. *Nature Human Behaviour*, *4*(5), 472–480. https://doi.org/10.1038/s41562-020-0833-x

Haenschen, K. (2020). Self-reported versus digitally recorded: Measuring political activity on facebook. *Social Science Computer Review*, *38*(5), 567–583. https://doi.org/10.1177/0894439318813586

Han, C., Kumar, D., & Durumeric, Z. (2022). On the infrastructure providers that support misinformation websites. *Proceedings of the International AAAI Conference on Web and Social Media*, *16*(1), 287–298. https://doi.org/10.1609/icwsm.v16i1.19292.

Hanley, H. W., Kumar, D., & Durumeric, Z. (2022). No calm in the storm: Investigating QAnon website relationships. *Proceedings of the International AAAI Conference on Web and Social Media*, *16*(1), 299–310. https://doi.org/10.1609/icwsm.v16i1.19293

Holz, T., Gorecki, C., Rieck, K., & Freiling, F. C. (2008). *Measuring and detecting fast-flux service networks*. Ndss.

Hounsel, A., Holland, J., Kaiser, B., Borgolte, K., Feamster, N., & Mayer, J. (2020). *Identifying disinformation websites using infrastructure features*. 10th USENIX Workshop on Free and Open Communications on the Internet (FOCI 20).

Jünger, J. (2021). A brief history of APIs: Limitations and opportunities for online research. In: *Handbook of computational social science* (vol. 2). Taylor & Francis.

Jungherr, A., Schoen, H., Posegga, O., & Jürgens, P. (2017). Digital trace data in the study of public opinion: An indicator of attention toward politics rather than political support. *Social Science Computer Review*, *35*(3), 336–356. https://doi.org/10.1177/0894439316631043

Klein, M., Van de Sompel, H., Sanderson, R., Shankar, H., Balakireva, L., Zhou, K., & Tobin, R. (2014). Scholarly context not found: One in five articles suffers from reference rot. *PLoS One*, *9*(12), Article e115253. https://doi.org/10.1371/journal.pone.0115253

Koehler, W. (1999). An analysis of web page and web site constancy and permanence. *Journal of the American Society for Information Science*, *50*(2), 162–180. https://doi.org/10.1002/(sici)1097-4571(1999)50:2<162::aid-asi7>3.0.co;2-b

Kreuter, F., Haas, G.-C., Keusch, F., Bähr, S., & Trappmann, M. (2020). Collecting survey and smartphone sensor data with an app: Opportunities and challenges around privacy and informed

consent. *Social Science Computer Review*, *38*(5), 533–549. https://doi.org/10.1177/0894439318816389

Krotov, V., & Silva, L. (2018). *Legality and ethics of web scraping*. Emergent Research Forum.

Kumar, D., Ma, Z., Durumeric, Z., Mirian, A., Mason, J., Halderman, J. A., & Bailey, M. (2017, April). Security challenges in an increasingly tangled web. In Proceedings of the 26th International Conference on world Wide web (pp. 677–684). https://doi.org/10.1145/3038912.3052686

Kumar, D. V., Kumar, B. T. S., & Parameshwarappa, D. (2015). URLs link rot: Implications for electronic publishing. *World Digital Libraries - An International Journal*, *8*(1), 59–66. https://doi.org/10.18329/09757597/2015/8105

Landers, R. N., Brusso, R. C., Cavanaugh, K. J., & Collmus, A. B. (2016). A primer on theory-driven web scraping: Automatic extraction of big data from the internet for use in psychological research. *Psychological Methods*, *21*(4), 475–492. https://doi.org/10.1037/met0000081

Li, F., Zhou, Y., & Cai, T. (2021). Trails of data: Three cases for collecting web information for social science research. *Social Science Computer Review*, *39*(5), 922–942. https://doi.org/10.1177/0894439319886019

Linell, P. (2004). *The written language bias in linguistics: Its nature, origins and transformations*. Routledge.

Lukito, Josephine, Brown, Megan A., Dahlke, Ross, Suk, Jiyoun, Yang, Yunkang, Zhang, Yini, Chen, Bin, Kim, Sang Jung, & Soorholtz, Kaiya (2023). The State of Digital Media Data Research, 2023. *Media & Democracy Data Coop*. http://dx.doi.org/10.26153/tsw/46177

Lyons, B. A. (2022). Why we should rethink the third-person effect: Disentangling bias and earned confidence using behavioral data. *Journal of Communication*, *72*(5), 565–577. https://doi.org/10.1093/joc/jqac021

McRoberts, S., Yuan, Y., Watson, K., & Yarosh, S. (2019). Behind the scenes: Design, collaboration, and video creation with youth. In: Proceedings of the 18th ACM International Conference on Interaction Design and Children, 173–184. https://doi.org/10.1145/3311927.3323134

Möller, J., van de Velde, R. N., Merten, L., & Puschmann, C. (2020). Explaining online news engagement based on browsing behavior: Creatures of habit? *Social Science Computer Review*, *38*(5), 616–632. https://doi.org/10.1177/0894439319828012

Moore, R. C., Dahlke, R., Bengani, P., & Hancock, J. T. (2023). *The consumption of Pink Slime journalism: Who, what, when, where, and why?* OSF Preprints. https://doi.org/10.31219/osf.io/3bwz6

Moore, R. C., Dahlke, R., & Hancock, J. T. (2023). Exposure to untrustworthy websites in the 2020 US election. *Nature Human Behaviour*, *7*(7), 1096–1105. https://doi.org/10.1038/s41562-023-01564-2

Myllylahti, M. (2014). Newspaper paywalls—the hype and the reality: A study of how paid news content impacts on media corporation revenues. *Digital Journalism*, *2*(2), 179–194. https://doi.org/10.1080/21670811.2013.813214

Myllylahti, M. (2017). What content is worth locking behind a paywall? Digital news commodification in leading australasian financial newspapers. *Digital Journalism*, *5*(4), 460–471. https://doi.org/10.1080/21670811.2016.1178074

Olmedilla, M., Martínez-Torres, M. R., & Toral, S. (2016). Harvesting big data in social science: A methodological approach for collecting online user-generated content. *Computer Standards & Interfaces*, *46*(2), 79–87. https://doi.org/10.1016/j.csi.2016.02.003

Pavlik, J. V. (2013). Innovation and the future of journalism. *Digital Journalism*, *1*(2), 181–193. https://doi.org/10.1080/21670811.2012.756666

Perdisci, R., & Lee, W. (2018). *Method and system for detecting malicious and/or botnet-related domain names*. Google Patents.

Perkel, J. M. (2015). The trouble with reference rot. *Nature*, *521*(7550), 111–112. https://doi.org/10.1038/521111a

Pickard, V., & Williams, A. T. (2014). Salvation or folly? The promises and perils of digital paywalls. *Digital Journalism*, *2*(2), 195–213. https://doi.org/10.1080/21670811.2013.865967

Praet, S., Guess, A. M., Tucker, J. A., Bonneau, R., & Nagler, J. (2022). What's not to like? Facebook page likes reveal limited polarization in lifestyle preferences. *Political Communication*, *39*(3), 311–338. https://doi.org/10.1080/10584609.2021.1994066

Reiss, M. V. (2022). Dissecting non-use of online news–systematic evidence from combining tracking and automated text classification. *Digital Journalism*, *11*(2), 363–383. https://doi.org/10.1080/21670811.2022.2105243

Revilla, M., Ochoa, C., & Loewe, G. (2017). Using passive data from a meter to complement survey data in order to study online behavior. *Social Science Computer Review*, *35*(4), 521–536. https://doi.org/10.1177/0894439316638457

Sjøvaag, H. (2016). Introducing the paywall: A case study of content changes in three online newspapers. *Journalism Practice*, *10*(3), 304–322. https://doi.org/10.1080/17512786.2015.1017595

Soffer, O. (2016). The oral paradigm and snapchat. *Social Media + Society*, *2*(3), 205630511666630. https://doi.org/10.1177/2056305116666306

Spence, P. R., & Burns, C. S. (2020). Retrieving arguments and support after publication: Archiving links in communication research. In *Communication Studies* (No. 5; Vol. 71, pp. 911–914). Taylor & Francis.

Stone-Gross, B., Cova, M., Cavallaro, L., Gilbert, B., Szydlowski, M., Kemmerer, R., Kruegel, C., & Vigna, G. (2009). Your botnet is my botnet: Analysis of a botnet takeover. Proceedings of the 16th ACM Conference on Computer and Communications Security, 635–647. https://doi.org/10.1145/1653662.1653738

Tyler, D. C., & McNeil, B. (2003). Librarians and link rot: A comparative analysis with some methodological considerations. *Portal: Libraries and the Academy*, *3*(4), 615–632. https://doi.org/10.1353/pla.2003.0098

Vázquez-Herrero, J., Direito-Rebollal, S., & López-García, X. (2019). Ephemeral journalism: News distribution through instagram stories. *Social Media + Society*, *5*(4), 205630511988865. https://doi.org/10.1177/2056305119888657

Villaespesa, E., & Wowkowych, S. (2020). Ephemeral storytelling with social media: Snapchat and instagram stories at the brooklyn museum. *Social Media + Society*, *6*(1), 205630511989877. https://doi.org/10.1177/2056305119898776

von Hohenberg, Bernhard Clemm, Stier, Sebastian, Cardenal, Ana S., Guess, Andrew M., Menchen-Trevino, Ericka, & Wojcieszak, Magdalena (2023). Analysis of Web Browsing Data: A Guide. *OSF Preprints*. https://doi.org/10.31219/osf.io/7hvap

Wells, C., & Thorson, K. (2017). Combining big data and survey techniques to model effects of political content flows in facebook. *Social Science Computer Review*, *35*(1), 33–52. https://doi.org/10.1177/0894439315609528

Wojcieszak, M., Leeuw, S. de, Menchen-Trevino, E., Lee, S., Huang-Isherwood, K. M., & Weeks, B. (2021). No polarization from partisan news: Over-time evidence from trace data. *The International Journal of Press/Politics*, *28*(3), 601–626. https://doi.org/10.1177/19401612211047194

Author Biographies

Ross Dahlke is a PhD Candidate in the Department of Communication at Stanford University (email: rdahlke@stanford.edu).

Deepak Kumar is a Postdoctoral Researcher in the Computer Science Department at Stanford University (email: kumarde@cs.stanford.edu).

Zakir Durumeric is an Assistant Professor in the Computer Science Department at Stanford University (email: zakir@cs.stanford.edu).

Jeffrey T. Hancock is the founding director of the Stanford Social Media Lab and the Harry and Norman Chandler Professor of Communication at Stanford University (email: hancockj@stanford.edu).