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

Ross Dahlke [1,2] Deepak Kumar [1,3] Zakir Durumeric [1,3] Jeffrey T. Hancock [1,2]

[1] Stanford University [2] Department of Communication [3] Department of Computer Science

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 persistence. 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 crawls or inaccessible via unsuccessful crawls. 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.

Introduction

Social science researchers are increasingly turning to observational web-tracking and digital trace data to understand patterns of exposure and effects of digital content. However, most social scientists do not collect digital trace data in real-time 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; 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 crawls into two main categories: accessible data in which the crawl is successful and inaccessible data in which the crawl is unsuccessful. We then further subcategorize accessible data from successful crawls 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), 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 – 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.

Literature Review

Accessible Data

From a technical perspective, accessible data can be accessed or retrieved through normal means, such as crawling a website. 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 scholars cannot find relevant reference material. This concern continues today (D Kumar et al., 2015; e.g., Klein et al., 2014; Perkel, 2015) and is shared across disciplines, for example, in communication (e.g., Dimitrova & Bugeja, 2007; Spence & Burns, 2020) and political science (e.g., Gertler & Bullock, 2017). Accessibility is important to scholars because it allows for the recreation and revisiting of the original content that scholars desire 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 crawl 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) in an effort 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 inaccessible.

These accessible-yet-restricted data are often under-considered. News organizations do not randomly construct paywalls; thus, content is not randomly inaccessible 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 be of interest to scholars. In addition, news organizations will occasionally temporarily drop their paywall for public emergencies, planned special events, and broader access for civically valuable content (Ananny & Bighash, 2016).

Of course, restricted data are not new. For example, one may have had to pay for print newspapers. What is new, however, is how researchers are attempting to access the data. While researchers in the past may have accessed the totality of news 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, inaccessible data pose additional problems for researchers above and beyond ephemerality 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 normal means. In the computer science security community, significant prior work has studied the ways in which adversarial actors cloak or hide malicious activity using Fast Flux Domains (Holz et al., 2008). These ephemeral 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 normal 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 normal means, making it potentially impossible for researchers to access and use the data without special methods or tools, such as a historical archive. 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., 2022) and to what effect (Dahlke & Hancock, 2022). One concern in misinformation research is that it has not accounted for ephemeral and inaccessible web-based misinformation. Many popular misinformation studies leverage lists of curated misinformation websites, but these websites are often unavailable 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, highlighting a possibility for 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 persistence 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 1,515 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 websites that are hard news, as defined by Baksy et al. (2015) and NewsGuard[1], and misinformation websites, as categorized by Moore et al. (2022). We assigned ideological labels to websites using NewsGuard's rating and classifications from Baksy et al. (2015). In addition, we only examined URLs that were to content webpages, i.e., 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 are 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.

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. In some cases, the browser crashed when visiting the URL. This

[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

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 crawl as unsuccessful. If the browser was able to retrieve some page content, we denoted that crawl 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 crawls 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 crawl 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 0.92 on the validation set. After applying this model to the entire set, we categorized 97,395 (91.3%) as successfully crawled with unrestricted content, 753 (0.7%) as successfully crawled with restricted content, 8,385 (7.9%) as successfully crawled with an error, and 152 (0.1%) as unsuccessfully crawled.

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).

Table 1: Percentage of Hard News and Misinformation URLs that are in each Category

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

Note:
$\chi^2(3) = 372.3$, $p < .001$

To investigate how stable our results are over time, we also crawled 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).

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 crawled 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 crawled but with restricted content and nearly three times as likely to be successfully crawled but returned an error. In contrast, misinformation web pages were nine times more likely to return an unsuccessful crawl than hard news pages.

Some of these findings are aligned with the conventional wisdom. For example, misinformation websites were more likely to be unsuccessfully crawled and, thus, inaccessible. However, some findings are surprising. One that stands out is that hard news is more likely to be successfully crawled 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 why 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 crawled 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 crawled 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 crawled 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

Table 2: Percentage of Hard News and Misinformation URLs that are in each category

URL Category	Ideology	Successful - Unrestricted	Successful - Restricted	Successful - Error	Unsccessful
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: $\chi^2(3) = 13.1$, $p = .005$

pages' categorizations change across the snapshots in Supplemental Materials C. We discuss the implications of these results below.

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 crawled 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 crawled but returned an error and to be unsuccessfully crawled. 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 crawls to *The New York Times*, a liberal hard news website, were successful but returned an error. Or, crawls 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 crawled but with restricted content or errors. However, misinformation web pages were much more likely to be unsuccessfully crawled. Looking at the ideological slant of the web pages, liberal hard news web pages are more likely to be successfully crawled but with an error than conservative hard news web pages. However, conservative misinformation

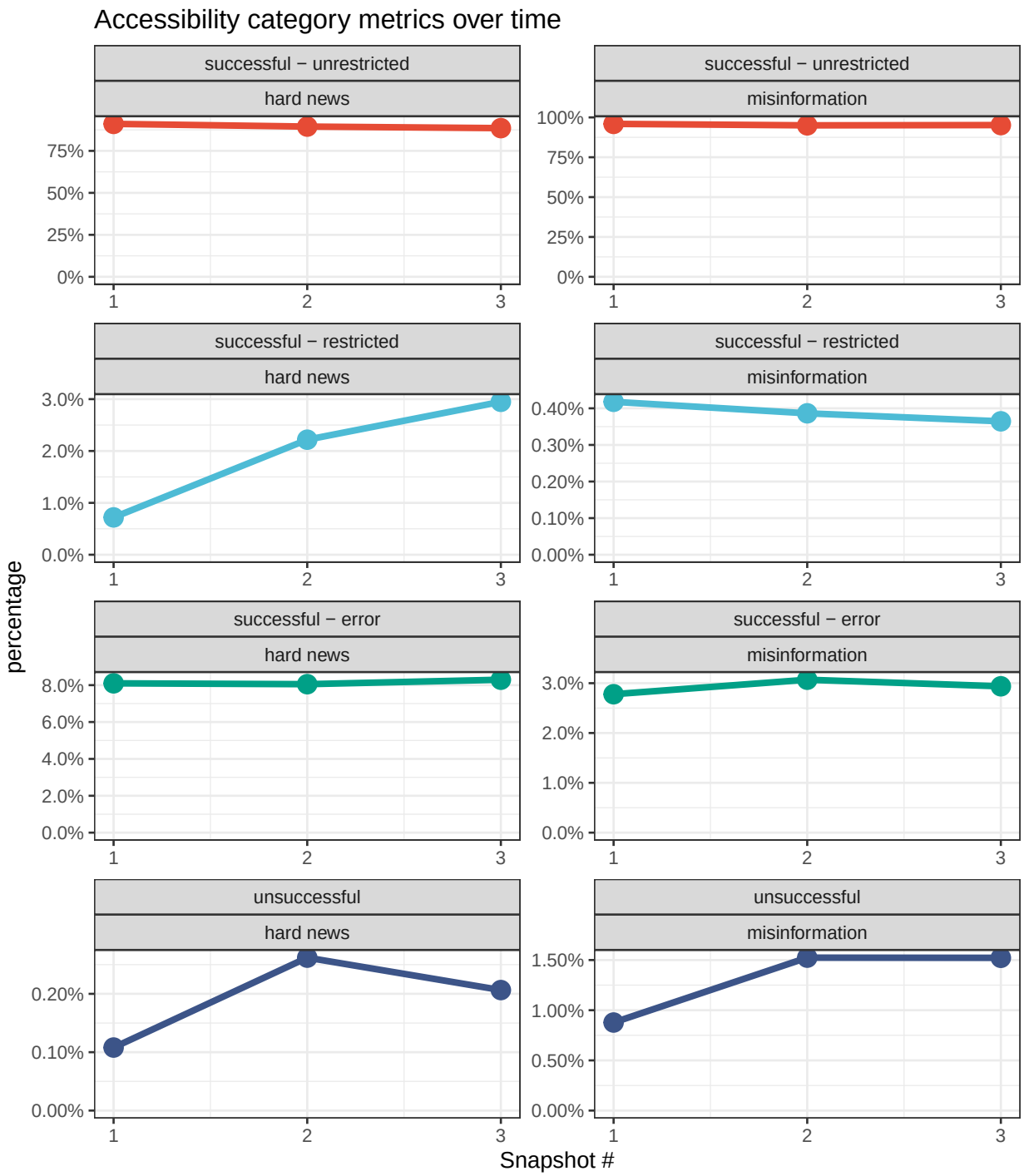

Figure 1: Percentage of URLs in our data set that are are in each category over time. On the x-axis is the snapshot number. On the y-axis is the percentage of the URLs that are in that category. Snapshot #1 was conducted one year after data collection. Snapshot #2 was conducted one-and-a-half years after data collection. Snapshot #3 was conducted two years after data collection.

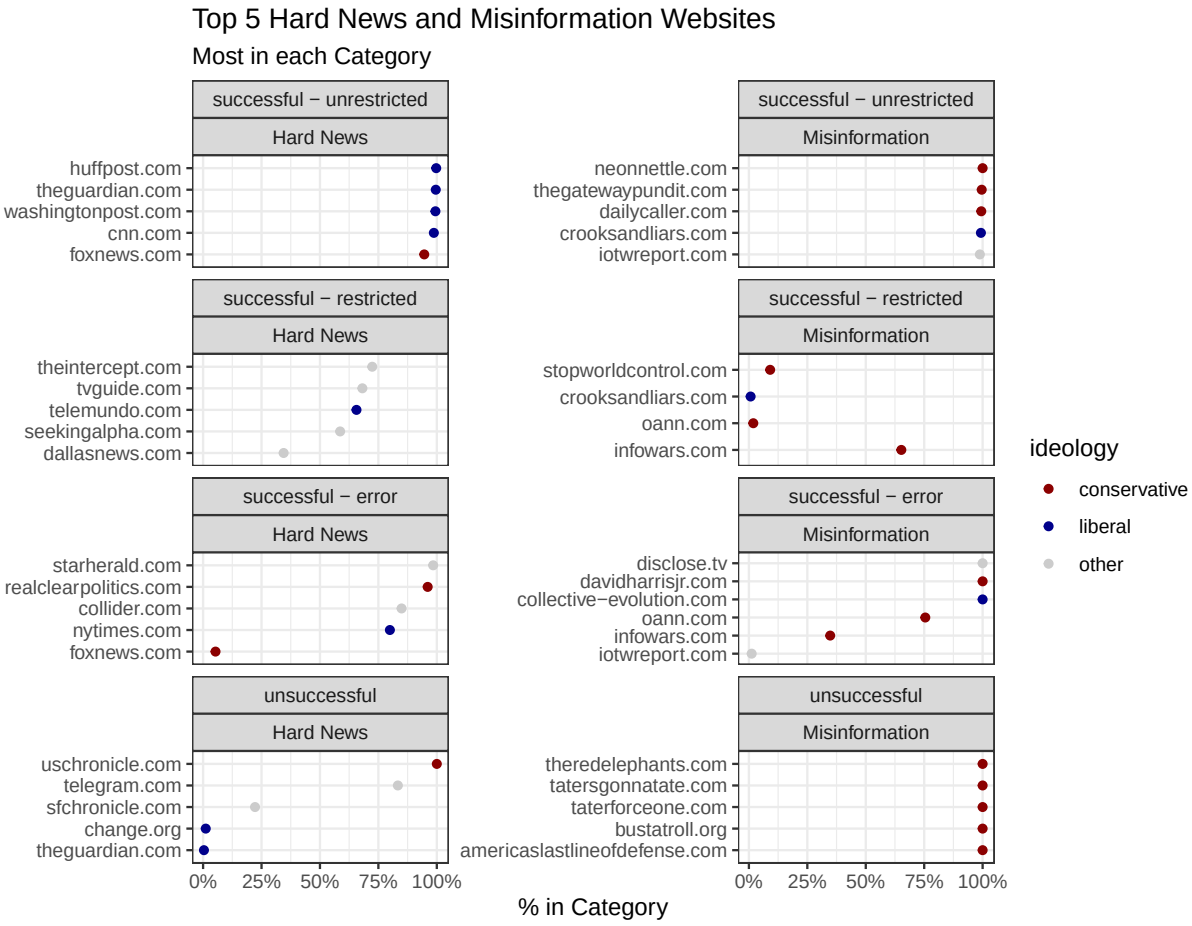

Figure 2: Graph of the top five hard news and misinformation websites that are ephemeral, inaccessible, and persistent On the x-axis is the percentage of the URLs from the given domain that fall into the category.

Table 3: PIE Table Template

URL Category	Successful - Unrestricted	Successful - Restricted	Successful - Error	Unsccessful
Type A	___%	___%	___%	___%
Type B	___%	___%	___%	___%

Note:
$\chi^2(\_\_\_) = \_\_\_, p = \_\_\_$

web pages were more likely to be crawled 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, restricted content makes up just 0.7% of total requests in the first snapshot but makes up 2.9% in the third snapshot.

These results have implications for misinformation research. Considering that misinformation is relatively rare (Dahlke et al., 2022; Guess et al., 2020; Moore et al., 2022), 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 in order 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 crawling 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. 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 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 unusability 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.

Supplemental Materials

A. Restricted Content Classifier

Often when conducting a web crawl, a web page is successfully returned but does not contain the content that one desires to study. For example, the content that may be returned is a paywall or some other sort of error message. To identify such URLs, we hand-coded returned content and trained a machine learning classifier that we applied to the remaining successfully returned URLs.

First, we trained two independent coders to identify "restricted" (i.e., content that is behind a paywall/ login or some other type of error). We instructed the coders to look for returned content that mentions a paywall, needing to pay to access content, a login page, or other messages that indicate an error or that the page was unavailable (e.g., "This page is not available right now"). The two coders achieved adequate intercoder reliability (Cohen's Kappa = .85).

After achieving adequate agreement, the coders hand-coded a random subset of 9,636 web pages. Of these web pages 11.6% were categorized as having restricted content, 88.4% were categorized as not having restricted content.

To train the classifier, we used 7,724 web page contents for the training set, 1,405 for the test set, and 507 for the validation set. The model achieved high accuracy with an F1 score of 0.92 on the validation set.

The data to train this classifier is available at osf.io/7beuv/.

B. Alternative statistical tests

As a robustness check, we examined whether the distribution of hard news and misinformation web pages remained significantly skewed when operating under different categorical groupings. We find that the distributions remain statistically significantly different when categorizing crawls into either successful or unsuccessful (Table S1), as well as successful and unrestricted content and other (i.e., successful but restricted, successful but error, and unsuccessful; Table S2).

Table S1: Percentage of Hard News and Misinformation URLs that are in each Category

URL Category	Successful	Unsuccessful
hard news	99.9%	0.1%
misinformation	99.1%	0.9%

Note:

$\chi^2(3) = 184.9$, $p < .001$

Table S2: Percentage of Hard News and Misinformation URLs that are in each Category

URL Category	Successful - Unrestricted	Other
hard news	91.1%	8.9%
misinformation	95.9%	4.1%

Note:

$\chi^2(3) = 134.8$, $p < .001$

C. Over-time results

We calculate the categorizations in the second (Table S3) and third (Table S4) snapshots to add more detail to our over-time analysis. In addition, we find that the distributions remain statistically significant across all three snapshots.

Also, we calculate the rates at which hard news and misinformation websites are in each of the categories across the three snapshots. (Table S5)

Table S3: Percentage of Hard News and Misinformation URLs that are in each Category in Snapshot 2

URL Category	Successful - Unrestricted	Successful - Restricted	Successful - Error	Unsccessful
hard news	89.5%	2.2%	8.1%	0.3%
misinformation	95.0%	0.4%	3.1%	1.5%

Note:

$\chi^2(3) = 446.7$, $p < .001$

Table S4: Percentage of Hard News and Misinformation URLs that are in each Category in Snapshot 3

URL Category	Successful - Unrestricted	Successful - Restricted	Successful - Error	Unsccessful
hard news	88.6%	2.9%	8.3%	0.2%
misinformation	95.2%	0.4%	2.9%	1.5%

Note:

$\chi^2(3) = 572.7$, $p < .001$

Table S5: Rates that hard news and misinformation websites are in each category across the three snapshots

Snapshot 1	Snapshot 2	Shapshot 3	Hard News %	Misinformation %
successful - unrestricted	successful - unrestricted	successful - unrestricted	86.45%	93.92%
successful - error	successful - error	successful - error	7.30%	2.52%
successful - unrestricted	successful - unrestricted	successful - restricted	1.73%	0.00%
successful - unrestricted	successful - restricted	successful - unrestricted	1.00%	0.02%
successful - unrestricted	successful - restricted	successful - restricted	0.97%	0.00%
successful - unrestricted	successful - unrestricted	successful - error	0.37%	0.13%
successful - restricted	successful - unrestricted	successful - unrestricted	0.37%	0.07%
successful - unrestricted	successful - error	successful - error	0.28%	0.17%
successful - error	successful - error	successful - unrestricted	0.23%	0.04%
successful - error	successful - unrestricted	successful - error	0.21%	0.00%
successful - error	successful - unrestricted	successful - unrestricted	0.18%	0.13%
successful - restricted	successful - restricted	successful - restricted	0.18%	0.35%
successful - unrestricted	successful - error	successful - unrestricted	0.16%	0.22%
successful - unrestricted	unsuccessful	successful - unrestricted	0.15%	0.72%
successful - unrestricted	successful - unrestricted	unsuccessful	0.12%	0.70%
unsuccessful	unsuccessful	unsuccessful	0.08%	0.72%
successful - restricted	successful - restricted	successful - unrestricted	0.05%	0.00%
successful - restricted	successful - unrestricted	successful - restricted	0.04%	0.00%
successful - restricted	successful - error	successful - error	0.02%	0.00%
successful - restricted	successful - error	successful - restricted	0.01%	0.00%
unsuccessful	successful - unrestricted	successful - unrestricted	0.01%	0.04%
unsuccessful	unsuccessful	successful - error	0.01%	0.00%
successful - restricted	successful - restricted	successful - error	0.01%	0.02%
successful - error	successful - error	unsuccessful	0.01%	0.04%
unsuccessful	successful - error	successful - error	0.00%	0.11%
successful - error	successful - error	successful - restricted	0.00%	0.00%
successful - error	unsuccessful	successful - error	0.00%	0.02%
successful - restricted	successful - unrestricted	successful - error	0.00%	0.00%
successful - unrestricted	successful - restricted	successful - error	0.00%	0.00%
successful - unrestricted	unsuccessful	successful - error	0.00%	0.00%
successful - unrestricted	unsuccessful	successful - restricted	0.00%	0.00%
successful - restricted	unsuccessful	successful - error	0.00%	0.00%
successful - restricted	unsuccessful	successful - restricted	0.00%	0.00%
successful - unrestricted	successful - error	successful - restricted	0.00%	0.00%
successful - unrestricted	unsuccessful	unsuccessful	0.00%	0.04%
successful - restricted	successful - error	successful - unrestricted	0.00%	0.00%
successful - unrestricted	successful - error	unsuccessful	0.00%	0.00%
unsuccessful	successful - unrestricted	successful - restricted	0.00%	0.02%

References

Ananny, M., & Bighash, L. (2016). Why drop a paywall? Mapping industry accounts of online news decommodification. *International Journal of Communication*, *10*, 22.

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.

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.

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.

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on facebook. *Science*, *348*(6239), 1130–1132.

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*, 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.

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.

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.

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

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*, 1934–1945.

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

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

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. *ICWSM*, 117–128.

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.

Clark, H. H. (1996). *Using language*. Cambridge university press.

D Kumar, V., Sampath Kumar, B., & Parameshwarappa, D. (2015). URLs link rot: Implications for electronic publishing. *World Digital Libraries-An International Journal*, *8*(1), 59–66.

Dahlke, R., & Hancock, J. (2022). *The effect of online misinformation exposure on false election beliefs*.

Dahlke, R., Moore, R., Forberg, P., & Hancock, J. (2022). *A mixed methods analysis of americans' QAnon website consumption*.

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

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: Los Angeles, 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.

Gertler, A. L., & Bullock, J. G. (2017). Reference rot: An emerging threat to transparency in political science. *PS: Political Science & Politics*, *50*(1), 166–171.

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.

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

Guess, A. M., Barberá, P., Munzert, S., & Yang, J. (2021). The consequences of online partisan media. *Proceedings of the National Academy of Sciences*, *118*(14), e2013464118.

Guess, A. M., Nyhan, B., & Reifler, J. (2020). Exposure to untrustworthy websites in the 2016 US election. *Nature Human Behaviour*, *4*(5), 472–480.

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

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*, 287–298.

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*, 299–310.

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, volume 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.

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), e115253.

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.

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.

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

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.

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.

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

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.

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

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

Moore, R., Dahlke, R., & Hancock, J. (2022). *Exposure to untrustworthy websites in the 2020 US election.*

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.

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.

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*, 79–87.

Pavlik, J. V. (2013). Innovation and the future of journalism. *Digital Journalism*, *1*(2), 181–193.

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.

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

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.

Reiss, M. V. (2022). Dissecting non-use of online news–systematic evidence from combining tracking and automated text classification. *Digital Journalism*, 1–21.

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.

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

Soffer, O. (2016). The oral paradigm and snapchat. *Social Media+ Society*, *2*(3), 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.

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.

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

Villaespesa, E., & Wowkowych, S. (2020). Ephemeral storytelling with social media: Snapchat and instagram

stories at the brooklyn museum. *Social Media+ Society*, *6*(1), 2056305119898776.

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.

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*, 19401612211047194.