PIE Metrics: Quantifying the Systematic Bias in the Ephemerality 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 unstated assumption the problematic assumption of persistence. To examine the extent to which this assumption may exist, we examine over 21 million URL-logged web browser visits from 1,515 participants over four months and record the degree to which hard news and misinformation URLs individuals visited were persistent, inaccessible, or ephemeral. While we find that the URLs collected are largely persistent, we find there are systematic biases in which URLs are ephemeral and inaccessible. For example, conservative misinformation URLs are more likely to be ephemeral than other types of misinformation. To standardize the reporting and understanding of the problematic assumption of persistence, we offer a set of metrics, $PersistenceRate$, $InaccessibilityRate$, $EphemeralityRate$ (PIE metrics), that future research should report when using 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). These post-facto content collection efforts, however, may be systematically biased by the inability to post hoc capture the content of many of these websites. For example, a website may have been deleted or behind a paywall. How much of the content digital scholars seek to analyze is ephemeral, which we define as content that computers can no longer connect to (i.e., the DNS records are missing), or inaccessible, which refers to content that no longer exists at the collected URLs (i.e., the website returns a 404 error or is behind a paywall)? Furthermore, why is some content ephemeral or inaccessible and others persistent?

To answer these questions, we leverage a dataset of 21 million URLs visited by a panel of 1,515 American adults to quantify the operative status–persistent, ephemeral, inaccessible–of hard news and misinformation web visits. 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 contributions: First, we estimate the levels of persistence, ephemerality, and inaccessibility of web pages that interest scholars studying digital behavioral data. Second, we investigate systematic differences in ephemeral content versus persistent web pages and show discrepancies across ideological content. Third, we investigate why this content is ephemeral and inaccessible. One possibility is that website content may expire due to financial constraints; another is that websites may adversarially "hide" content to avoid automatic detection (e.g., web cloaking). Ultimately, we recommend that scholars adopt a standardized set of reporting metrics that we call the PIE metrics ($PeristenceRate$, $InaccessibilityRate$, and $EphemeralityRate$) 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). Now, 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).

Most of these data are collected post hoc and, therefore, can be studied because they are *persistent.* In the field of communication, persistent communication is permanent, static, and atemporal (Linell, 2004, p. 8). Scholars can revisit preserved written text indefinitely, which has led to the exponential growth of big data forms of research in the social sciences. However, these data are prone to biases that social science researchers need to grapple with, for example, considering whose data is not being recorded and, thus, analyzed (Hargittai, 2020). More obviously, computational social science research that relies on digital trace data may be conducted only on persistent data because possibly only persistent media are being recorded. If specific media leave no digital trace but nonetheless play a role in people's experiences, these omissions could have consequences on all types of studies of digital trace data. Another issue is that just because digital trace data is persistent does not necessarily mean that it is accessible to researchers.

We call the reliance on digital trace data in computational social science the *problematic assumption of persistence*. This assumption is often unstated but assumes that the digital traces available to a researcher are representative and complete. We argue here that while a great deal of digital trace data is persistent and reasonably captures social behavior or experiences, there are also trace data that are ephemeral and inaccessible. Below we lay out these two other forms of trace data that we argue may undermine assumptions that trace data are representative and complete.

Ephemeral Communication

Despite the problematic assumption of persistence, communication scholars have long argued that human communication exists in one of two states: persistent or ephemeral (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.

Early internet scholars documented the extent to which web pages were persistent or ephemeral. For example, early estimates found that websites are generally persistent, with about 17.2% of web pages being ephemeral (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). Persistence is important to scholars because it allows for the recreation and revisitation of the original content that scholars desire to study.

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).

Inaccessible Communication

Past social science scholarship has considered the states of persistence and ephemerality of data and their implications for research. However, we argue that a third state of digital trace data is also important to computational social science: inaccessible data. Inaccessible data are not ephemeral in the sense that they continue to exist, but they are not fully persistent because they are not easily accessible. For example, paywall journalism creates communication that are often inaccessible. 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 ephemeral. 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.

This in-between state of inaccessible communication, persistent but not accessible, is often underconsidered. 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, inaccessible 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 can simply record a webpage snapshot before the page gets taken down and becomes ephemeral. Researchers must also decide how to access the web page's contents in addition to the capture step for inaccessible pages.

Persistence, Inaccessibility, Ephemerality, and the Study of Misinformation

In the present paper, we examine the inaccessibility relative to ephemerality and persistence 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 ephemerality and inaccessibility are systematic.

Quantifying Ephemerality and Inaccessibility on the internet

Some applied studies have already dealt with URL ephemerality and inaccessibility. For example, past research (Bastos & Mercea, 2019), in analyzing URLs on Twitter, found that over 50% of the hyperlinks they examined were "Dead links" to external (non-Twitter) websites. Using actual web pages that a representative panel of American adults visited, we re-engage with scholarly work on quantifying the internet's persistent, ephemeral, and inaccessible states.

This quantification is vital to social scientists studying human behavior on the internet because this content may not be randomly distributed across the persistent, ephemeral, and inaccessible categories. 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, likely just the persistent content, without fully considering the ephemeral and inaccessible content. 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, 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 three research questions:

RQ1: To what extent are hard news and misinformation website visits persistent, ephemeral, and inaccessible?

RQ2: Are there systematic biases in the websites and types of websites that are persistent, ephemeral, and inaccessible?

RQ3: Why may these biases exist?

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 participants. 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. All participants consented to the terms of the research, and YouGov compensated the participants. We collected survey responses and URL-level tracking data from 1,238 participants. These participants visited approximately 21 million websites throughout our data collection period.

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

[1] newsguardtech.com

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 107,783 unique URLs.

Method

One year after collecting URL logs, we visited each web page using a headless Google Chrome web browser. We did this to most closely simulate the real-world browsing experience of end-users using an Internet browser. We labeled the URLs we could not connect to at all as ephemeral. We labeled websites that responded to our request, but returned some sort of error code (e.g., 404 page not found) as inaccessible.

We then placed each URL into one of three buckets: ephemeral, inaccessible, or persistent. We labeled the URLs where the browser itself crashed when trying to connect, and we received no response data as ephemeral. For the inaccessible category, we sought to identify URLs that did return content but were not the content the end-user originally observed, for example, the content behind a paywall or login form.

To identify such content, we trained a machine learning classifier that could discern between inaccessible content and persistent content. For our training data, we hand-coded a random subset of 9,636 webpages (IRR, Cohen's Kappa = .85) that returned content for whether the page contained a message restricting access (e.g., "This page is not available right now.") and did not return the original content. We then used this hand-coded set to fine-tune a publicly available Huggingface BERT classifier to identify inaccessible vs. accessible 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 high accuracy with an F1 score of 0.92 on the validation set. After applying this model to the entire set, we categorized any remaining sites that returned an error message as inaccessible. We labeled the remaining sites as persistent. Out of our 107,783 initial sites, we categorized 165 (.15%) as ephemeral, 1,838 (1.7%) as inaccessible, and 102,804 (98.1%) as persistent.

We employ a standard chi-squared test on the distributions of persistence, inaccessibility, and ephemerality of various categories of 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.

Results

To answer RQ1, quantifying rates of ephemerality and inaccessibility, we find low rates of ephemerality (Table 1): only 0.1% of hard news websites and 0.9% of misinformation webpages are ephemeral. We also find low,

[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

Table 1: Percentage of Hard News and Misinformation URLs that are Persistent, Ephemeral, and Inaccessible

URL Category	Persistent	Ephemeral	Inaccessible
hard news	98.1%	0.1%	1.8%
misinformation	97.94%	0.94%	1.13%

Note:
$\chi^2(2) = 288.4$, $p < .001$

Table 2: Percentage of Hard News and Misinformation URLs that are Persistent, Ephemeral, and Inaccessible by Ideological Slant of Website

Ideological Slant	Persistent	Ephemeral	Inaccessible
hard news			
conservative	98.2%	0.1%	1.7%
liberal	98.7%	0.1%	1.2%
other	97.4%	0.1%	2.4%
misinformation			
conservative	97.39%	1.09%	1.51%
liberal	99%	0.0%	1%
other	98.56%	0.92%	0.52%

Note:
Hard news webpages: $\chi^2(3) = 27.2$, $p < .001$; Misinformation webpages: $\chi^2(3) = 9.7$, $p = .008$

albeit higher than ephemerality, rates of inaccessibility, with 1.8% of hard news and 1.1% of misinformation web pages being inaccessible. We also find that these results are relatively stable over time (see Supplemental Materials A), although the percentage of URLs that are inaccessible an ephemeral does slight increase after the initial snapshot, highlighting the importance of capturing content as quickly as possible.

For RQ2, which asks about the potential biases in ephemerality and inaccessibility, we find significant differences in ephemerality and inaccessibility rates in conservative versus liberal web pages (Table 2). Conservative hard news webpages are more likely to be ephemeral than liberal ones. Similarly, conservative misinformation webpages are more likely to be ephemeral and inaccessible than liberal ones. In other words, there are systematic biases in the ideological bent of the types of webpages that can be recovered for post hoc analysis.

Furthermore, specific domains are more likely to be ephemeral or inaccessible. As seen in Fig. 2, some websites are almost entirely ephemeral or inaccessible, and some domains' webpages are a mix between ephemeral, inaccessible, and persistent. Said another way, there are misinformation and hard news websites that are systematically difficult for researchers to record the content of.

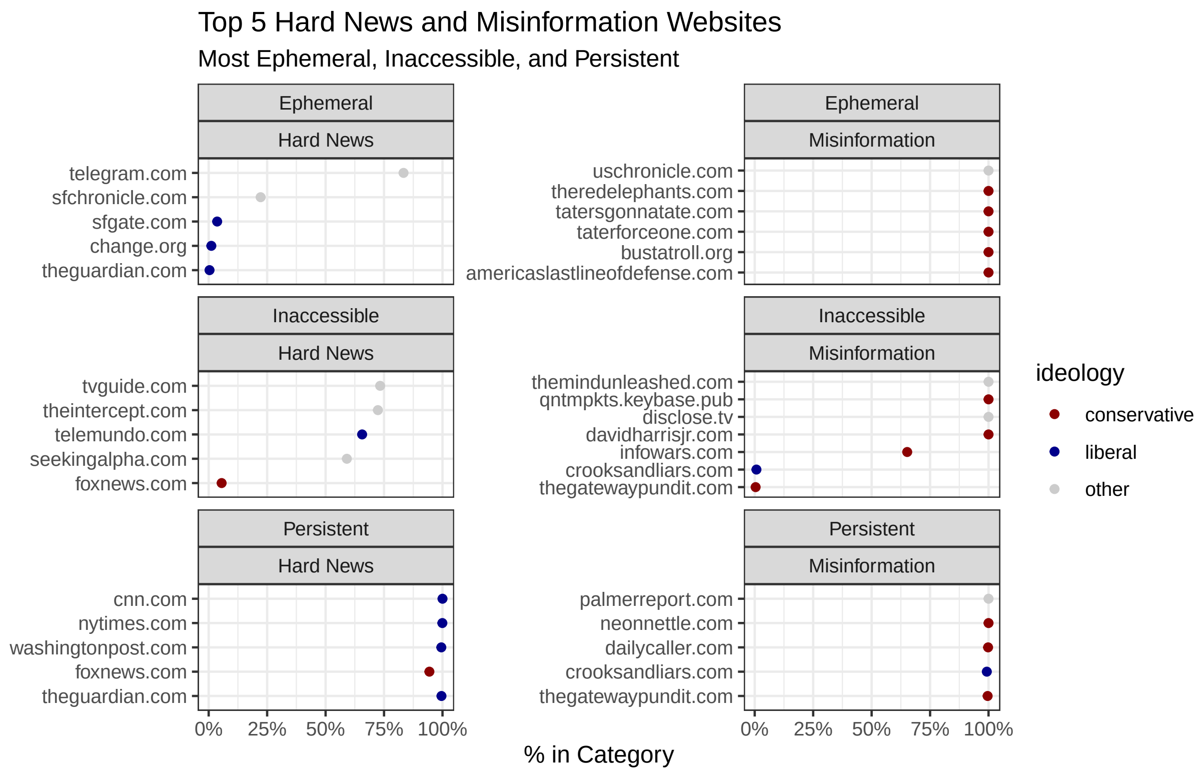

Figure 1: Graph of the top five hard news and misinformation websites that are ephemeral, inaccessible, and persistent On the x-axis the percentage of the URLs from the given domain that fell into the category.

Table 3: Ephemeral URL Errors

Error Type	Percentage	Explanation
Timeout	33.4%	The webpage did not load after one minute of waiting
HTTP Failure	23.5%	The webpage returned an error that rendered the browser unable to continue
Name Not Resolved	9.9%	The domain name for this URL no longer exists
Cert Common Name Invalid	7.5%	The website no longer has a valid HTTPS certificate.
Connection Aborted	5.7%	The connection was stopped by the server.
Connection Refused	3.9%	The server refused to respond to the request issued by the browser
Certificate Date Invalid	2.5%	The website no longer has a valid HTTPS certificate
SSL Protocol Error	2.4%	The server failed to properly establish an HTTPS connection.

Note:
Breakdown of the percentage of ephemeral URLs that returned each error code.

Ephemeral Website Error

To answer RQ3, why some websites are ephemeral, we turn to the technical details of the URLs we attempted to scrape. Websites may not return content at requested URLs for a myriad of reasons – for example, the page may no longer exist, the website may no longer exist, or changes to the website's configuration may render the page no longer accessible. We track and aggregate the top errors thrown by our crawling infrastructure for ephemeral URLs in Table 3. The most common reason a site is ephemeral is due to a timeout (33.4%). We stay on a single webpage for up to 60 seconds as a timeout window – after this point, we determine there is likely some unknown external factor (e.g., web server configuration error) that is preventing us from retrieving the page.

Other errors are more specific – 23.5% of ephemeral URLs failed due to an HTTP Error, which typically meant the server response failed in an unrecoverable way. While it is hard to ascertain exactly why a website operator would take down a website, it highlights that many such ephemeral articles are removed over time, and post hoc analyses would miss these URLs. Other errors point to fundamental web server failures – 9.9% of errors are Name Not Resolved, meaning that the domain itself has been taken down since the YouGov Pulse data was collected. Our examination of these errors suggests a complex set of factors that dictate URL ephemerality.

Table 4: Inaccessible URL Error Codes

Error Code	Percentage	Explanation
404	56.7%	404 means missing page – which could mean the page has disappeared from the server
200	18.9%	Page returned content, but typically contains a paywall or is otherwise inaccessible
500	10.4%	Unexpected Server Error
403	6.3%	403 means the page has forbidden access; the user may have had privilege that we do not have as researchers
410	4.8%	Resource has vanished and it is unlikely to come back (GONE)
400	2.4%	Server will not process the request

Note:
Breakdown of the percentage of inaccessible URLs that returned each error code.

Inaccessible Website Errors

A significant number of websites are inaccessible, meaning that the content of the page either sits behind a paywall or is no longer available. To better characterize these errors, we investigated the most common HTTP status codes for inaccessible websites (Table 4). The majority of status codes (56.7%) were 404 not found, which means the original page has gone missing and no longer appears on the server. The second most prominent error was a 200 OK (18.9%), which means that while the page was accessible by our crawler, the content of the page was behind a paywall. Other, less common errors include 500 (10.4%), 403 (6.3%), 410 (4.8%), and 400 (2.4%), all of which in some way either restrict access to the webpage or render it unavailable.

Discussion

The present study examined the operative status–the persistence, inaccessibility, and ephemerality–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 websites are more likely than misinformation websites to be inaccessible, but that misinformation websites are generally more likely to be ephemeral. When looking at the ideological slant of the websites, conservative misinformation was the most likely to be ephemeral. Broadly, ephemeral errors are due to misconfigured servers that either never return any content or, in some cases, cease to exist on the Internet altogether. Inaccessible errors often stem from paywalls, however, in some cases, websites may restrict certain articles or take them down altogether.

Table 5: PIE Table Template

Category	Persistent	Inaccessible	Ephemeral
Type A	___%	___%	___%
Type B	___%	___%	___%

Note:
Type A webpages: $\chi^2(\_\_\_) = \_\_\_$, $p = \_\_\_$; Type B webpages: $\chi^2(\_\_\_) = \_\_\_$, $p = \_\_\_$

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. In particular, we have identified three key metrics that quantify potential error associated with a web page's operative status. We encourage future research that uses scraped web data to report the PIE metrics. These metrics are: $PersistenceRate = \frac{p}{t}$ where p is the number of persistent web pages and t is the number of total web pages scraped; $InaccessibilityRate = \frac{i}{t}$ where i is the number of inaccessible web pages; and $EphemeralityRate = \frac{e}{t}$ where e is the number of ephemeral web pages.

Then, these PIE metrics should be reported in a consistent manner through a table, as modeled in Table 5, where there are at least two categories of websites (Type A, Type B, Type C, etc.). This sort of test is flexible to handle granular levels, for example, 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 distribution of the PIE metrics significantly differ across subgroups. If the distributions are significantly different, that suggests that there is systematic bias in one's data.

When this 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 ephemeral sites via an online archive), 2) conduct an error analysis to examine why some categories' PIE metrics are different, and 3) note in the limitation 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, similarly to how sampling metrics are always reported in human participant studies, we argue here that the PIE 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 PIE metrics across the internet and websites of interest to scholars.

Conclusion

We examine the persistence, inaccessibility, and ephemerality of web scraping data from web browsing data of all misinformation and hard news websites that 1,515 individuals visited across 21M URL-level visits. We find significant amounts of systematic bias in the scraped data. Misinformation, particularly conservative misinformation, web pages are more likely to be ephemeral. Hard news, specifically liberal hard news, web pages are more likely to be inaccessible. We suggest that future researchers should take care to consider and report the systematic biases in their own data by reporting the PIE metrics, $PersistenceRate$, $InaccessibilityRate$, and $EphemeralityRate$ in a standard way that makes clear the potential biases in one's data and allows for easy interpretation across studies.

Supplemental Materials

A. Stability of Results Over Time

We also tested the time stability of our results (see Figure S1). To do so, we repeated the same process outlined above at three different time periods post data collection: at one year, one-and-a-half years, and two full years. The rates of inaccessible websites and ephemeral websites slightly increase after the initial snapshot; this is likely because many websites (especially news websites) will transition older articles to archived content. However, it does highlight the importance of collecting content-related information as quickly as possible to the data collection date, as some content has a short life cycle on the Internet. Future work should examine not only which websites are likely to be inaccessible and ephemeral but also how these biases may be exacerbated as time goes on.

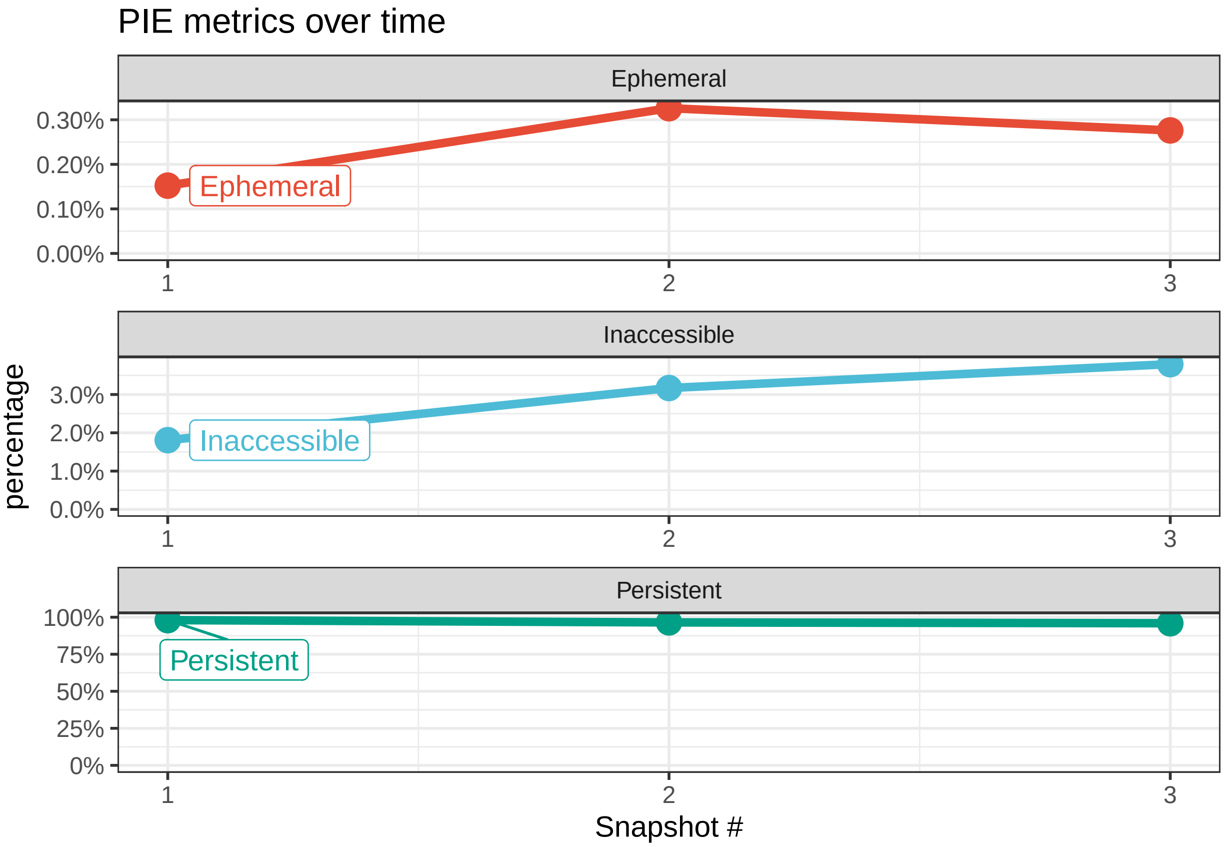

Figure S1: Percentage of URLs in our data set that are persistent, inaccessible, and ephemeral over time. On the x-axis is the snapshot number. Snapshot #1 was conducted one year after data collection. Snapshot #2 was conducted one-and-a-half years after data collection. Snapshot #3 was conducted two years after data collection.

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.

Bastos, M. T., & Mercea, D. (2019). The brexit botnet and user-generated hyperpartisan news. *Social Science Computer Review*, *37*(1), 38–54.

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.

Hargittai, E. (2020). Potential biases in big data: Omitted voices on social media. *Social Science Computer Review*, *38*(1), 10–24.

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.