Understanding Web Archiving Services and Their (Mis)Use on Social Media

Savvas Zannettou*, Jeremy Blackburn‡, Emiliano De Cristofaro†, Michael Sirivianos*, Gianluca Stringhini‡

*Cyprus University of Technology, †University College London, ‡University of Alabama at Birmingham
sa.zannettou@edu.cut.ac.cy, blackburn@uab.edu, {e.decristo, g.stringhini}@ucl.ac.uk, michael.sirivianos@cut.ac.cy

Abstract

Web archiving services play an increasingly important role in today’s information ecosystem, by ensuring the continuing availability of information, or by deliberately caching content that might get deleted or removed. Among these, the Wayback Machine has been proactively archiving, since 2001, versions of a large number of Web pages, while newer services like archive.is allow users to create on-demand snapshots of specific Web pages, which serve as time capsules that can be shared across the Web. In this paper, we present a large-scale analysis of Web archiving services and their use on social media, shedding light on the actors involved in this ecosystem, the content that gets archived, and how it is shared. We crawl and study: 1) 21M URLs from archive.is, spanning almost two years; and 2) 356K archive.is plus 391K Wayback Machine URLs that were shared on four social networks: Reddit, Twitter, Gab, and 4chan’s Politically Incorrect board (/pol/) over 14 months. We observe that news and social media posts are the most common types of content archived, likely due to their perceived ephemeral and/or controversial nature. Moreover, URLs of archiving services are extensively shared on “fringe” communities within Reddit and 4chan to preserve possibly contentious content. Lastly, we find evidence of moderators nudging or even forcing users to use archives, instead of direct links, for news sources with opposing ideologies, potentially depriving them of ad revenue.

Introduction

In today’s digital society, the availability and persistence of Web resources are very relevant issues. A substantial number of URLs shared on the Web becomes unavailable after some time as websites are shutdown or redesigned in a way that does not preserve old URLs – a phenomenon known as “link rot” (Koehler 2004). Moreover, content might be taken down by authorities on a legal basis, deleted by users who have shared it on social media, removed as per the “right to be forgotten”, etc. Overall, the ephemerality of Web content often prompts debate with respect to its impact on the availability of information, accountability, or even censorship.

In this context, an important role is played by services like the Wayback Machine (archive.org), which proactively archives large portions of the Web, allowing users to search and retrieve the history of more than 300 billion pages. At the same time, on-demand archiving services like archive.is have also become popular: users can take a snapshot of a Web page by entering its URL, which the system crawls and archives, returning a permanent short URL serving as a time capsule that can be shared across the Web.

Archiving services serve a variety of purposes beyond addressing link rot. Platforms like archive.is are reportedly used to preserve controversial blogs and tweets that the author may later opt to delete (Mondal et al. 2016). Moreover, they also reduce Web traffic toward “source URLs” when the original content is still accessible, thus depriving them of potential ad revenue streams (users do not visit the original site, but just the archived copy). In fact, anecdotal evidence has emerged that alt-right communities target outlets they disagree with by nudging their users to share archive URLs instead (Koebliger 2014), or discrediting them by pointing at earlier versions of articles (Ralph 2017).

Given the role in helping content persist, their use on social networks, as well as anecdotal evidence of their misuse in contexts where information could be weaponized (MacFarquhar 2016), archiving services are arguably impactful actors that should be thoroughly analyzed. To this end, this paper aims to shed light on the Web archiving ecosystem, aiming to answer the following research questions: How are archive URLs disseminated across popular social networks? What kind of content gets archived, by whom and why? Are archiving services misused in any way?

To answer these questions, we perform a large-scale quantitative analysis of Web archives, based on two data sources: 1) 21M URLs collected from the archive.is live feed, and 2) 356K archive.is plus 391K Wayback Machine URLs that were shared on four social networks: Reddit, Twitter, Gab, and 4chan’s Politically Incorrect board (/pol/). Our main findings include:

1. News and social media posts are the most common types of content archived, likely due to their (perceived) ephemeral and/or controversial nature.
2. URLs of archiving services are extensively shared on “fringe” communities within Reddit and 4chan to preserve possibly contentious content, or to refer to it without increasing the Web traffic to the source. We also find that /pol/ and Gab users favor archive.is over Wayback Machine (respectively, 15x and 16x), highlighting a particular use case in “controversial” online communities.
3. Web archives are exploited by users to bypass censorship policies in some communities: for instance, /pol/ users post archive.is URLs to share content from 8chan and Facebook, which are banned on the platform, or to circumvent accidental censorship of some news sources because of substitution filters (e.g., ‘smh’ becomes ‘baka’, so links to smh.com.au are unusable).

4. Reddit bots are responsible for posting a very large portion of archive URLs in the subreddits we study (respectively, 44% and 85% of archive.is and Wayback Machine URLs). This is due to moderators aiming to alleviate the effects of link rot on the platform; however, this pro-active archival of content also impacts traffic to archived sites originating from Reddit.

5. The _Donald subreddit systematically targets ad revenue of news sources with conflicting ideologies: moderation bots block URLs from those sites and prompt users to post archive URLs instead (e.g., nydailynews.com have 46% of their content censored). According to our conservative estimates, popular news site like the Washington Post lose yearly approx. $70K from their ad revenue because of the use of archiving services on Reddit.

### Related Work

**Web archives.** (Alnoamany et al. 2014) analyze 6M access logs from the Wayback Machine, aiming to understand what users are looking for, and why they use it. They find that users visit the site predominantly via referrals, and that they mostly look for English pages, while most popular country-specific domains are from Japan, Russia, and Germany. (Alonso et al. 2017) simulate a Web archiving service, studying social discourse through the URLs as well as relevant entities and metadata, by analyzing millions of tweets as well as a case study related to fake news. (Ainsworth et al. 2011) measure how much content is available on Web archiving services: they sample URL shorteners and search engines, query 12 public archives, and find that 35%-90% of URLs have at least one archived copy. Finally, (Hale, Blank, and Alexander 2017) assess whether the Wayback Machine archives a purely random sample of Web pages, finding some bias towards more visible and prominent pages.

**Archived content.** (Nikiforakis et al. 2012) study the evolution of JavaScript using historical data of 3M Web pages from the Wayback Machine, while (Lerner et al. 2016) analyze how Web trackers have evolved in the previous decade. (Soska and Christin 2014) predict whether an uncompromised website will become malicious also using the Wayback Machine. (Holzmann, Nejdll, and Anand 2016) study how the German Web has evolved in the past 18 years using data from the Wayback Machine for 100 popular domains, highlighting the exponential growth of the number of pages. Then, (Hackett, Parmanto, and Zeng 2004) characterize the accessibility of websites using a random sample of sites from the Wayback Machine between 1997 and 2002.

**Remarks.** In contrast to previous work, our analysis focuses on how archived content is shared on social networks, how archiving services are used, and by whom. To the best of our knowledge, our work is the first large-scale measurement to do so, as well as to provide an in-depth analysis of on-demand Web archiving services (such as archive.is).

### Background

We now provide an overview of the Web archiving services, as well as the social networks studied in this paper.

**Web Archives**

archive.is offers a free, on-demand archival service of Web pages: a user visits the service and enters a URL to be archived. It also acts as a link shortener which obfuscates the source URL, by generating a 5-character URL. For instance, http://archive.is/HVbU shows the snapshot of Google’s homepage, archived on July, 03, 2012 at 07:03:24.

**Wayback Machine.** Launched in 2001, the Wayback Machine archives a large portion of Web content, storing periodic snapshots of various pages. It mainly works through a proactive crawler, which visits various sites and captures a snapshot of the content. However, users can also trigger information archival on demand. When a page is archived, the _archive URL_ includes the _source URL_ as well as the archival time, e.g., https://web.archive.org/web/20100205062719/http://www.google.com/ has Google’s home page as of February 5, 2010, at 06:27:19 (UTC). In the rest of the paper, we refer to the URLs generated by archiving services as _archive URLs_ and to the archived URL as _source URLs_.

We opt to study the Wayback Machine and archive.is for a few reasons. First, they are popular services: as of January 2018, their Alexa Global Rank is, resp., 300 and 2,920. We also choose these two because of some important differences between them. The Wayback Machine is run by a 501(c)(3) non-profit organization, while archive.is is hosted by Russian provider Hostkey (only accessible via HTTP in Russia). Moreover, the former respects robots exclusion standards (even retroactively) and gives website owners the right to request removal of pages from the archive, while the latter only complies (albeit inconsistently) with DMCA take-down requests. Finally, archive.is is reportedly used in “fringe” Web communities within 4chan and Reddit, which are known for generating (Wendling 2016) and incubating (Jackson 2016) fake news, and for their influence on the information ecosystem (Zannettou et al. 2017).

**Social Networks**

**Twitter** is a micro-blogging social network where users broadcast short “tweets” to their followers. Search and discussion around specific topics is facilitated by hashtags, while the action of “retweeting” rebroadcasts a tweet.

**Reddit** is a social news aggregator site that lets users post URLs along with a title. Posts get up- or down-voted and this determines the order in which they are displayed on the site. Communities on Reddit are primarily formed via so-called “subreddits,” i.e., forums created by users and dedicated to specific topics (e.g., /r/politics or /r/The_Donald).

**4chan** is an imageboard discussion forum. We focus on the Politically Incorrect board (/pol/) – the main board for politics and world events – because archive.is is among the most
Gab is a social network launched in 2016 to “champion free speech, individual liberty, and the free flow of information online.” It is a hybrid of Twitter and Reddit: users can broadcast 300 character messages (called “gabs”) to their followers, while a voting system determines the popularity of content. Gab has been criticized for high degree of racism (Benson 2016), hate (Price 2017; Zannettou et al. 2018), and for attracting alt-right users that are banned from mainstream communities like Twitter (Wilson 2016).

Datasets

We now present our datasets as well as our data collection methodology. We perform two crawls: 1) archive.is URLs obtained from the live feed page and 2) Wayback Machine and archive.is URLs posted on four social networks, namely, Twitter, Reddit, Gab, and 4chan’s /pol/.

archive.is live feed. To gather a large dataset of archive.is generated URLs, we use the live feed page (http://archive.is/livefeed/), which provides a view of the archive based on archival time (e.g., the first page lists URLs archived in the previous 10 minutes). In Aug 2017, we crawl the first 100K pages of the live feed, acquiring 45.2M URLs, archived between Oct 7, 2015 and Aug 26, 2017.

Next, we visit the archive.is URLs, and scrape the content to get the archival time and the source URL. To avoid issues for the site operators, we throttle our crawler and do not visit all 45.2M URLs. Instead, we randomly sample them while ensuring temporal coverage, visiting 21.5M (48%) archive URLs, corresponding to 20.6M unique source URLs from 5.3M unique domains. Note that given the substantial size of our sample, which guarantees temporal coverage over almost two years, the resulting dataset is representative of the archive. In other words, our sampling strategy does not likely introduce substantial biases affecting our results.

Archive URLs posted on social networks. We search for archive.is and Wayback Machine URLs on Twitter, Reddit, and /pol/, between Jul 1, 2016 and Aug 31, 2017, and on Gab between Aug 1, 2016–Aug 31, 2017. We obtain the 4chan dataset from (Hine et al. 2017), the Gab one from (Zannettou et al. 2018), the Reddit one from pushshift.io, while, for Twitter, we rely on the 1% Streaming API.

Overall, the resulting dataset includes 50K posts from /pol/, 528K posts from Reddit, 7K posts from Gab, and about 9K tweets. Note that we have some gaps due to failure of our data collection infrastructure, specifically, there are 70 and 13 days missing for Twitter and /pol/, respectively.

Table 1: Overview of our datasets: number and percentage of posts that include archive URLs, unique number of archive URLs, source URLs, and source domains.

<table>
<thead>
<tr>
<th>Platform</th>
<th>Archive</th>
<th>#Posts with Archive</th>
<th>Archive</th>
<th>Source URLs</th>
<th>Source Filtered Domains</th>
</tr>
</thead>
<tbody>
<tr>
<td>Live Feed</td>
<td>archive.is</td>
<td>21,537,554</td>
<td>20,608,834</td>
<td>5,388,112</td>
<td>-</td>
</tr>
<tr>
<td>Reddit</td>
<td>archive.is</td>
<td>327,050 (2.9 · 10^{-6})</td>
<td>310,392</td>
<td>291,382</td>
<td>15,994 (35.70%)</td>
</tr>
<tr>
<td>Wayback</td>
<td>320,379 (2.8 · 10^{-6})</td>
<td>357,081</td>
<td>343,851</td>
<td>21,214 (17.20%)</td>
<td></td>
</tr>
<tr>
<td>/pol/</td>
<td>archive.is</td>
<td>46,912 (1.1 · 10^{-6})</td>
<td>36,277</td>
<td>33,824</td>
<td>3,970 (4.67%)</td>
</tr>
<tr>
<td>Wayback</td>
<td>3,848 (9.7 · 10^{-7})</td>
<td>2,325</td>
<td>2,207</td>
<td>976 (83.12%)</td>
<td></td>
</tr>
<tr>
<td>Gab</td>
<td>archive.is</td>
<td>6,602 (3.4 · 10^{-7})</td>
<td>5,943</td>
<td>5,773</td>
<td>1,300 (5.54%)</td>
</tr>
<tr>
<td>Wayback</td>
<td>478 (5.1 · 10^{-7})</td>
<td>361</td>
<td>349</td>
<td>240 (61.18%)</td>
<td></td>
</tr>
<tr>
<td>Twitter</td>
<td>archive.is</td>
<td>6,750 (3.1 · 10^{-7})</td>
<td>3,772</td>
<td>3,669</td>
<td>845 (8.23%)</td>
</tr>
<tr>
<td>Wayback</td>
<td>1,905 (9.0 · 10^{-7})</td>
<td>1,290</td>
<td>1,257</td>
<td>846 (7.49%)</td>
<td></td>
</tr>
</tbody>
</table>

Basic Statistics. In Table 1, we report statistics from our archive.is live feed crawl as well as the crawl of archive.is and Wayback Machine URLs shared on Twitter, Reddit, /pol/, and Gab. We report the number of posts with archive URLs, along with the percentage over the total number of posts, as well as the number of unique archive URLs, unique source URLs, unique source domains, and the percentage of URLs that are filtered out. Specifically, besides malformed URLs, we exclude, for archive.is, URLs unreachable between Aug 29 and Oct 7, 2017, while for Wayback Machine those pointing to types of information other than Web pages (e.g., images, videos, software, etc.).

Overall, /pol/ and Gab users often share Wayback Machine URLs that point to non-Web pages: around 83% and 61% of the total, respectively, suggesting that archive.is is used mostly for the dissemination of Web pages, while Wayback Machine is preferred for other content. Also, a high percentage of malformed archive.is URLs are shared on Reddit (35%), due to bots trying to pro-actively archive resources but failing. From the normalized percentages, we observe that Twitter users rarely share URLs from archiving services, while Reddit users do so from both archiving services. On /pol/ and Gab, we find 15 and 16 times, respectively, more archive.is URLs than Wayback Machine ones.

Ethical Considerations. Although we only collect publicly available data, we have obtained clearance by our institutional ethics review board. Datasets and backups are stored encrypted, and we make no attempt to harm, re-identify or de-anonymize users, following standard ethical principles (Rivers and Lewis 2014).

Cross-Platform Analysis

In this section, we present a cross-platform analysis of archive URLs collected from the archive.is live feed, as well as Wayback Machine and archive.is URLs shared on the four platforms. We focus on understanding what kind of content gets archived, the related temporal characteristics, and on assessing the availability of archived content from the source.

Source Domains

Live Feed. In Fig. 1(a), we plot the CDF of the number of distinct URLs per domain in our archive.is live feed dataset. The vast majority (90%) of domains only appear
once, while a few domains yield a large number of archive URLs – e.g., there are 1.2M distinct archive.is URLs for which twitter.com is the source domain. In Table 2, we report the top 20 source domains as well as the top 20 domain suffixes (Sx). Surprisingly, the top domain (11.8%) is actually the Wayback Machine’s archive.org. Mainstream social networks like Twitter and Facebook are also included, likely due to their (perceived) ephemeral nature, i.e., users want to preserve social network posts before they are removed or deleted. As for the suffixes, we observe that common ones, such as .com and .org, are the majority, followed by domains from Germany (.de) and Japan (.jp) with 7% and 5.6% of the URLs, respectively. This suggests that a substantial portion of archive.is’s user base might be in Germany and Japan.

**Social Networks.** In Fig. 1(b)–1(e), we plot the CDF of the number of distinct URLs per source domain in each dataset, finding that over 40% of the source domains only appear once. Wayback Machine generally archives more URLs per source domain than archive.is, although for Reddit the distributions are quite similar. Then, in Tables 3–6, we report the top 20 source domains observed on each platform, along with their archival fraction (AF), i.e., the number of times a source domain appears in an archive over the total number of times it appears in the dataset (either archived or not).

On all platforms except for Gab, the most popular domain archived through archive.is is the platform itself; e.g., archives of tweets are the most shared ones on Twitter. This also happens for Wayback Machine URLs, but only on Reddit. On Reddit, this may be due to meta-subreddits focused on the preservation and discussion of dramatic happenings, e.g., flame wars and intra-Reddit conflict, that would otherwise be lost when deleted by moderators after some time. These meta-subreddits tend to make use of bots that automatically archive drama submitted by their members.

Overall, we notice a strong presence of both mainstream (e.g., Washington Post) and alternative (e.g., Breitbart) news sources archived and shared on Reddit, /pol/, and Gab. Moreover, on /pol/, archive.is is often used for links to hypothes.is, a service that lets users annotate news articles, possibly due to the fact that /pol/ users often “unravel” conspiracy theories by researching and commenting on news articles. On Twitter, where the footprint of archive URLs is relatively low, we find a relatively large number of Japanese domains, which might possibly indicate a stronger presence of Japanese Twitter users relying on archives.

The AFs are quite low overall, implying that archiving services disseminate a small fraction of most domains. However, on /pol/, specific domains have extremely high AFs. For instance, we find that facebook.com (AF = 0.96) and 8ch.net (AF = 1.0) are marked as spam from /pol/, and posts including links to them are rejected, a phenomenon we refer to as **platform-specific censorship**. We manually analyze other domains with high AF values like hypothes.is, chetlyzarko.com, tdbming.com, etc., without finding evidence of censorship on /pol/. There is also “accidental” censorship on /pol/. For instance, the Australian newspaper smh.com.au, is affected because of a substitution filter (used for fun), which replace one word with another, as the word “smh” is automatically replaced on /pol/ with “baka.”

**URL Characterization.**

We now proceed to characterize the type of archived content. To this end, we extract the domain categories of source URLs using the free Virus Total API (virustotal.com), which we choose since it consolidates categories from multiple services (e.g., Bit Defender and Alexa). Although categorization is done at domain-level, results are presented at a per-URL level (a URL is assigned the same category as its domain) to capture the popularity of each domain.

**Live Feed.** Due to throttling enforced by the API, we are not able to categorize all the 20.6M source URLs in our archive.is live feed dataset. Therefore, we first aggregate URLs into their domain, then, we follow a sampling approach using: 1) the top 100K most popular domains in our dataset, which correspond to 15M source URLs, and 2) a sample of 121K domains drawn according to their empirical distribution in our archive datasets, resulting in 1.4M (7%) archive URLs.

In Fig. 2, we report the top 15 categories obtained from Virus Total for both samples. Note that Virus Total is unable to provide a category for 1% and 7% of the URLs for the two sets of domains that we checked, respectively. From Fig. 2(a), we observe that the most popular category is Reference Materials (23%), which is due to the fact that, as dis-
cussed earlier, many archive.is URLs archive Wayback Machine URLs. Other popular categories include Social Networks (15%), News Sources (14%), Education (13%), and Business (12%). Adult Content accounts for 4% of source URLs. Fig. 2(b) shows that, for the empirically distributed sample, the top 15 categories are slightly different, including Business (21%), News (13%), and Adult Content (12%).

Social Networks. Unlike the live feed dataset, we perform URL characterization for all source URLs (aggregated by domain) found on Reddit, /pol/, Gab, and Twitter, again using the Virus Total API. In Fig. 3, we report the top categories and their corresponding percentages for both archiving services (specifically, the union of categories that appear in the top 10 categories for each service). The Virus Total API is unable to provide a category for, on average, 1.5% and 9% of the archive.is and Wayback Machine URLs found on Reddit, /pol/, Gab, and Twitter, respectively. Overall, both archiving services are often used to disseminate URLs from news sources, social networks, and marketing sites on all social networks. However, there are interesting differences for the two archiving services: Education and Government URLs appear as top categories for the Wayback Machine (see Fig. 3(b), 3(c), and 3(d)), while sites that contain obscene language appear only for archive.is (see Fig. 3(c)). This suggests that the latter is used more extensively for “questionable” content. Moreover, we observe that Adult Content is among the top categories for all social networks except Twitter, while Gab and Reddit users often share archive URLs for domains related to Boards and Forums. Also, on /pol/, archive.is is used to archive and disseminate pages with obscene language, which is somewhat in line with previous observations (Hine et al. 2017) showing that /pol/ conversations often include hate speech.

Temporal Dynamics

Next, we study, from a temporal point of view, how archive URLs are created and shared on social networks.

Live Feed. In Fig. 4, we plot the day and hour of day of the creation of the archive.is URLs. Each day, between 1K and 10K URLs are archived (Fig. 4(a)), mostly between 11AM and 4PM UTC time, with a peak at 2PM (Fig. 4(b)), which seems to suggest that a great number of users may be located in Europe and the US. According to Alexa, the top country
for archive.is is the US, with 37% of the visitors.

Social Networks. Next, we measure the time interval between the archiving of a URL and its appearance on one of the four social networks. In Fig. 5, we plot the CDF of these time intervals, finding that the interval between archiving and sharing times of a URL ranges from a few seconds (in which case, Reddit/4chan/twitter/Gab users themselves might be creating the archive) to years. Reddit is the “fastest” platform for Wayback Machine URLs, mainly because of bots that actively archive URLs (as we show later in the paper), while for archive.is it is Gab.

We also focus on the top source domains shared via archive URLs: Figs 6–7 plot the CDF of the slack time of the top four domains for archive.is and Wayback Machine URLs, respectively. We define slack time as the time difference between the archival time on archiving services and the appearance on other social media. On Reddit, the top domains archived via Wayback Machine follow very similar distributions, likely due to bots, while for archive.is URLs, distributions vary, with the fastest domain being reddit.com itself. On Twitter, slack times vary for URLs archived via archive.is, with the fastest domain being Twitter and the slowest nhk.or.jp. The same applies for the Wayback Machine, with the fastest domain being Twitter and the slowest ameblo.jp. We also find that, on /pol/, archive.is URLs pointing to 4chan are considerably slower, suggesting that users are more interested in archiving the URL for persistence rather than sharing the content within /pol/.

Based on anecdotal observations, we believe users might be archiving threads with “evidence” for conspiracy theories/false narratives, and using them in the future to perpetuate mis/disinformation. This is not the case for news sources like the Washington Post or Guardian, as /pol/ users might be more focused on reducing Web traffic to the source domain instead (indeed we find users explicitly mentioning this when manually examining posts). Finally, on Gab, the faster domain is Twitter, and Reddit the slowest.

Original Content Availability

We then assess the availability of the original archived content; this allow us to determine whether users are archiving URLs that are subsequently deleted. To this end, we make an HTTP request for each source URL in our datasets, on Oct 14–21, 2017 for the live feed dataset, on Oct 4–5, 2017 for Reddit, Twitter, /pol/ datasets and on Jan 3, 2018 for Gab dataset. We treat each URL as unavailable if we receive HTTP codes 404/410/451/5xx, or if the request times out.

Live Feed. We find that 12% of the source URLs corresponding to archive URLs on archive.is live feed are no longer available. Domains with most unavailable content include twitter.com (6%), nhk.or.jp (6%), googleusercontent.com (3%), aaaaarg.fail (3%), and 4chan.org (3%).

Social Networks. In Reddit, source URLs corresponding to both archive.is and Wayback Machine are still available to a large degree (93% and 89% of them, respectively). This can be explained by the fact that Reddit bots archive URLs without considering the content. In /pol/, 82% and 66% of the original content is available for archive.is and Wayback Machine URLs, while on Gab it is 87% and 48%. Percentages decrease further for Twitter, with 76% and 49% for
archive.is and Wayback Machine URLs, respectively.

We also find that the top domains for which content is no longer available differ across platforms. Except for Gab, the top unavailable domain are the social networks themselves: 10%, 54%, and 28%, for Reddit, /pol/, and Twitter, respectively. URLs from cache servers (i.e., googleusercontent.com) and Twitter are also frequently unavailable; 9% and 10% in Reddit, 5% and 4% in /pol/, 8% and 28% in Twitter, and 12% and 19% in Gab, for googleusercontent.com and Twitter, respectively. We also note the presence of unavailable 8ch.net URLs (another ephemeral imageboard) with 5% and 4% on /pol/ and Gab, respectively.

Take-Aways

Overall, we find that archiving services play an important role in the information ecosystem, as they are used to preserve news sources as well as ephemeral or controversial content. Also, users on fringe communities such as /pol/ and Gab favor less popular Web archiving services like archive.is to archive and disseminate Web pages. This prompts questions as to why less popular, and seemingly less durable, archiving services are favored by more controversial communities like /pol/ and Gab. Although this would be out of the scope of this work, we do find one potential answer in that these communities also use archiving services to bypass platform-specific censorship policies.

We also observe that temporal dynamics of how archive URLs are shared on social networks differ according to their content: for instance, on /pol/, content from news sources has a considerably larger time lag between first appearing on the platform and archival compared to 4chan threads. Lastly, a non-negligible percentage of archived content is no longer available at the source; in particular, a substantial percentage of posts from social networks like Twitter are eventually deleted from the platform, yet remain stored in the archives.

Social-Network-based Analysis

In this section, we present a social-network-specific analysis by considering the fundamental differences of each platform. We analyze the users involved in the dissemination of archive URLs as well as the content that is shared along with those URLs. Lastly, we discuss a case study of ad revenue deprivation on Reddit. Due to space limitations, we exclude Twitter from our analysis, given the relatively small footprint of archive URLs on that platform (see Table 1).

User Base

Reddit. Our analysis shows that archiving services are extensively used by Reddit bots. In fact, 31% of all archive.is URLs and 82% of Wayback Machine URLs in our Reddit dataset are posted by a specific bot, namely, SnapshillBot (which is used by subreddit moderators to preserve “drama-related” happenings discussed earlier or just as a subreddit specific policy to preserve every submission). Other bots include AutoModerator, 2016VoteBot, yankbot, and autoldr. We also attempt to quantify the percentage of archive URLs posted from bots, assuming that, if a username includes “bot” or “auto,” it is likely a bot. This is a reasonable strategy since Reddit bots are extensively used for moderation purposes, and do not usually try to obfuscate the fact that they are bots.1 Using this heuristic, we find that bots are responsible for disseminating 44% of all the archive.is and 85% of all the Wayback Machine URLs that appear on Reddit between Jul 1, 2016 and Aug 31, 2017. We also use the score of each Reddit post to get an intuition of users’ appreciation

---

1This is somewhat evident from the list of Reddit bots available at https://www.reddit.com/r/autowikibot/wiki/redditbots
Table 7: Top 15 subreddits sharing archive.is and Wayback Machine URLs.

<table>
<thead>
<tr>
<th>Subreddit (archive.is) (%)</th>
<th>Subreddit (Wayback) (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>The_Donald 24.48%</td>
<td>EnoughTrumpSpam 31.82%</td>
</tr>
<tr>
<td>KotakuInAction 15.83%</td>
<td>MGTOW 7.38%</td>
</tr>
<tr>
<td>EnoughTrumpSpam 12.06%</td>
<td>SnapshillBotEx 7.19%</td>
</tr>
<tr>
<td>MGTOW 3.48%</td>
<td>undelete 5.90%</td>
</tr>
<tr>
<td>undelete 2.74%</td>
<td>SubredditDrama 5.50%</td>
</tr>
<tr>
<td>SubredditDrama 2.61%</td>
<td>Drama 5.03%</td>
</tr>
<tr>
<td>Drama 2.33%</td>
<td>Gamingcirclejerk 3.47%</td>
</tr>
<tr>
<td>Gamingcirclejerk 1.57%</td>
<td>ShitAmericansSay 1.63%</td>
</tr>
<tr>
<td>conspiracy 1.44%</td>
<td>TopMindsOfReddit 1.51%</td>
</tr>
<tr>
<td>MensRights 1.12%</td>
<td>TheBluePill 1.25%</td>
</tr>
<tr>
<td>savedyouaclick 1.00%</td>
<td>Buttonc0000 1.15%</td>
</tr>
<tr>
<td>politics 0.98%</td>
<td>AgainstHateSubreddits 1.00%</td>
</tr>
<tr>
<td>DerekSmart 0.76%</td>
<td>subredditcancer 0.99%</td>
</tr>
<tr>
<td>ShitAmericansSay 0.75%</td>
<td>The_Donald 0.95%</td>
</tr>
<tr>
<td>PoliticsAll 0.72%</td>
<td>badeconomics 0.75%</td>
</tr>
</tbody>
</table>

Figure 8: CDF of the scores of posts that include archive.is and Wayback Machine URLs.

Figure 9: CDF of cosine similarity of words obtained from LDA topics on Reddit and /pol/ threads.

/1. Subreddit (archive.is) (%) Subreddit (Wayback) (%)

Reddits. We then study how specific subreddits share URLs from archiving services. In Table 7, we report the top subreddits that share the most archive URLs from archive.is and the Wayback Machine. Among these, we find a variety of subreddits ranging from politics (e.g., EnoughTrumpSpam, The_Donald) to gaming (e.g., Gamingcirclejerk) and “drama-related” communities (e.g., SubredditDrama and Drama). Several subreddits prefer to use archive.is rather than the Wayback Machine, e.g., KotakuInAction, which historically covers the GamerGate controversy (Chatzakou et al. 2017), The_Donald, which discusses politics with a focus on Donald Trump, and Conspiracy, which focuses on various conspiracy theories.

Gab. On Gab, each post has a score that determines the popularity of the content. In Fig. 8(b), we report the CDF of the scores in posts that contain archive.is and Wayback Machine URLs, between August 2016 and August 2017. Once again, we also include a baseline, which is the scores for all the posts with URLs. We find that posts with Wayback Machine URLs have higher scores than those with archive.is URLs, and the baseline. Specifically, the mean score for Wayback Machine is 90, while for archive.is and the baseline the mean score is 35 and 30, respectively. This trend mirrors the one observed on Reddit for posts not authored by bots.

Next, we focus on the content that gets shared along with archive URLs on social platforms. We aim to evaluate if users share the same information for a given archive URL on multiple platforms. We do so using Latent Dirichlet Allocation (LDA). Due to space limitations, we only study Reddit and /pol/. Before running LDA, we exclude /pol/ and Reddit threads that contain less than 100 posts, and select only threads that have archive URLs appearing in both Reddit and /pol/ datasets; there are 425 such threads on /pol/ and 299 on Reddit. Next, we run LDA on all the posts within these threads and extract terms for 10 topics per thread. In Fig. 9, we plot the CDF of the cosine similarities on the terms extracted from LDA topics on the two platforms when sharing the same archive URLs. We observe that 80% of the terms have similarity under 0.3, which is expected given the fact that the two communities discuss topics in a different way. By manually observing terms with high similarity scores, we find that a number of them relate to well-known conspiracy theories, like the Seth Rich murder (Jonah Engel Bromwich 2017) or Pizzagate (Wendling 2016), as well as general discussions around politics (e.g., tensions between North Korea and the USA). Once again, this highlights that archiving services are used to preserve content related to controversial stories and conspiracy theories. A more detailed analysis is deferred to the extended version of the paper.

Content Analysis

Next, we focus on the content that gets shared along with archive URLs on social platforms. We aim to evaluate if users share the same information for a given archive URL on multiple platforms. We do so using Latent Dirichlet Allocation (LDA). Due to space limitations, we only study Reddit and /pol/. Before running LDA, we exclude /pol/ and Reddit threads that contain less than 100 posts, and select only threads that have archive URLs appearing in both Reddit and /pol/ datasets; there are 425 such threads on /pol/ and 299 on Reddit. Next, we run LDA on all the posts within these threads and extract terms for 10 topics per thread. In Fig. 9, we plot the CDF of the cosine similarities on the terms extracted from LDA topics on the two platforms when sharing the same archive URLs. We observe that 80% of the terms have similarity under 0.3, which is expected given the fact that the two communities discuss topics in a different way. By manually observing terms with high similarity scores, we find that a number of them relate to well-known conspiracy theories, like the Seth Rich murder (Jonah Engel Bromwich 2017) or Pizzagate (Wendling 2016), as well as general discussions around politics (e.g., tensions between North Korea and the USA). Once again, this highlights that archiving services are used to preserve content related to controversial stories and conspiracy theories. A more detailed analysis is deferred to the extended version of the paper.

Ad Revenue Deprivation

During our experiments, we find evidence that at least one Reddit bot, AutoModerator, is used to remove links to unwanted domains and nudge users to share archive.is instead. In particular, it posts: “Your submission was removed because it is from cnn.com, which has been identified as a
Table 8: Number and percentage of submissions deleted from The_Donald with links to different news sources.

<table>
<thead>
<tr>
<th>News Source</th>
<th>Count (%)</th>
<th>News Source</th>
<th>Count (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>washingtonpost.com</td>
<td>3,814</td>
<td>change.org</td>
<td>96</td>
</tr>
<tr>
<td>cnn.com</td>
<td>3,354</td>
<td>huffpost.com</td>
<td>62</td>
</tr>
<tr>
<td>nydailynews.com</td>
<td>1,070</td>
<td>fusion.net</td>
<td>60</td>
</tr>
<tr>
<td>huffingtonpost.com</td>
<td>978</td>
<td>cnn.it</td>
<td>58</td>
</tr>
<tr>
<td>nationalreview.com</td>
<td>774</td>
<td>alternet.org</td>
<td>26</td>
</tr>
<tr>
<td>theblaze.com</td>
<td>704</td>
<td>buzzfeed.com</td>
<td>16</td>
</tr>
<tr>
<td>salon.com</td>
<td>588</td>
<td>todayivies.com</td>
<td>4</td>
</tr>
<tr>
<td>vice.com</td>
<td>373</td>
<td>fullycucked.com</td>
<td>4</td>
</tr>
<tr>
<td>vox.com</td>
<td>372</td>
<td>usanewssite.g</td>
<td>10.52%</td>
</tr>
<tr>
<td>weeklystandard.com</td>
<td>233</td>
<td>northcrane.com</td>
<td>1</td>
</tr>
<tr>
<td>politifact.com</td>
<td>185</td>
<td></td>
<td>0.13%</td>
</tr>
</tbody>
</table>

This kind of notification appears in five different subreddits that discuss mainly politics and news, specifically, The_Donald, Mr_Trump, TheNewRight, Vote_Trump, and Republicans. In particular, in The_Donald, there are 13K such comments. AutoModerator blocks URLs from 23 news sources likely to be considered as anti-Trump by that community. In Table 8 we report the number of submissions deleted for each of the sources, along with the percentage over all submissions that include that source. Mainstream news outlets like Washington Post and CNN are the top domains that get removed from The_Donald (3.8K and 3.3K submissions, respectively), and this happens slightly less than half the times (44% and 39% of the submissions, respectively). Interestingly, only URLs posted via the URL submission field are censored by AutoModerator, but not URLs that are inserted as part of the title field.

Table 9: Top 20 domains with the largest ad revenue losses because of the use of archiving services on Reddit. We report an estimate of the average monthly visits from Reddit and the monthly ad loss.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Visits</th>
<th>Loss ($)</th>
<th>Domain</th>
<th>Visits</th>
<th>Loss ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>washingtonpost.com</td>
<td>79,880</td>
<td>5,928</td>
<td>wsj.com</td>
<td>11,389</td>
<td>845</td>
</tr>
<tr>
<td>cnn.com</td>
<td>70,483</td>
<td>5,231</td>
<td>breitbart.com</td>
<td>11,357</td>
<td>842</td>
</tr>
<tr>
<td>nytimes.com</td>
<td>46,442</td>
<td>3,446</td>
<td>bbc.com</td>
<td>10,708</td>
<td>794</td>
</tr>
<tr>
<td>huffingtonpost.com</td>
<td>27,125</td>
<td>2,013</td>
<td>salon.com</td>
<td>10,364</td>
<td>769</td>
</tr>
<tr>
<td>thehill.com</td>
<td>18,643</td>
<td>1,383</td>
<td>buzzfeed.com</td>
<td>10,359</td>
<td>768</td>
</tr>
<tr>
<td>theguardian.com</td>
<td>16,376</td>
<td>1,215</td>
<td>foxnews.com</td>
<td>9,638</td>
<td>715</td>
</tr>
<tr>
<td>politico.com</td>
<td>15,774</td>
<td>1,170</td>
<td>yahoo.com</td>
<td>9,497</td>
<td>704</td>
</tr>
<tr>
<td>dailymail.co.uk</td>
<td>14,442</td>
<td>1,071</td>
<td>latimes.com</td>
<td>9,277</td>
<td>688</td>
</tr>
<tr>
<td>dailycaller.com</td>
<td>12,735</td>
<td>945</td>
<td>vox.com</td>
<td>8,976</td>
<td>667</td>
</tr>
<tr>
<td>google.com</td>
<td>11,576</td>
<td>859</td>
<td>washingtontimes.com</td>
<td>8,862</td>
<td>657</td>
</tr>
</tbody>
</table>

Loss, for each domain, based on the average CPM value as well as the conservative estimate of the visits using the up- and down-votes. Overall, replacing URLs with archive URLs, as done, e.g., by the AutoModerator bot, yields an estimate of $30K per month in revenue loss (for the top 20 domains in terms of views). This is detailed in Table 9, where we break down the estimate for each of the top 20 revenue-deprived domains.

On a purely pragmatic level, consider that our conservative estimate of ad revenue deprivation is around $70K per year for the Washington Post alone. Although a more detailed impact analysis is out of the scope of this work, we suspect that even $70K could have a real world effect, e.g., on intern budgets or even early career hires.

Take-Aways

In summary, our social-network-specific analysis shows, among other things, that moderation bots on Reddit proactively leverage Web archiving services to ensure that content shared on their community persists. In particular, we find that 44% and 85% of archive.is and Wayback Machine URLs are shared by Reddit moderation bots, respectively. Also, Web archiving services are extensively used for the archival and dissemination of content related to conspiracy theories (e.g., Pizzagate) as well as other world events related to politics (e.g., tensions between North Korea and the USA), thus suggesting that these services play an important role in the (false) information ecosystem and need to be taken into account when designing systems to detect and contain the cascade of misinformation on the Web. Finally, we find evidence that moderators from specific subreddits force users to misuse Web archiving services so as to ideologically target certain news sources by depriving them of traffic and potential ad revenues. We also provide a best-effort conservative estimate of ad revenue loss of popular news sources showing that they can lose up to $70K per year.

Discussion & Conclusion

This paper presented a large-scale analysis of the use of Web archiving services, such as archive.is and the Wayback Machine, on social media. Our study is based on two data crawls: 1) 21M URLs obtained from the archive.is live feed, covering almost two years, and 2) 356K archive.is plus 391K Wayback Machine URLs that were shared, over 14 months, on Reddit, Twitter, Gab, and 4chan’s Politically Incorrect

---

https://www.statista.com/statistics/308015
board (/pol/). We showed that these services are extensively used to archive and disseminate news, social network posts, and controversial content—in particular by users of fringe communities within Reddit and 4chan. We also found that users not only rely on them to ensure persistence of Web content, but also to bypass certain censorship policies of some social networks. Some subreddits, as well as 4chan’s /pol/, actually nudge or force users to share archive URLs instead of direct link to news sources they perceive as having contrasting ideologies, taking away potentially hundreds of thousands of dollars in ad revenue. Overall, our measurements highlight the importance of archiving services in the Web’s information and ad ecosystems, and the need to carefully consider them when studying social media.

Our work also points to several open research avenues. For instance, future work could better understand the role of archiving services in the dis/misinformation ecosystem, e.g., with respect to the content that gets archived and the context in which archive URLs are disseminated. Moreover, further work could shed light on the actors archiving specific URLs in specific contexts, as well as how much traction they get on Web communities like Twitter and Reddit. Finally, we believe that a deeper dive into the socio-technical and ethical implications of archiving services is warranted: they serve a crucial role in ensuring that Web content persists, but do so without regard to (and often in spite of) the rights and consent of content producers.

Acknowledgments. This project has received funding from the European Union’s Horizon 2020 Research and Innovation program under the Marie Skłodowska-Curie ENCASE project (GA No. 691025). This work reflects only the authors’ views and the Commission are not responsible for any use that may be made of the information it contains.

References


Alnoamany, Y.; Alsum, A.; Weigle, M. C.; and Nelson, M. L. 2014. Who and what links to the Internet Archive. JDL.


Barford, P.; Canadi, I.; Krushevskaja, D.; Ma, Q.; and Muthukrishnan, S. 2014. Adscape: harvesting and analyzing online display ads. In WWW.

Benson, T. 2016. Inside the Twitter for racists: Gab the site where Milo Yiannopoulos goes to troll now. https://goo.gl/Yqv4Ue.


Gilbert, E. 2013. Widespread underprovision on Reddit. In CSCW.


Holzmann, H.; Nejdl, W.; and Anand, A. 2016. The Dawn of today’s popular domains: A study of the archived German Web over 18 years. In JCDL.


