Savvas Zannettou, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini,
Understanding Web Archiving Services and Their (Mis)Use on Social Media,
Proceedings of ICWSM 2018.

Presentation 1
Grant Atkins
September 24, 2020

Old Dominion University
Web Archiving Forensics
CS 895

Purpose of the paper

- 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 (e.g. avoiding ads, blocking site traffic)?

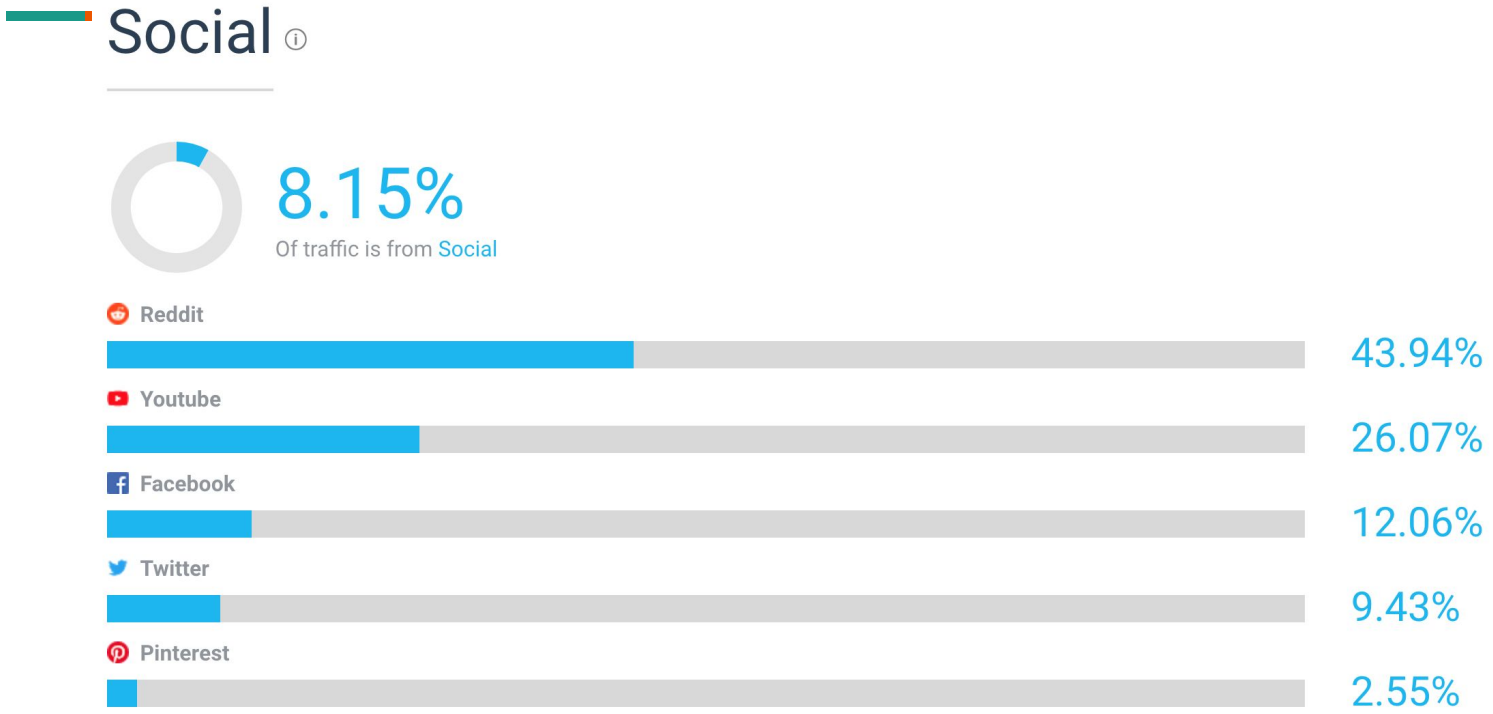
The Vice Ad Argument (2014)

MOTHERBOARD
TECH BY VICE

Dear GamerGate: Please Stop Stealing Our Shit

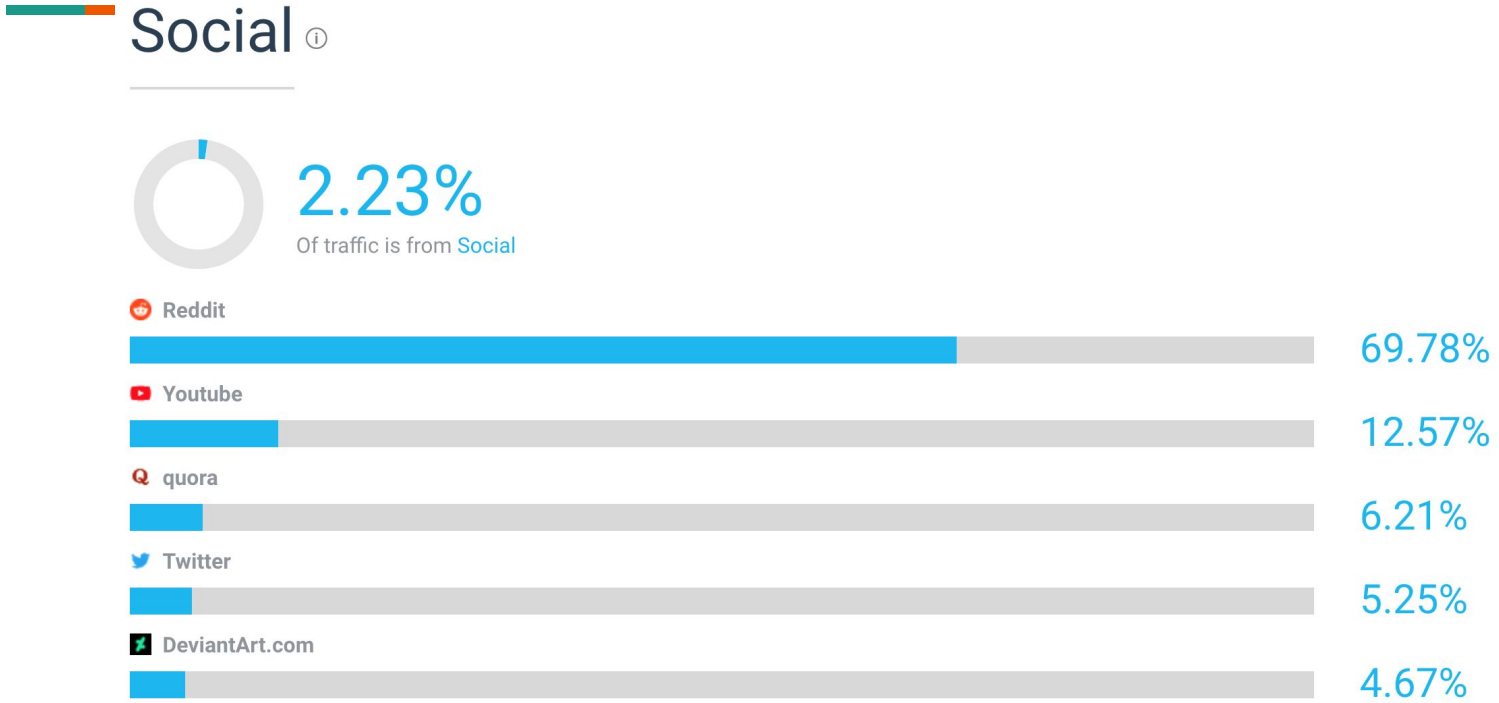
https://www.vice.com/en_us/article/ypw5mj/dear-gamergate-please-stop-stealing-our-shit

Looking at archive site traffic today - Archive.org



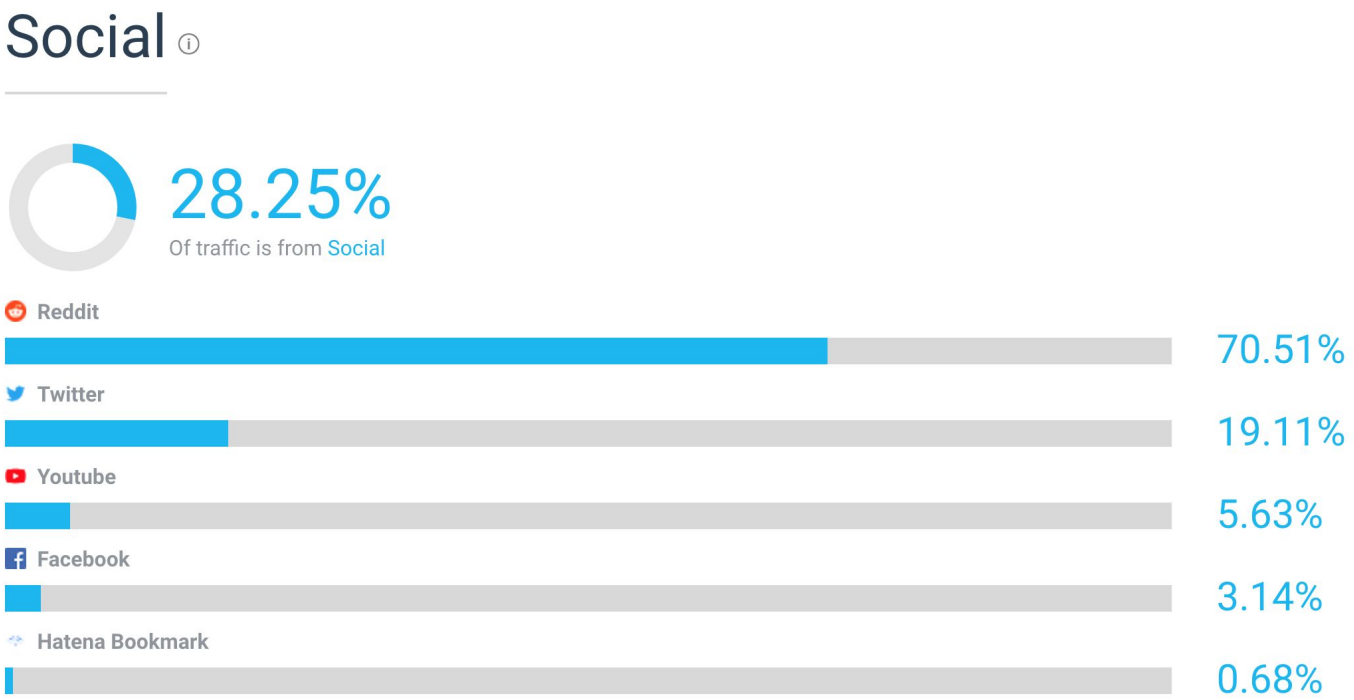
https://www.similarweb.com/website/archive.org/#social

Looking at archive site traffic today - Archive.today



https://www.similarweb.com/website/archive.today/#social

Looking at archive site traffic today - Archive.is



https://www.similarweb.com/website/archive.is/#social

other flavors include archive.vn, archive.md, etc

The Dataset

Platform	Archive	#Posts with Archive URLs (%all posts)	Archive URLs	Source URLs	Source Domains	Filtered
Live Feed	archive.is		21,537,554	20,608,834	5,388,112	-
Reddit	archive.is	327,050 ($2.9 \cdot 10^{-4}$%)	310,392	291,382	15,994	35.70%
	Wayback	320,379 ($2.8 \cdot 10^{-4}$%)	387,081	343,851	21,124	17.20%
/pol/	archive.is	46,912 ($1.1 \cdot 10^{-3}$%)	36,277	33,824	3,970	4.67%
	Wayback	3,848 ($9.7 \cdot 10^{-5}$%)	2,325	2,207	976	83.12%
Gab	archive.is	6,602 ($3.4 \cdot 10^{-4}$%)	5,943	5,773	1,300	5.54%
	Wayback	478 ($5.1 \cdot 10^{-5}$%)	361	349	240	61.18%
Twitter	archive.is	6,750 ($3.1 \cdot 10^{-6}$%)	3,772	3,669	845	8.23%
	Wayback	1,905 ($9.0 \cdot 10^{-7}$%)	1,290	1,257	846	7.49%

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. We also filter URLs that are malformed, unreachable, or point to resources other than Web pages.

The Sample

- 21.5M Archive URLs (URI-M)
- 20.6M Unique Source URLs (URI-R)
- 5.3M Unique Domains

Cross Platform Analysis

Evaluating Archive.is Live Feed URLs

Domain	(%)	Sx	(%)	Domain	(%)	Sx	(%)
archive.org	11.82%	.com	38.29%	ru-board.com	0.50%	.pl	1.24%
twitter.com	5.73%	.org	17.64%	asstr.org	0.49%	.ch	1.23%
quora.com	3.18%	.de	7.02%	ruliweb.com	0.43%	.eu	1.01%
livejournal.com	2.17%	.jp	5.61%	4chan.org	0.40%	.se	0.80%
reddit.com	1.81%	.net	3.19%	googleusercontent.com	0.40%	.cz	0.69%
facebook.com	1.31%	.ru	3.10%	ameblo.jp	0.39%	.br	0.66%
nhk.or.jp	0.78%	.nl	2.56%	wordpress.com	0.38%	.at	0.63%
youtube.com	0.65%	.uk	1.51%	yahoo.co.jp	0.38%	.es	0.57%
wikipedia.org	0.52%	.it	1.39%	aaaaarg.fail	0.37%	.be	0.55%
tumblr.com	0.51%	.fr	1.39%	blogspot.nl	0.36%	.ca	0.51%

Table 2: Top 20 domains and suffixes of the source URLs in the archive.is live feed dataset.

Which archive URLs are Social Media sites using?

- For archive.is, the top domain for each platforms point to their own platform mementos (e.g., archives of reddit posts are the most shared ones on Reddit)
- For Wayback Machine URLs, only Reddit has its own platform as the top domain (flame wars and intra-Reddit conflict are often captured before takedowns)

Archival Fraction

Number of times a source domain appears in an archive

Total number of times it appears in the dataset (either archived or not).

Domain (archive.is)	(%)	AF	Domain (Wayback)	(%)	AF
reddit.com	31.21%	<0.01	reddit.com	36.88%	<0.01
pastebin.com	6.80%	0.08	imgur.com	7.05%	<0.01
twitter.com	5.89%	<0.01	twitter.com	5.19%	<0.01
imgur.com	3.02%	<0.01	redd.it	4.79%	<0.01
washingtonpost.com	2.46%	0.02	youtube.com	3.90%	<0.01
youtube.com	2.33%	<0.01	washingtonpost.com	1.54%	0.01
redd.it	2.14%	<0.01	youtu.be	1.19%	<0.01
nytimes.com	1.76%	0.01	nytimes.com	0.98%	<0.01
cnn.com	1.64%	0.02	cnn.com	0.90%	<0.01
wikipedia.org	1.37%	<0.01	reddituploads.com	0.89%	0.06
huffingtonpost.com	0.93%	0.02	archive.is	0.61%	<0.01
theguardian.com	0.78%	<0.01	streamable.com	0.61%	<0.01
googleusercontent.com	0.65%	0.08	thehill.com	0.54%	0.01
politico.com	0.64%	0.02	wikipedia.org	0.52%	<0.01
wsj.com	0.61%	0.03	politico.com	0.49%	0.02
dailymail.co.uk	0.54%	0.01	theguardian.com	0.46%	<0.01
4chan.org	0.53%	0.16	rawstory.com	0.45%	0.06
facebook.com	0.52%	<0.01	huffingtonpost.com	0.44%	<0.01
thehill.com	0.43%	0.01	bbc.com	0.44%	0.01
breitbart.com	0.40%	0.01	kickstarter.com	0.37%	0.02

Table 3: Top 20 source domains of archive.is and Wayback Machine URLs, and archival fraction (AF), in the Reddit dataset.

Domain (archive.is)	(%)	AF	Domain (Wayback)	(%)	AF
twitter.com	25.02 %	<0.01	justpaste.it	11.90 %	0.02
facebook.com	3.65 %	<0.01	twitter.com	6.90 %	0.01
togetter.com	3.58 %	<0.01	dailymail.co.uk	1.95 %	0.13
seesaa.net	2.97 %	**0.91**	nikkansports.com	1.50 %	0.18
justpaste.it	2.19 %	0.01	mikelofgren.net	1.20%	**1.00**
yahoo.co.jp	2.03 %	0.21	blogspot.com	1.10%	0.09
googleusercontent.com	1.77 %	**0.98**	whitehouse.gov	1.05%	0.02
time.com	1.75 %	0.01	journalists-in-russia.org	1.00%	**1.00**
monjiro.net	1.66 %	0.51	pcdepot.co.jp	0.90%	**0.90**
pastebin.com	1.45 %	0.04	rydon.co.uk	0.85%	**1.00**
google.com	1.39 %	0.01	yeniakit.com.tr	0.85%	0.16
jimin.jp	1.35 %	**0.95**	cdse.edu	0.75%	**0.93**
notepad.cc	1.33 %	0.47	tetsureki.com	0.75%	**1.00**
ameblo.jp	1.16 %	<0.01	donaldjtrump.com	0.75%	0.04
nhk.or.jp	1.16 %	0.33	reidreport.com	0.75%	**1.00**
magi.md	1.16 %	0.49	ameblo.cjp	0.70%	<0.01
opensecrets.org	1.05 %	0.67	jreast.co.jp	0.70%	**0.93**
fc2.com	0.99 %	0.27	eastandard.net	0.65%	**1.00**
dailyshincho.jp	0.93 %	**0.94**	yahoo.co.jp	0.60%	0.01
reddit.com	0.89 %	0.03	livedoor.jp	0.60%	0.07

Table 5: Top 20 source domains of archive.is and Wayback Machine URLs, and archival fraction (AF), in the Twitter dataset.

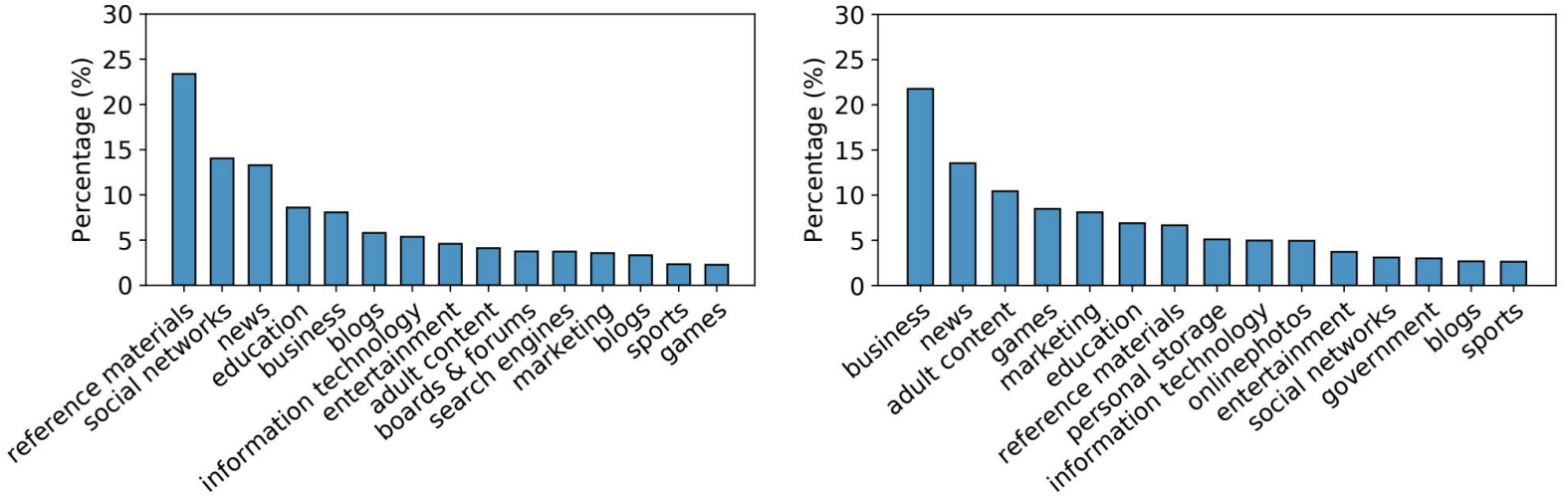
Domain (archive.is)	(%)	AF	Domain (Wayback)	(%)	AF
4chan.org	9.35%	0.54	justice4germans.com	7.50%	**0.94**
theguardian.com	3.78%	0.13	chetlyzarko.com	3.90%	**1.00**
washingtonpost.com	3.70%	0.20	twitter.com	2.82%	<0.01
nytimes.com	3.46%	0.16	dailymail.co.uk	2.47%	<0.01
cnn.com	2.78%	0.14	revcom.us	2.16%	0.66
twitter.com	2.75%	0.01	reddit.com	1.98%	<0.01
independent.co.uk	2.37%	0.13	tumblr.com	1.85%	0.02
breitbart.com	1.96%	0.08	thebilzerianreport.com	1.57%	0.72
reddit.com	1.85%	0.09	jeffreyepsteinscience.com	1.55%	**1.00**
dailymail.co.uk	1.72%	0.05	cnn.com	1.51%	<0.01
facebook.com	1.69%	**0.96**	tdbimg.com	1.43%	**1.00**
huffingtonpost.com	1.37%	0.20	huffingtonpost.com	1.43%	0.01
thehill.com	1.21%	0.16	metapedia.org	1.22%	0.04
politico.com	1.04%	0.13	nytimes.com	1.15%	<0.01
bbc.com	1.01%	0.08	washingtonpost.com	1.11%	<0.01
8ch.net	0.98%	**1.00**	theguardian.com	1.08%	<0.01
googleusercontent.com	0.91%	0.59	independent.co.uk	1.08%	<0.01
hypothes.is	0.87%	**0.98**	wordpress.com	1.06%	<0.01
telegraph.co.uk	0.85%	0.03	idrsolutions.com	1.01%	0.86
theatlantic.com	0.81%	0.24	wikileaks.com	1.01%	<0.01

Table 4: Top 20 source domains of archive.is and Wayback Machine URLs, and archival fraction (AF), in the /pol/ dataset.

Domain (archive.is)	(%)	AF	Domain (Wayback)	(%)	AF
twitter.com	12.28%	<0.01	dailymail.co.uk	20.98%	< 0.01
nytimes.com	4.71%	0.03	washingtonpost.com	7.08%	0.01
washingtonpost.com	4.17%	0.03	infowars.com	5.54%	<0.01
reddit.com	3.10%	0.03	brandenburg.de	4.35%	0.10
googleusercontent.com	2.43%	0.18	twitter.com	3.63%	< 0.01
breitbart.com	1.82%	< 0.01	huffingtonpost.com	3.08%	<0.01
cnn.com	1.63%	0.01	abcnews.go.com	2.54%	< 0.01
4chan.org	1.59%	0.07	salon.com	1.72%	0.01
dailymail.co.uk	1.44%	<0.01	alexa.com	1.63%	0.03
theguardian.com	1.29%	< 0.01	news.com.au	1.54%	<0.01
wsj.com	1.22%	0.01	tu-dortmunt.de	1.45%	0.80
bbc.com	1.15%	0.01	causes.com	1.27%	0.50
huffingtonpost.com	1.14%	0.03	vigilantcitizen.com	1.18%	0.02
google.com	1.01%	< 0.01	reddit.com	1.08%	<0.01
facebook.com	0.92%	< 0.01	sahra-wagenknecht.de	0.99%	0.78
latimes.com	0.85%	0.01	quillette.com	0.99%	0.02
yahoo.com	0.81%	< 0.01	derwesten.de	0.99%	<0.01
dailycaller.com	0.77%	< 0.01	politico.com	0.91%	<0.01
thehill.com	0.74%	< 0.01	mikelofgren.net	0.81%	**0.90**
wikileaks.org	0.73%	0.01	alexanderhiggins.com	0.81%	0.02

Table 6: Top 20 source domains of archive.is and Wayback Machine URLs, and archival fraction (AF), in the Gab dataset.

Archive.is live feed domain categories



(a) Top 100K Domains (b) Sample of 121K Domains

Figure 2: Top 15 domain categories for the archive.is live feed.

Social Media Archive URL Domain Categories

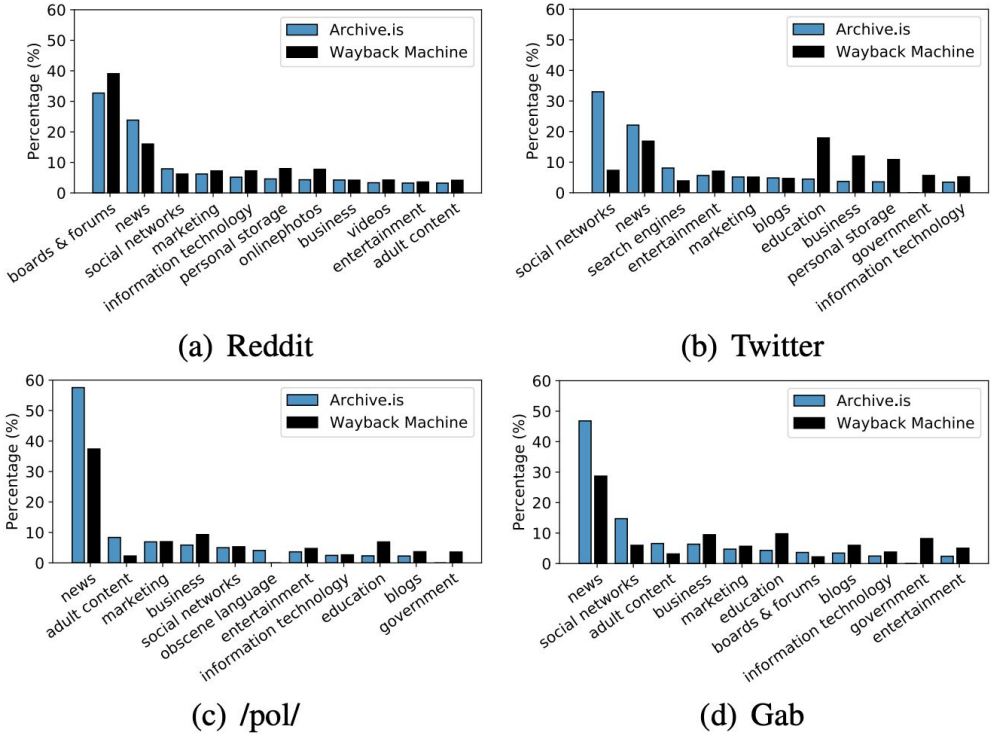
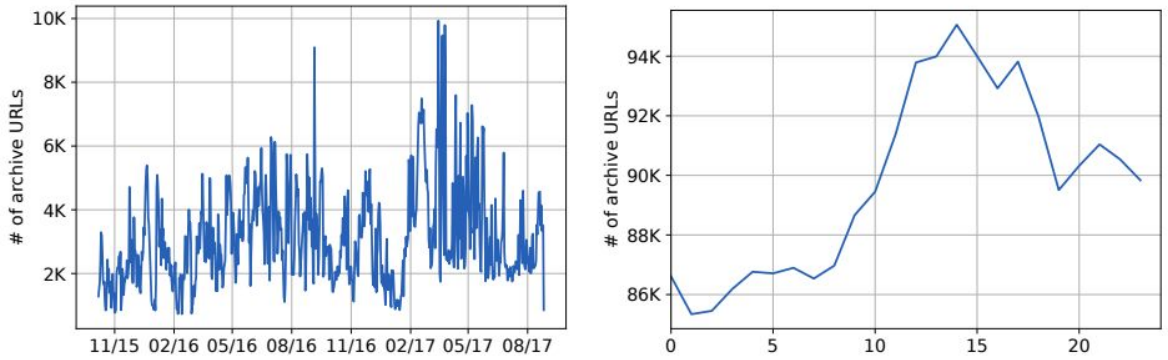


Figure 3: Top domain categories for archive URLs appearing on the four social networks.

Observations from Domain categorizations

- Education and Government URLs appear as top categories for the Wayback Machine
- Sites that contain obscene language appear only for archive.is
- Adult Content is among the top categories for all social networks except Twitter

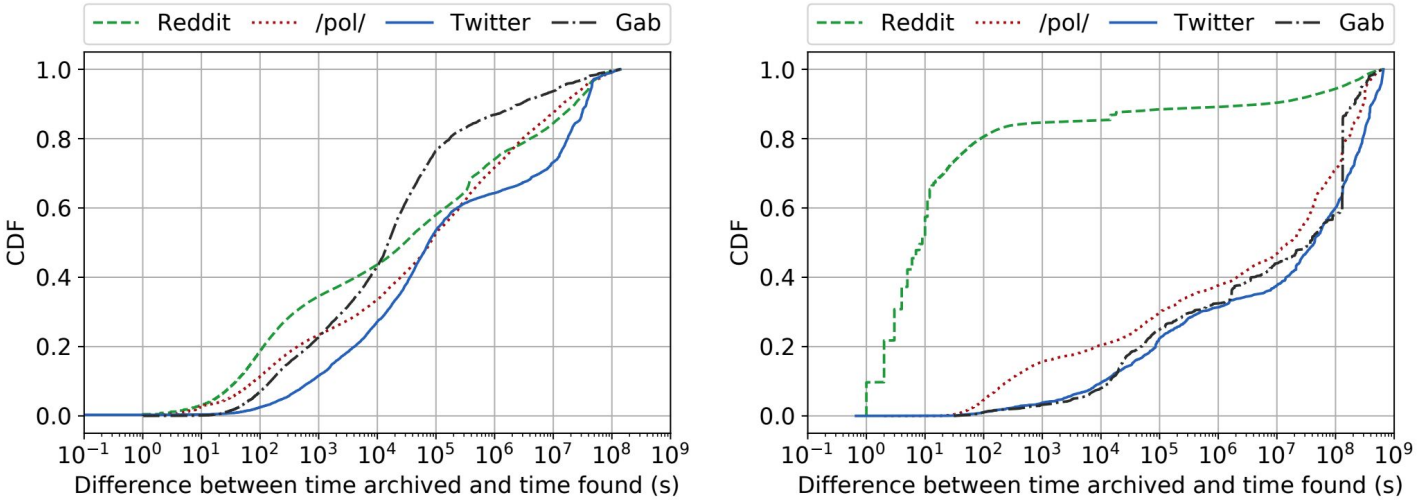
Archive.is Archival Rate



(a) Date (b) Hour of Day

Figure 4: Temporal analysis of the archive.is live feed dataset, reporting the number of URLs that are archived (a) each day and (b) based on hour of day.

Time difference:
Time found vs. Time it was Archived



(a) archive.is (b) Wayback Machine

Figure 5: CDF of the time difference between the archival time and the time appeared on each of the four social networks.

Live Web URIs no longer available (404/410/451/5xx)

- Archive.is lived feed: 12% gone
- Social Network URLs referencing web archives:

Social Network	Archive.is % alive	Wayback % alive
Reddit	93%	89%
Gab	87%	48%
Twitter	76%	49%
4chan	82%	66%

Social-Network-based Analysis

Reddit Archiving User Base

- 31% of all archive.is URLs and 82% of Wayback Machine URLs are posted by a specific bot, namely, **SnapshillBot**
- Other bots:
  - AutoModerator,
  - 2016VoteBot,
  - yankbot, and
  - autotldr

Reddit Sub Communities

Subreddit (archive.is)	(%)	Subreddit (Wayback)	(%)
The_Donald	24.48%	EnoughTrumpSpam	31.82%
KotakuInAction	15.83%	MGTOW	7.38%
EnoughTrumpSpam	12.06%	SnapshillBotEx	7.19%
MGTOW	3.48%	undelete	5.90%
undelete	2.74%	SubredditDrama	5.50%
SubredditDrama	2.61%	Drama	5.03%
Drama	2.33%	Gamingcirclejerk	3.47%
Gamingcirclejerk	1.57%	ShitAmericansSay	1.63%
conspiracy	1.44%	TopMindsOfReddit	1.51%
MensRights	1.12%	TheBluePill	1.25%
savedyouaclick	1.00%	Buttcoin_1000	1.15%
politics	0.98%	AgainstHateSubreddits	1.06%
DerekSmart	0.76%	subredditcancer	0.99%
ShitAmericansSay	0.75%	The_Donald	0.95%
PoliticsAll	0.72%	badeconomics	0.75%

Table 7: Top 15 subreddits sharing archive.is and Wayback Machine URLs.

Gab, 4chan, and Twitter User Bases

- Gab and Twitter both rely on specific users to point out archived content
- 4Chan is anonymous users and only 200 posts per community. USA majority of provided URLs.

URL Blocking Robots

- At least 1 reddit robot, AutoModerator, is used to remove links and point to mementos (URI-M).
- Robots mainly appearing in politic oriented subreddits
- Mainstream news outlets (e.g., CNN, Washington Post) are the top domains removed from **The_Donald** subreddit

Estimating Ad Revenue Missed

News Source	Count	(%)	News Source	Count	(%)
washingtonpost.com	3,814	44.13%	change.org	96	7.52%
cnn.com	3,354	39.39%	huffpost.com	62	13.39%
nydailynews.com	1,070	46.32%	fusion.net	60	44.77%
huffingtonpost.com	978	43.77%	cnn.it	58	44.61%
nationalreview.com	774	45.58%	alternet.org	26	20.01%
theblaze.com	704	46.74%	infostormer.com	16	27.11%
buzzfeed.com	588	45.97%	dailynewsbin.com	4	26.67%
salon.com	373	44.88%	todayvibes.com	4	7.27%
vice.com	372	45.14%	usanewsbets.ga	4	10.52%
vox.com	323	45.23%	fullycucked.com	1	1.78%
weeklystandard.com	253	46.25%	northcrane.com	1	0.13%
politifact.com	185	33.09%			

Table 8: Number and percentage of submissions deleted from The_Donald with links to different news sources.

Estimated Monthly Ad Revenue Missed

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

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.

Findings

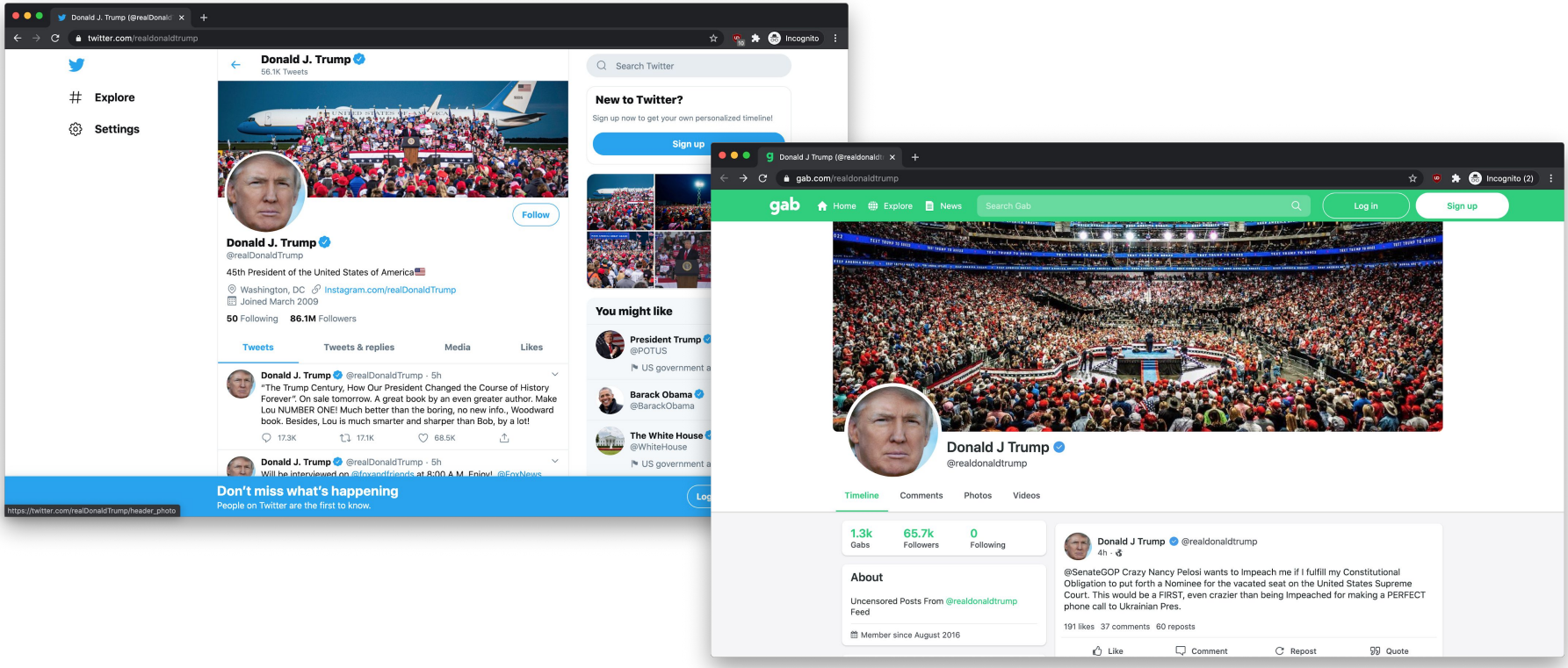
- News and social media posts are the most common types of content archived, likely due to their (perceived) ephemeral and/or controversial nature.
- URLs of archiving services are extensively shared on “fringe” communities within Reddit and 4chan to preserve possibly contentious content
- Web archives are exploited by users to bypass censorship policies in some communities (e.g., sharing archive.is URLs banned from Facebook)
- Reddit bots are responsible for posting a very large portion of archive URLs in the subreddits they study
- 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

Extra Slides

Extra references:

- The paper itself: https://arxiv.org/abs/1801.10396
- Archive.is live feed: http://archive.is/livefeed/?img=0&onlyfailures=0&submitter=any&grid=0&urlinterval=*&time=now&pagesize=600&page=1
- The WSDL review: https://ws-dl.blogspot.com/2018/04/2018-04-12-web-archives-are-used-for.html

Twitter vs. Gab



Things Grant didn't Enjoy

Web archives. [2] 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. [3] 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. [1] 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, [12] assess whether the Wayback Machine archives a purely random sample of Web pages, finding some bias towards more visible and prominent pages.

Who?

Memento wasn't referenced, but they have a whole section related to temporal captures of URIs.