The Web Centipede: Understanding How Web Communities Influence Each Other Through the Lens of Mainstream and Alternative News Sources

Savvas Zannettou[⋆], Tristan Caulfield[‡], Emiliano De Cristofaro[‡], Nicolas Kourtellis[†],
Ilias Leontiadis[†], Michael Sirivianos[⋆], Gianluca Stringhini[‡], and Jeremy Blackburn[†]
[⋆]Cyprus University of Technology, [‡]University College London, [†]Telefonica Research
sa.zannettou@edu.cut.ac.cy {t.caulfield,e.decristofaro,g.stringhini}@ucl.ac.uk
{nicolas.kourtellis,ilias.leontiadis,jeremy.blackburn}@telefonica.com michael.sirivianos@cut.ac.cy

ABSTRACT

As the number and diversity of news sources on the Web grows, so does the opportunity for alternative sources of information production. The emergence of mainstream social networks like Twitter and Facebook makes it easier for misleading, false, and agenda driven information to quickly and seamlessly spread online, deceiving people or influencing their opinions. Moreover, the increased engagement of tightly knit communities, such as Reddit and 4chan, compounds the problem as their users initiate and propagate alternative information not only within their own communities, but also to other communities and social media platforms across the Web. These platforms thus constitute an important piece of the modern information ecosystem which, alas, has not been studied as a whole.

In this paper, we begin to fill this gap by studying mainstream and alternative news shared on Twitter, Reddit, and 4chan. By analyzing millions of posts around a variety of axes, we measure how mainstream and alternative news flow between these platforms. Our results indicate that alt-right communities within 4chan and Reddit can have a surprising level of influence on Twitter, providing evidence that "fringe" communities may often be succeeding in spreading these alternative news sources to mainstream social networks and the greater Web.

1. INTRODUCTION

After the Boston Marathon bombings in 2013, a large number of tweets started to claim that the bombings were a "false flag" perpetrated by the United States government [24]. The #GamerGate controversy started as a blogpost by a jaded ex-boyfriend that was twisted and turned into a pseudo-political campaign of targeted online harassment [6]. More recently, the PizzaGate conspiracy, a debunked theory connecting some restaurants and members of the US Democratic Party with a child-sex ring, even led to a shooting in a North Carolina restaurant [27]. What these examples have in common is that they were propagated in no small part via the use of "alternative" news sites like Infowars and "fringe" Web communities like 4chan.

Overall, the Web and online social networks have greatly reduced the barrier of entry for such alternative news sources. Due to the negligible cost of distributing information over social media, fringe sites can quickly gain traction with large audiences. At the same time, the explosion of information sources also hinders the effective regulation of the sector, while further muddying the water when it comes to the evaluation of news information by readers.

While there are many plausible motives for the rise in alternative narratives [23], ranging from libelous (e.g., to harm the image of a particular person or group), political (e.g,. to influence voters), profit (e.g., to make money from advertising), or just trolling [1], the manner in which they proliferate throughout the Web is still unknown. Although previous work has examined information cascades, rumors, and hoaxes [10, 13, 21], to the best of our knowledge, very little work provides a holistic view of the modern information ecosystem. This knowledge is crucial to understand the risks associated with alternative news and to design appropriate detection and mitigation strategies.

Anecdotal evidence and press coverage suggest that alternative news dissemination might start on "fringe" websites, eventually reaching mainstream online social networks and news outlets.[1] Nevertheless, to the best of our knowledge, this phenomenon has not been rigorously studied and no thorough analysis has focused on how news moves from one online service to another like an interconnected centipede. In this paper, we address this gap by providing the first large-scale measurement of how mainstream and alternative news flows through multiple social media platforms. More specifically, we focus on the relationship between three fundamentally different social media platforms, Reddit, Twitter, and 4chan, chosen because of their fundamental differences as well as their generally accepted "driving" of substantial portions of the online world.

Contributions. This paper makes several contributions. First, we undertake a large-scale measurement and compar-

[1] http://bbc.in/2pQP5KH

ison of the occurrence of mainstream and alternative news sources across three social media platforms (4chan, Reddit, and Twitter). Next, we provide an understanding of the temporal dynamics of how URLs from news sites are posted on the different social networks. Finally, we present a measurement of the *influence* between the platforms that provides insight into how information spreads throughout the greater Web.

Overall, our findings indicate that Twitter, Reddit, and 4chan are used quite extensively for the dissemination of both alternative and mainstream news. We also find relatively heavy use of time capsule services, such as archive.is, by users of Reddit and 4chan but not those of Twitter. By utilizing a statistical model for influence, Hawkes processes, we show that each of the platforms (and in the case of Reddit, sub-communities) have varying degrees of influence on each other, and this influence differs with respect to mainstream and alternative news sources.

Paper Organization. The rest of the paper is organized as follows. The next section reviews related work, then, in Section 3, we discuss the social networks and the information sources studied in this paper. Section 4 presents a general characterization of each platform, while Section 5 discusses our temporal findings. Section 6 reports our measurements of the influence between the platforms, while the paper concludes in Section 7.

2. RELATED WORK

In this section, we review prior work on disinformation propagation dynamics in social networks as well as on detecting false information sources.

Disinformation dynamics. Kwon et al. [14] study the characteristics of rumor propagation on Twitter. They analyze a corpus of 1.7B tweets, covering three and a half years of Twitter data, and extract 104 viral events, which are then annotated by human coders. They study debunked stories from false information busting sites such as snopes.com and compare temporal, linguistics, and structural characteristics to legitimate information. Friggeri et al. [10] also use stories that Snopes determined as false to study the propagation and the evolution of false information on Facebook, finding it to be quite bursty, and that posts containing a comment with a link to snopes.com were more likely to be deleted. Kumar et al. [13] study the presence of hoaxes in Wikipedia articles. They report that while most are detected quickly and have little impact, some end up cited widely on the Web.

Shao et al. [21] introduce Hoaxy, a platform providing information about the dynamics of false information propagation on Twitter. They also study a sample of 1.4M tweets, finding that the diffusion of fact-checking content lags that of false information by 10-20 hours and that the top 1% users with the most tweets share a much higher ratio of false information. Finn et al. [9] present TwitterTrails, a website allowing users to study propagation of false information on Twitter, i.e., to visualize indications of bursty activity, community skepticism, temporal characteristics of propagation, as well as re-tweets networks. Also, Del Vicario et al. [8] analyze how Facebook users perceive and react to conspiracy theories vs. scientific stories, finding two polarized and homogeneous communities that have similar content consumption patterns but exhibit different cascade dynamics.

Situngkir [22] empirically studies an Indonesian hoax on Twitter, finding that it spread broadly and quickly (within two hours), and that it would have spread more if a conventional media outlet did not publicly deny it. He also argues that hoaxes can propagate easily if there is collaboration between the recipients of the hoax. Arif et al. [3] also present a case study based on a hostage crisis in Sydney, analyzing 5.4M tweets from three main perspectives: (i) volume (i.e., number of rumor-related messages per time interval), (ii) exposure (i.e., number of individuals exposed to the rumor), and (iii) content production (i.e., whether the content is written by the user or is re-shared). Andrews et al. [2] study two crisis-related incidents on Twitter aiming to determine the effect of “official” accounts with respect to the containment of rumors. Authors show that official account can significantly contribute to stopping the propagation of the rumor by actively engaging in conversations related to the incidents. Finally, Mendoza et al. [17] study the dissemination of false rumors vs. confirmed news on Twitter the days following the 2010 earthquake in Chile, concluding that an aggregate analysis on the flow of tweets can effectively distinguish the former from the latter.

Detecting false information sources. Shah et al. [20] formulate the problem of finding the source of false information as a maximum likelihood estimation problem, using a metric called rumor centrality. They evaluate it for all nodes in the network using a simple linear time message-passing algorithm, and the node with the highest rumor centrality is deemed to be the most likely source. Authors show, experimentally, that the model can distinguish the source of false information with a maximum error of 7 hops for general networks and 4 for tree networks. Wang et al. [26] study the problem from a statistical point of view, proposing a source detection framework, also based on rumor centrality, which supports multiple snapshots of the network during the false information spread. They show that using two network snapshots instead of one can significantly improve detection.

Budak et al. [5] study the notion of competing campaigns in a social network and address the problem of influence limitation to counteract the effect of misinformation. Nguyen et al. [18] look for the k users that are most suspected to have originated false information, using a reverse diffusion process along with a ranking process. Seo et al. [19] aim to identify the source of rumors in online social networks by injecting monitoring nodes across the social graph. They propose an algorithm that observes the information received by the monitoring nodes in order to identify the source. They indicate that with sufficient number of monitoring sources they can recognize the source with high accuracy.

Finally, Starbird [23] performs a qualitative analysis on tweets pertaining to shooting events and conspiracy theories, using graph analysis on the domains linked from the tweets, and provides insight on how various websites work to promote conspiracy theories and push political agendas.

Remarks. In contrast to prior work, this paper provides insights on disinformation dynamics on social networks from a comprehensive (i.e., multi-service) point of view. In other words, we study major OSN and various news sites that actively contribute to information diffusion across the Web, specifically, analyzing the dynamics and information flow of Reddit, 4chan, Twitter, and several news sites.

3. DATASET

In this section, we provide some background information on the three social media platforms we study, the selection of news sources, and details on the collected data.

3.1 Platforms and News Sources

Twitter. Twitter is a micro-blogging, directed social network where users can easily broadcast 140-character 'tweets' to their followers. Some of its features include the hashtag (basically, a keyword preceded by #), which makes it easier for users to find and weigh in on tweets around a theme, as well as retweeting, i.e., rebroadcasting a tweet.

Reddit. The so-called "front page of the Internet" is a social news aggregator, where users post URLs to content along with a title, and other users can upvote or downvote the post. Votes determine the ranking of the posts, i.e., the order in which they are displayed on the site. There is also a threaded comments section for users to discuss a post, and comments are also subject to the voting system. Although users can mark each other as friends, the community structure is not defined by the friendship relation. Rather, communities on Reddit are formed via the "subreddit" concept. Users can create their own subreddits, choosing the topic as well as the moderation policy. This has led to a plethora of communities, ranging from video games to news and politics, pornography, and even meta-communities focusing on interactions people have in other subreddits.

4chan. 4chan is a type of discussion forum known as an *imageboard*: users create a new thread by making a post with a single image attached, and perhaps some text, in one of several boards (69 as of May 2017) for different topics of interest. Other users can add posts to the thread, with or without an image, and quote or reply to posts. Users are not required to provide a username to access or post to 4chan, and the default "Anonymous" is the preferred and overwhelmingly used identity. Another key characteristic of 4chan is ephemerality: there is only a finite number of threads that can be active at a given time on a board. When a new thread is created, an old one is purged based on their ranking within the "bump" system [12]. Although several boards have a temporary archive for purged posts, all threads are permanently deleted after 7 days. Finally, 4chan is known for its extremely lax moderation: although boards are divided into safe and not safe for work categories, volunteer "janitors" and paid employees generally are not concerned with the language used or the tone of the discussions, as long as the discussion falls within the general topic of the board. Since 4chan's primary mode of operation is "anonymous", it inherently lacks many of the "social" features of other social media platforms, and there is no concept of friends/followers.

Platform	Total Posts	% Alt.	% Main.
Twitter	587 Million	0.022%	0.070%
Reddit (posts + comments)	332 Million	0.023%	0.181%
4chan	42 Million	0.050%	0.197%

Table 1: Total number of posts crawled and percentage of posts that contain URLs to our list of alternative and mainstream sites.

Platform	Posts/Comments	Alt. URLs	Main. URLs
Twitter	486,700	42,550	236,480
Reddit (6 selected subreddits)	620,530	40,046	301,840
Reddit (all other subreddits)	1,228,105	24,027	726,948
4chan (/pol/)	90,537	8,963	40,164
4chan (/int/, /sci/, /sp/)	7,131	615	5,513

Table 2: Overview of our datasets with the number of posts/comments that contain a URL to one of our information sources, as well as the number of unique URLs linking to alternative and mainstream news sites in our list.

While we use 4chan's sports (/sp/), international (/int/), and science (/sci/) boards as a baseline, we are primarily interested in the Politically Incorrect board, or /pol/. /pol/ focuses on the discussion of politics and world events, and has been often linked to the alt-right [4] as well as exhibiting a high degree of racist and hate speech content [12].

News sites. Our analysis uses a set of news sites that can confidently be labeled as either "mainstream" or "alternative" news. More specifically, we create a list of 99 news sites including 45 mainstream and 54 alternative ones. For the former, we select 45 among the Alexa top 100 news sites, leaving out those based on user-generated content, those serving specialized content (e.g., finance news), as well as non-English sites. For the latter, we use Wikipedia[2] and FakeNewsWatch.[3] We also add two state-sponsored alternative news domains: sputniknews.com and rt.com, as they have recently attracted public attention due to their posting of controversial, and seemingly agenda-pushing stories [7].[4]

3.2 Datasets

We gather information from posts, threads, and comments on Twitter, Reddit, and 4chan that contain URLs from the 99 news sites. With a few gaps (described below), our datasets cover activity on the three platforms we measure between June 30, 2016 and February 28, 2017. Table 1 shows the total number of posts/comments crawled and the percentage of the post that contains links to the URLs from the aforementioned news domains. When compared to Twitter, users of

[2] https://en.wikipedia.org/wiki/List_of_fake_news_websites

[3] http://fakenewswatch.com/

[4] The complete list of the 99 sites is available at https://drive.google.com/open?id=0ByP5a__khV0dM1ZSY3YxQWF2N2c.

Subreddit (Alt.)	(%)	Subreddit (Main.)	(%)
The_Donald	35.37 %	politics	12.9 %
politics	8.21 %	worldnews	6.24 %
news	3.85 %	The_Donald	4.53 %
conspiracy	3.84 %	news	4.23 %
Uncensored	2.66 %	TheColorIsBlue	3.06 %
Health	2.10 %	TheColorIsRed	2.48 %
PoliticsAll	1.54 %	willis7737_news	2.27 %
Conservative	1.45 %	news_etc	1.94 %
worldnews	1.41 %	AskReddit	1.37 %
WhiteRights	1.21 %	canada	1.31 %
KotakuInAction	1.04 %	EnoughTrumpSpam	1.20 %
HillaryForPrison	0.94 %	NoFilterNews	1.16 %
TheOnion	0.94 %	BreakingNews24hr	1.07 %
AskTrumpSupporters	0.84 %	conspiracy	0.89 %
POLITIC	0.81 %	todayilearned	0.83 %
rss_theonion	0.67 %	thenewsrightnow	0.78 %
the_Europe	0.67 %	europe	0.77 %
new_right	0.6 %	ReddLineNews	0.75 %
AskReddit	0.59 %	hillaryclinton	0.73 %
AnythingGoesNews	0.51 %	nottheonion	0.73 %

Table 3: Top 20 subreddits w.r.t. mainstream and alternative URLs occurrence and their percentage in Reddit (all subreddits).

Domain (Alt.)	(%)	Domain (Main.)	(%)
breitbart.com	55.58 %	nytimes.com	14.07 %
rt.com	19.18 %	cnn.com	11.23 %
infowars.com	8.99 %	theguardian.com	8.86 %
sputniknews.com	3.95 %	reuters.com	6.67 %
beforeitsnews.com	2.34 %	huffingtonpost.com	5.67 %
lifezette.com	2.28 %	thehill.com	5.15 %
naturalnews.com	1.54 %	foxnews.com	4.89 %
activistpost.com	1.45 %	bbc.com	4.76 %
veteranstoday.com	1.11 %	abcnews.go.com	2.94 %
redflagnews.com	0.63 %	usatoday.com	2.87 %
prntly.com	0.49 %	nbcnews.com	2.86 %
dcclothesline.com	0.4 %	time.com	2.57 %
worldnewsdailyreport.com	0.36 %	washinghtontimes.com	2.52 %
therealstrategy.com	0.3 %	bloomberg.com	2.5 %
disclose.tv	0.23 %	wsj.com	2.31 %
clickhole.com	0.2 %	cbsnews.com	2.26 %
libertywritersnews.com	0.2 %	thedailybeast.com	2.05 %
worldtruth.tv	0.14 %	forbes.com	1.87 %
thelastlineofdefence.org	0.07 %	nypost.com	1.85 %
nodisinfo.com	0.05 %	cncb.com	1.54 %

Table 4: Top 20 mainstream and alternative domains and their percentage in the 6 selected subreddits.

4chan and Reddit are sharing frequently links to mainstream news whereas 4chan users are more likely to post links to alternative news sites. Table 2 provides a summary of our datasets, which we present in more detail below. Note that we break Reddit and 4chan datasets into two different instances, as further discussed in Section 4.

Twitter. We collect the 1% of all publicly available tweets with URLs from the aforementioned news domains between June 2016 and February 2017 using the Twitter Streaming API.[5] In total, we gather 487k tweets containing 279k unique URLs. Since tweets are retrieved at the time they are posted, we do not get information such as the number of times they are re-tweeted or liked. Therefore, between March and May 2017, we re-crawl each tweet to retrieve this data. Due to a failure in our collection infrastructure, we have some gaps in the Twitter dataset, specifically between Oct 28–Nov 2 and Nov 5–16, 2016, as well as Nov 22, 2016 – Jan 13, 2017, and Feb 24–28, 2017.

Reddit. We obtain all posts and comments on Reddit between June 2016 and February 2017, using data made available on Reddit itself.[6] We collect approximately 42M posts, 390M comments, and 300k subreddits. Once again, we filter posts and comments that contain URLs from one of the 99 news sites, which yields a dataset of 1.8M posts/comments and approximately 1.1M URLs.

4chan. For 4chan, we use all threads and posts made on the Politically Incorrect (/pol/) board, as well as /sp/ (sports), /int/ (international), and /sci/ (science) boards for comparison, using the same methodology as [12]. The resulting dataset includes 97k posts and replies, including 56k alternative and mainstream URLs, between of June 30, 2016 and February 28, 2017. We have some small gaps due to our crawler failing, specifically, Oct 15–16 and Dec 16–25, 2016 as well as Jan 10–13, 2017.

4. GENERAL CHARACTERIZATION

In this section, we present a general characterization of the mainstream and alternative news URLs found on each of the three platforms. We also shed light on time capsule services and how they contribute to the dissemination of news.

4.1 Platform Analysis

Reddit. We start by identifying communities around news and politics. In Table 3, we report the top 20 subreddits with the most URLs, along with their percentage. Note that we omit automated ones (e.g., /r/AutoNewspaper/) where news articles are posted without user intervention. Many of the subreddits are indeed related to news and politics – e.g., 'The_Donald' is mostly a community of Donald Trump supporters, while 'worldnews' is focused around globally relevant events. We also find the presence of the 'conspiracy' subreddit, which has been involved in disinformation campaigns including Pizzagate as well as 'AskReddit,' where both mainstream and alternative news sources are used to answer questions submitted by users. Although the latter is intended for open-ended questions that spark discussion, we can anecdotally confirm that commenters often try to push their agenda even on non-political threads. In the end, based on their propensity to include news URLs, we single out six subreddits for further exploration: The_Donald, conspiracy, AskReddit, politics, worldnews, and news.

In order to get a better view of the popularity of news sites on Reddit, we study the occurrence of each news outlet. Specifically, we find 76k URLs (40k unique) from alternative news and 600k (301k unique) from mainstream news domains. Table 4 reports the top 20 mainstream/alternative news sites and their percentage in the six subreddits. The top 20 domains for mainstream news account for 89% of all mainstream URLs in our data, while for alternative domains the percentage is 99%. Known alt-right news outlets, such as breitbart.com and infowars.com, are predominantly

[5] https://dev.twitter.com/streaming/overview

[6] https://www.reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment/

	Tweets	Retrieved (%)	Avg. Retweets	Avg. Likes
Alternative	110,629	92,104 (83.2%)	341 ± 1,228	0.82 ± 15.6
Mainstream	376,071	329,950 (87.7%)	404 ± 2,146	0.96 ± 55.6

Table 5: Basic statistics of the occurrence of alternative and mainstream URLs in the tweets in our dataset.

Domain (Alt.)	(%)	Domain (Main.)	(%)
breitbart.com	46.04 %	theguardian.com	19.04 %
rt.com	17.56 %	nytimes.com	10.07 %
infowars.com	17.25 %	bbc.com	8.99 %
therealstrategy.com	5.63 %	forbes.com	6.24 %
sputniknews.com	4.11 %	thehill.com	4.95 %
beforeitsnews.com	2.26 %	cbc.ca	4.82 %
redflagnews.com	2.04 %	foxnews.com	4.79 %
dcclothesline.com	1.37 %	wsj.com	4.04 %
naturalnews.com	1.29 %	bloomberg.com	3.48 %
clickhole.com	0.53 %	reuters.com	2.85 %
activistpost.com	0.41 %	usatoday.com	2.02 %
disclose.tv	0.39 %	thedailybeast.com	2.02 %
prntly.com	0.26 %	nbcnews.com	1.96 %
worldtruth.tv	0.25 %	nypost.com	1.95 %
libertywritersnews.com	0.15 %	cbsnews.com	1.89 %
worldnewsdailyreport.com	0.06 %	abcnews.go.com	1.78 %
mediamass.net	0.04 %	time.com	1.71 %
newsbiscuit.com	0.03 %	cnbc.com	1.40 %
react365.com	0.02 %	washingtontimes.com	1.34 %
the-daily.buzz	0.02 %	washingtonexaminer.com	1.33 %

Table 6: Top 20 mainstream and alternative news sites in the Twitter dataset and their respective percentage.

present, as well as state-sponsored alternative domains like sputniknews.com and rt.com, which have recently been in the spotlight for disseminating false information and propaganda [7]. The fact that many such URLs appear in our dataset may indeed be an indication that Reddit significantly contributes to the dissemination of controversial stories.

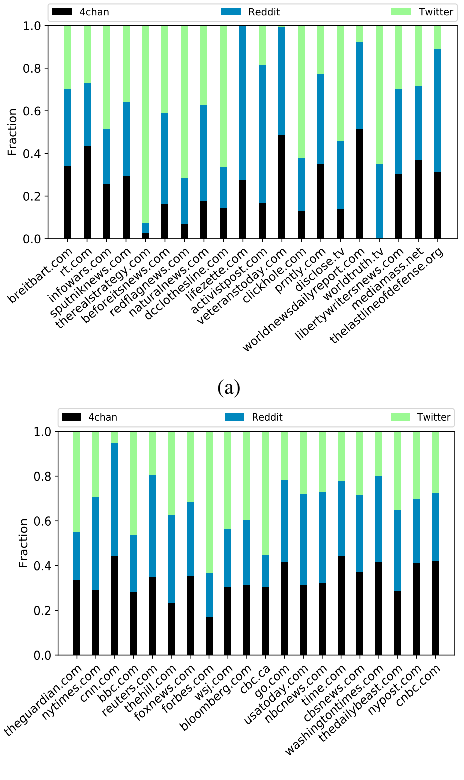
Twitter. In our Twitter dataset, we find 129k (42k unique) URLs of alternative news domains and 413k (236k unique) URLs of mainstream ones. Basic statistics are summarized in Table 5. Recall that we re-crawl tweets to get the number of retweets and likes, and a small percentage of them are no longer available as they were either deleted or the associated account was suspended. This percentage is slightly higher for tweets with URLs from alternative news, possibly due to the fact that some users tend to remove controversial content when a particular false story is debunked [10]. Also, alternative and mainstream news tend to get a significant number of retweets, at about the same rate (on average, 404 and 341 retweets per tweet, respectively). A similar pattern is observed for likes.

Then, in Table 6, we report the top 20 mainstream and alternative news domains, along with their percentage, in our Twitter dataset. These cover 86% and 99% of all URLs, respectively. Similar to Reddit, we note the presence of many popular alt-right as well as state-sponsored news outlets.

4chan. In our 4chan dataset, we find 21k (9k unique) URLs to alternative news outlets and 82k (40k unique) to mainstream news. Table 7 reports the percentage of URLs of the top 20 domains for each type of news. These cover 87% and 99% of, respectively, all mainstream and alternative news URLs. Again, we observe that, by far, the most influential alternative news domains are breitbart.com, rt.com, infowars.com, and sputniknews.com. For the mainstream news, we observe that theguardian.com is the most present one, followed by nytimes.com, cnn.com, and bbc.com. We also obtained similar statistics for the domain popularity in the other boards of 4chan but we omit them for brevity.

Domain (Alt.)	(%)	Domain (Main.)	(%)
breitbart.com	53.00 %	theguardian.com	14.10 %
rt.com	28.22 %	nytimes.com	10.07 %
infowars.com	9.12 %	cnn.com	9.90 %
sputniknews.com	3.36 %	bbc.com	5.45 %
veteranstoday.com	1.07 %	foxnews.com	5.35 %
beforeitsnews.com	0.91 %	reuters.com	5.10 %
lifezette.com	0.86 %	time.com	3.42 %
naturalnews.com	0.61 %	abcnews.go.com	3.40 %
worldnewsdailyreport.com	0.46 %	huffingtonpost.com	3.29 %
prntly.com	0.41 %	thehill.com	3.04 %
activistpost.com	0.38 %	wsj.com	2.82 %
dcclothesline.com	0.29 %	washinghtontimes.com	2.77 %
redflagnews.com	0.20 %	bloomberg.com	2.75 %
libertywritersnews.com	0.16 %	cbc.ca	2.66 %
therealstrategy.com	0.16 %	nypost.com	2.65 %
clickhole.com	0.11 %	cbsnews.com	2.44 %
disclose.tv	0.10 %	nbcnews.com	2.32 %
now8news.com	0.06 %	usatoday.com	2.25 %
firebrandleft.com	0.05 %	cnbc.com	2.13 %
nodisinfo.com	0.05 %	forbes.com	1.68 %

Table 7: Top 20 mainstream and alternative news sites in the 4chan (/pol/) dataset and their respective percentage.



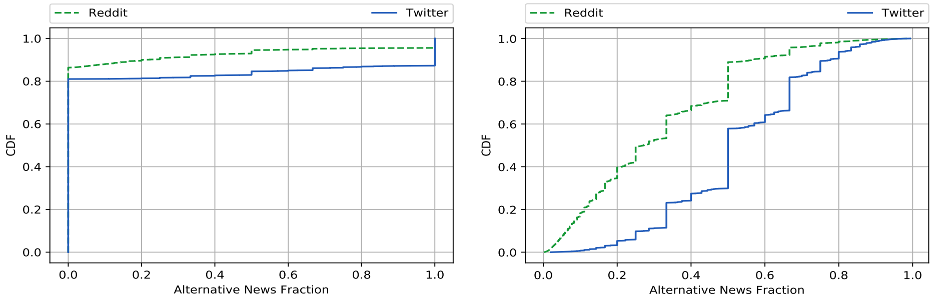
(a)

(b)

Figure 1: Top 20 domains and each platform's fraction for (a) alternative and (b) mainstream news.

4.2 Popular Domains

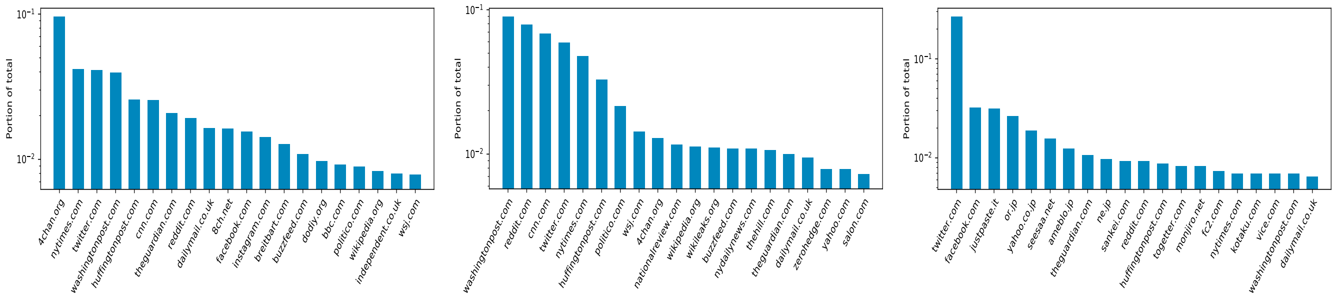
Next, we compare how popular domains, in both categories, appear on the three platforms (i.e., Twitter, the 6 subreddits, /pol/), as plotted in Fig. 1. We find that the



(a)

(b)

Figure 2: CDF of the fraction of URLs from alternative news and overall news URLs for (a) all users in our Twitter and Reddit datasets, and (b) users that shared URLs from both mainstream and alternative news.



(a)

(b)

(c)

Figure 3: Top 20 domains found after resolving archive.is URLs on (a) 4chan, (b) Reddit, and (c) Twitter.

top 4 alternative domains – breitbart.com, rt.com, infowars.com, sputniknews.com – influence the three platforms more or less in the same way. However, some outlets appear predominantly in some platforms but not in others – e.g., therealstrategy.com is popular only on Twitter, while lifezette.com and veteranstoday.com are popular on Reddit and 4chan, but not on Twitter.

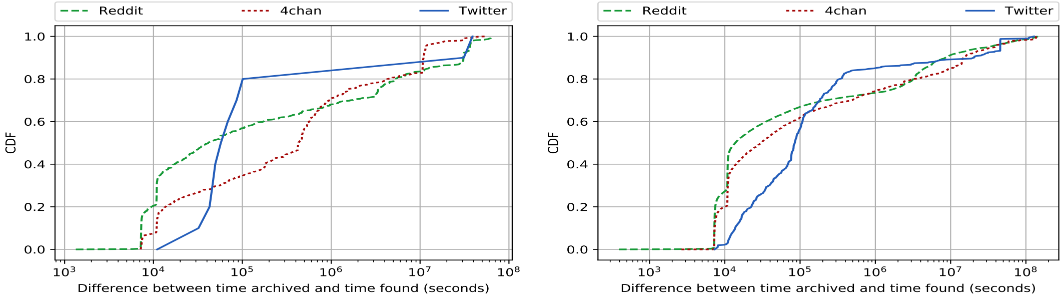
We believe the primary reason for this has to do with the reasonably well known phenomenon of Twitter bots. While we cannot say with any certainty that bots do not exist on 4chan, and bots are acceptable on Reddit (as long as they stay within the terms of service), they are certainly more prevalent on Twitter. Thus, if the reason that a particular domain is popular on Twitter is primarily due to the influence of bots, it follows that it would not be popular on Reddit and 4chan.

We also measure the fraction of news URLs that are alternative, *per user*, as plotted in Fig. 2. We report this fraction only for Reddit and Twitter users, since on 4chan posts are anonymous. We find that 80% of the users of both platforms share only URLs from mainstream news, while, 13% of Twitter users – that likely are bots [25] – only post URLs to alternative news. We also observe from Fig. 2(b), which shows the ratio for users sharing URLs from both categories, that there is a wide distribution, especially on Reddit, between people that rarely share alternative news (fraction close to 0) and those who share them almost all the time (fraction close to 1). Moreover, we find that that Twitter users share more alternative news as just 5% of these users have a fraction below 0.2, which might be also attributed to the presence of bots.

4.3 Time Capsules

Time capsule services are used to generate a short URL pointing to a snapshot of a web-page, so that users can access the content of that page even if it is later deleted, and without redirecting to the original site. Besides preserving content, time capsules can be used to obfuscate the original URL as well as to prevent additional traffic (and possibly ad revenue) from reaching the original web-page. We focus on archive.is, the most popular time capsule service used on 4chan and Reddit, aiming to analyze what content users habitually snapshot through the service, and what is the sharing behavior of such URLs.

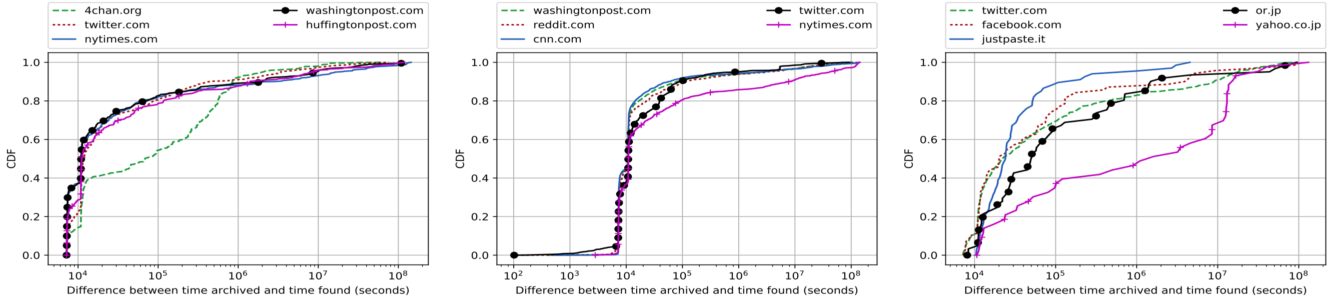
We retrieve all the archive.is URLs found on our 4chan, Reddit, and Twitter datasets: we find 5.2k (2.3k unique) URLs on Twitter, 27.3k (10.7k unique) on 4chan, and 92.1k (26k unique) on Reddit. These numbers indicate that archive.is is not so popular among the twitter community (just 0.001% of tweets contain such URLs) when compared to Reddit (0.028% of posts/comments) and 4chan (0.078% of posts). Next, we crawl these URLs to extract the time and the original page URL. We find that the fraction of alternative vs. mainstream news is 5%, 6% and 9% in Twitter, Reddit and 4chan respectively. In Fig. 3, we report, for each platform, the top 20 domains archived using archive.is.



(a)

(b)

Figure 4: CDF of the archival time and the first occurrence of archived URLs pointing to (a) alternative and (b) mainstream domains.



(a)

(b)

Figure 5: Time difference of top 5 archived domains in (a) 4chan; (b) Reddit and (c) Twitter.

Note that the site itself is among the most popular original domains, as the time capsules may be used for persistence, e.g., to access a 4chan thread after it is removed, or a tweet that has been deleted by the user. We also observe a strong presence of both alternative and mainstream news domains.

Fig. 4 plots the slack time between the archival time and the occurrence of the archive.is URL within a specific platform, comparing URLs pointing to mainstream and alternative news domains. With the former, Reddit and 4chan exhibit similar times, whereas, with the latter, Reddit is significantly faster than 4chan. To verify if different original web pages correspond to different slack times, we also plot the top 5 domains for each platform separately – see Fig. 5 – and find that archive.is URLs pointing to 4chan are considerably slower than the other domains. This indicates that users are more interested to archive the URL for persistence rather than sharing the content within 4chan. A similar behavior is observed on Twitter, with the slowest domain being yahoo.co.jp. For Reddit, we do not find any noticeable differences between the top domains.

5. TEMPORAL DYNAMICS

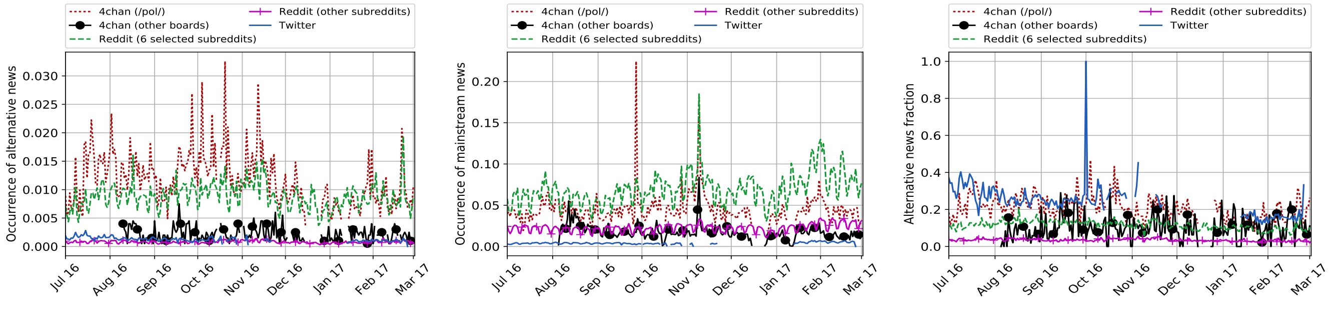
In this section, we present the results of a cross-platform temporal analysis of the way news are posted on Twitter, Reddit, and 4chan.

5.1 URLs Occurrence

In Fig. 6, we measure the daily occurrence of URLs over the three platforms normalized by the average daily number of URLs shared in each community.[7] We find that /pol/ and the 6 selected subreddits exhibit a much higher percentage of URLs occurrence to alternative news compared to the other communities (Fig. 6(a)), whereas, for mainstream news, the sharing behavior is more similar (Fig. 6(b)) across platforms. There are also some interesting spikes, likely to be related to the 2016 US elections, on the date of the first presidential debate and the election day itself. These findings indicate that the specific sub-communities are heavily utilized for the dissemination of alternative news. We also study the fraction between alternative and overall news URLs (Fig. 6(c)), highlighting that the latter (dominated by mainstream news) are overall more "popular" than the former. The Twitter spike in Fig. 6(c) appears to be an artifact of a failure in our collection infrastructure.

As some users repost the same URL many times within the same platform, we decide to study such reposting behavior and extract insight while comparing platforms. In Fig. 7, we plot the CDF of the time difference between the first occurrence of a URL and its next occurrences on the same platform. Both alternative and mainstream URLs are recycled over time within the platform (even after several months), while Twitter exhibits smaller time differences between the first occurrence and the next ones than the other two platforms. In all three platforms, there is an inflection point at the 24h period, which probably signifies the day-to-day behavior of news propagation within a platform, and this is true for both alternative and mainstream news. Finally,

[7] Gaps in the plot correspond to gaps in our dataset due to crawler failure.



(a)

(b)

(c)

Figure 6: Normalized daily occurrence of URLs for (a) alternative news, (b) mainstream news, and (c) fraction of alternative news over overall news.

Comparison	Type of News	#URLs where platform 1 is faster	#URLs where platform 2 is faster
Reddit vs Twitter	Mainstream	18,762	11,416
	Alternative	5,232	4,301
4chan vs Twitter	Mainstream	2,938	4,700
	Alternative	778	2,099
4chan vs Reddit	Mainstream	5,382	14,662
	Alternative	1,455	3,695

Table 8: Statistics of URLs for the comparisons of time difference between platforms

mainstream news seem to propagate faster in these platforms than alternative news, especially in Reddit; for Twitter and 4chan the difference is not evident.

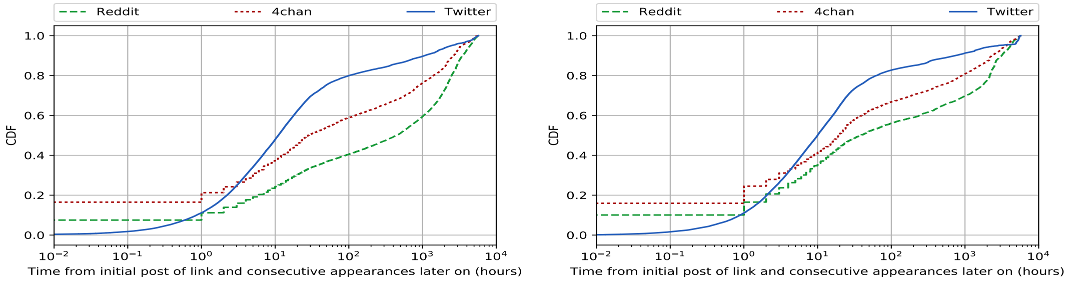
We also study the inter-arrival time of reposted URLs. Fig. 8 shows the CDF of the mean inter-arrival time of URLs that appear more than one time in each platform. Each platform exhibits unique behavior, confirmed by a two sample KolmogorovSmirnov test showing significant differences between the distributions ($p < 0.01$ for each pairwise comparison). However, 4chan and Reddit exhibit similar time-related sharing behavior for both mainstream and alternative URLs, and Twitter has smaller mean inter-arrival time overall. Interestingly, Reddit appears to have a duality in reposting behavior: for URLs with small inter-arrival time, it follows the faster pace of Twitter, whereas, for URLs with longer inter-arrival times, it follows 4chan.

5.2 Cross Platform Analysis

We now look at URLs that appear on more than one platform and study the time at which they are shared. Fig. 9 plots the CDF of the time difference (in seconds) between the first occurrence of a URL on pairs of platforms, while Table 8 reports the numbers of involved URLs for each comparison, We make the following observations. First, when comparing pairs of distributions for a given category of URLs, they are statistically different, with p-value $< 10^{-4}$. Second, alternative news appear on different platforms faster than mainstream news. This is consistent regardless of the pair of platforms we consider, and the sequence of appearances (i.e., first in platform A and then B, vs. first in B and then in A). Third, we notice the presence of a "turning point" with respect to the delay between URL appearance on each platform, which seems to be consistent across all pairs of platforms and types of news, and matches the 24h period observed earlier. Finally, there is a cross point when comparing URLs first posted on platform A and then on B, and URLs which were posted first in B and then A (i.e., when the lines for the same type of URLs cross). Such a point signifies which portion of URLs appear faster in one platform than the other. For Twitter-Reddit comparison, alternative (respectively, mainstream) news appear faster on Twitter than Reddit for 80% (resp., 50%) of the time, with these URLs being slower in propagation, since the switching point is at $\sim$1 hour (resp., 5 hours). Similarly, for Twitter-4chan comparison, alternative (mainstream) news appear faster on Twitter than 4chan for 70% (65%) of the time, with the switching point being at 1 day (2 days). Finally, for Reddit-4chan comparison, alternative (mainstream) news appear faster on Reddit than 4chan for 65% (40%) of the time, with the switching point being at 18 hours (12 hours).

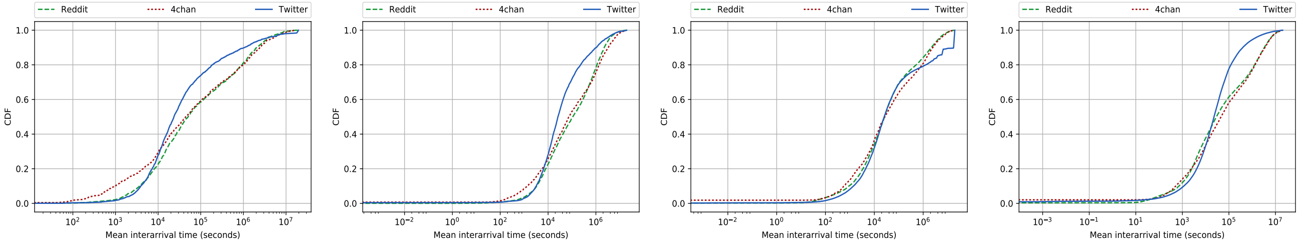
Next, given the set of unique URLs across all platforms and the time they appear for the first time, we analyze their appearance in one, two, or three platforms, and the order in which this happens. For each URL, we find the first occurrence on each platform and build corresponding "sequences," e.g., if a URL first appears on Reddit and subsequently on 4chan, the sequence is Reddit→ 4chan (R→4). Table 9 reports the distribution of the sequences of appearances considering only the first hop, i.e., up to the first two platforms in the sequence. We observe that the majority of URLs only appear on one platform: 82% of alternative URLs and 89% of mainstream URLs. Also, both alternative and mainstream URLs tend to appear on Reddit first and later appear on either Twitter or 4chan, and on Twitter before 4chan.

We also study the temporal dynamics of URLs that appear on all three platforms, with triplets of sequences. Table 10 reports the distribution of these sequences. The most common sequences are similar for both alternative and mainstream URLs: R→T→4, R→4→T, and T→R→4 are the top 3 sequences. As already mentioned, Reddit "outperforms" both other platforms in terms of the speed of sharing mainstream and alternative URLs, as evidenced by the fact that it is at the head of the sequence for 51% and 59% of alternative



(a) (b)

Figure 7: CDF of time difference (in hours) between the first occurrence of a URL and its next occurrences on each platform for (a) alternative and (b) mainstream news.



(a) (b) (c) (d)

Figure 8: CDF for mean inter-arrival time for the URLs that occur more than once for (a) common alternative news URLs; (b) common mainstream news URLs; (c) all alternative news URLs, and (d) all mainstream news URLs.

Sequence	Alternative (%)	Mainstream (%)
4 only	3,236 (4.4%)	18,654 (3.7%)
4→R	1,118 (1.5%)	4,606 (0.9%)
4→T	315 (0.5%)	861 (0.17%)
R only	24,292 (33.3%)	230,602(46.1%)
R→4	2,181 (3.0%)	11,307 (2.3%)
R→T	4,769 (6.5%)	16,685 (3.35%)
T only	32,443 (44.5%)	204,836 (41%)
T→4	585 (0.8%)	1,345 (0.26%)
T→R	3,964 (5.5%)	10,640 (2.12%)

Table 9: Distribution of URLs according to the sequence of first appearance within platforms for all URLs, considering only the first hop.

and mainstream URLs, respectively.

Finally, we analyze the source of the URLs for each of the three platforms, using graph model and analysis techniques. We create two directed graphs, one for each type of news, $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E})$, where $\boldsymbol{V}$ represents alternative or mainstream domains, as well as the three platforms, and $\boldsymbol{E}$ the set of sequences that consider only the first-hop of the platforms. For example, if a breitbart.com URL appears first on Twitter and later on Reddit, we add an edge from breitbart.com to Twitter, and from Twitter to Reddit. We also add weights on these edges based the number of such unique URLs. By examining the paths, we can discern which domains URLs tend to appear first on each of the platforms.

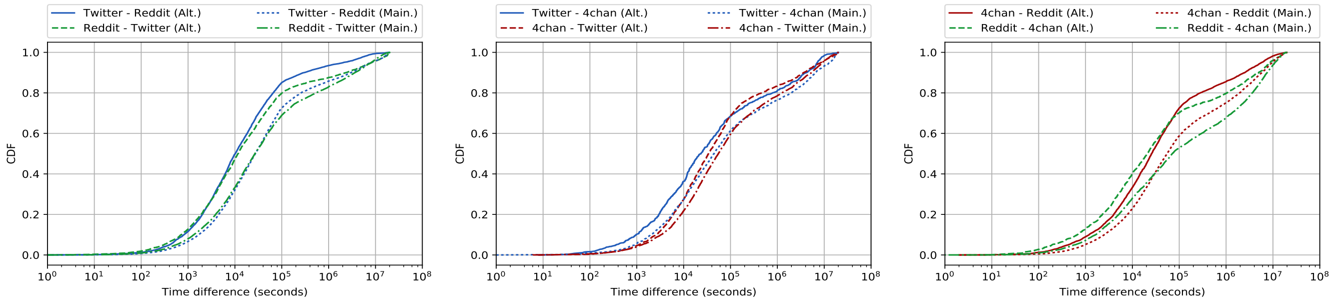
Fig. 10 shows the graphs built for alternative and mainstream domains. Comparing the outgoing edges' thickness, one can see that breitbart.com URLs appear first in Reddit more often than on Twitter and more frequently than they do on 4chan. However, for other popular alternative domains, such as infowars.com, rt.com, and sputniknews.com, URLs appear first on Twitter more often than Reddit and 4chan. Also, 4chan is rarely the platform where a URL first spawns. For the mainstream news domains, we note that URLs from nytimes.com and cnn.com tend to appear first more often on Reddit than Twitter and 4chan, however, URLs from other domains like bbc.com and theguardian.com tend to appear first more often on Twitter than Reddit. Similar to the alternative domains graph, there is no domain where 4chan dominates in terms of first URL appearance.

Sequence	Alternative (%)	Mainstream (%)
4→R→T	128 (5.5%)	552 (8.9%)
4→T→R	145 (6.2%)	290 (4.7%)
R→4→T	335 (14.4%)	1,525 (24.5%)
R→T→4	841 (36.3%)	2,189 (35.3%)
T→4→R	192 (8.2%)	486 (7.8%)
T→R→4	673 (29%)	1,166 (18.8%)

Table 10: Distribution of URLs according to the sequence of first appearance within a platform for URLs common to all platforms. "4" stands for 4chan, "R" for Reddit, and "T" for Twitter.

6. INFLUENCE ESTIMATION

Thus far, our measurements have shown relative differences in how news media is shared on Reddit, Twitter, and 4chan. In this section, we set to provide meaningful evidence of how the individual platforms influence the media shared on other platforms. We do so by using a mathemati-

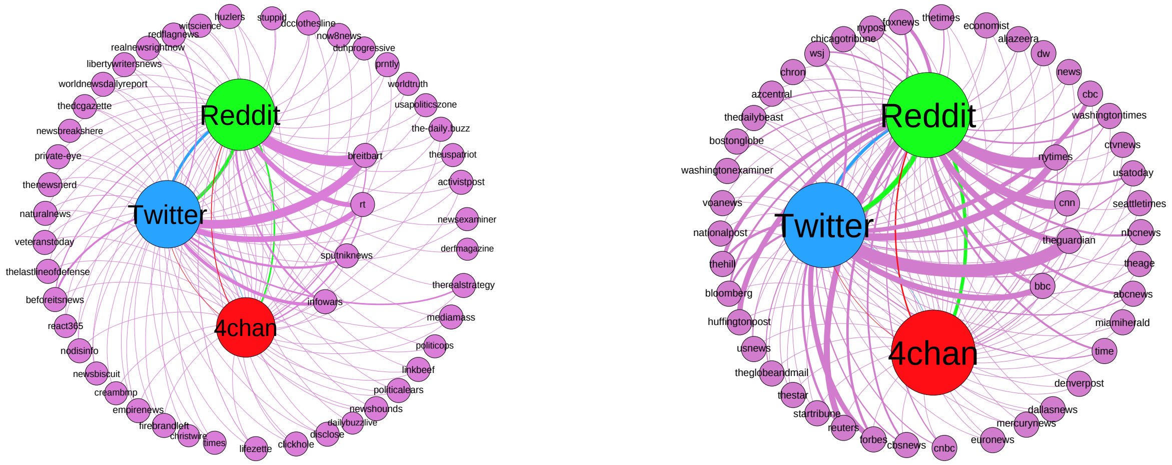


(a)

(b)

(c)

Figure 9: CDF of the difference between the first occurrence of a URL between (a) Reddit and Twitter, (b) 4chan and Twitter, and (c) 4chan and Reddit.



(a)

(b)

Figure 10: Graph representation of news ecosystem (a) alternative news domains and (b) mainstream news domains. Edges are colored the same as their source node.

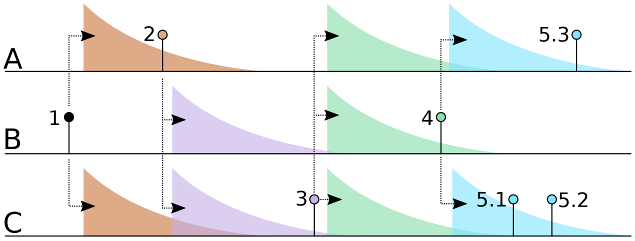


Figure 11: A depiction of a Hawkes model showing the interaction between events on 3 processes.

cal technique known as Hawkes processes. First, we provide a high level intuition of the analysis.

The three platforms we measure do not exist in a vacuum, but rather within the greater ecosystem of the Web. Imagine, however, that each of the platforms was entirely self-contained, with a completely disjoint set of users. In such a scenario, there would be a natural rate at which URLs will be posted, and it would be possible to model this using standard Poisson processes. However, our platforms are clearly *not* independent. While they do exhibit their own *background* URL posting rates and internal influence, they are also affected by each other, as well as by the greater Web.

A Hawkes model consists of a number, K, of point processes, each with a "background rate" of events $\lambda_{0,k}$. An event on one process can cause an impulse response on other processes, increasing their rates. Fig. 11 depicts a sequence of events on a Hawkes process with three processes. An initial event 1 is caused by the background rate of process B. This event causes an impulse response on the rates of the other processes, A and C, eventually causing event 2 on process A. A process can cause an additional impulse response to itself, as seen with event 3, and multiple events can be caused in response to a single event, as seen with event 4 causing events 5.1, 5.2, and 5.3.

For a discrete-time Hawkes model, time is divided into a series of bins of duration Δt, and events occurring within the same time bin do not interact with each other. The rate of each k-th process, $\lambda_{t,k}$ is given by:

$$\lambda_{t,k} = \lambda_{0,k} + \sum_{k'=1}^{K} \sum_{t'=1}^{t-1} s_{t',k'} \cdot h_{k' \to k}[t - t']$$

where $s \in \mathbb{N}^{T \times K}$ is the matrix of event counts (how many events occur for process k at time t) and $h_{k' \to k}[t - t']$ is an impulse response function that describes the amplitude of influence that events on process k' have on the rate of process k.

Following [15], the impulse response function $h_{k \to k'}[t - t']$ can be decomposed into a scalar weight $W_{k \to k'}$ and a probability mass function $G_{k \to k'}[d]$. The weight specifies the strength of the interaction from process k to process k' and the probability mass function specifies how the interaction changes over time:

$$h_{k \to k'}[d] = W_{k \to k'} G_{k \to k'}[d]$$

The weight value $W_{k \to k'}$ can be interpreted as the expected number of child events that will be caused on process k' after an event on process k. The probability mass function $G_{k \to k'}$ specifies the probability that a child event will occur at each specific time lag $d\Delta t$, up to a maximum lag Δt_{max}.

This interpretation of $W_{k \to k'}$ is useful because it will allow us to compare how much influence platforms have on each other. For instance, we will be able to examine whether a URL posted on Twitter or on Reddit is more likely to cause the same URL to be posted on 4chan, or if there is a difference in influence from one platform to another between URLs for mainstream and alternative news sites.

6.1 Methodology

We now provide more details about our experiments. Once again, we consider 4chan (/pol/), Twitter, and the 6 selected subreddits from Reddit. We study the Hawkes process in the subreddit granularity in order to get a better understanding of the various platforms and subreddits. We aim to examine how these platforms and subreddits influence each other, so we model the arrival of URLs, in posts or tweets, with a Hawkes model with $K = 8$ point processes—one each for Twitter, /pol/, and each of the subreddits. The model is fully connected, i.e., it is possible for each process to influence all the others, as well as itself, which describe behavior where participants on a platform see a URL and re-post it on the same platform. For example, with Twitter, this value ($W_{\text{Twitter} \to \text{Twitter}}$) would likely be quite high, given that tweets are commonly re-tweeted a number of times: the initial tweet containing a URL is likely to cause a number of re-tweets, also containing the URL, on the same platform.

We select URLs that have at least one event in Twitter, /pol/, and at least one of the subreddits, and we model each URL individually. The missing Twitter data affects 3177 (37%) of the URLs. To lessen the impact of the missing data, we remove the 10% of URLs (895) from those that overlap any of the missing days with the shortest total duration from the first event recorded until the last event. This increases the smallest amount of Twitter data included in the remaining URLs and allows us to keep long-duration URLs, which are more likely to overlap with the missing dates, and contain a large proportion of the total events.

The number of remaining URLs and events included for each platform are shown in Table 11. For each URL, we create a matrix $s \in \mathbb{N}^{T \times 8}$ containing the number of events (URL posts) per minute for each of the platforms/subreddits. Here, T is the number of minutes from the first recorded post of the URL on any platform, to the last recorded post of a URL on any platform, and this value can be different for each URL. We select $\Delta t = 1 \text{minute}$ as a reasonable compromise between accuracy and computational cost. Using this bin size, 92% of events are in a bin by themselves, and another 5.4% share a bin, but only with other events from the same platform or subreddit, meaning that timing interactions between the platforms are not lost.

Next, we fit Hawkes processes using Gibbs sampling as described in [16]. By setting $\Delta t_{max} = 60 \cdot 12 = 720$, we say that a given event can cause other events within a 12-hour time window. Tests with other values (6, 12, 24, and 48 hours) gave similar results. After fitting the model, we have the values for the W matrix – i.e., the weights of the interactions between events on different processes. These weights can then be interpreted as the expected number of events. For example, $W_{twitter \to /pol/} = 0.1$ would mean that an event on Twitter will cause n events on /pol/, where n is drawn from a Poisson distribution with rate parameter 0.1. Finally, we also get $\lambda_{0,k}$ for each process, which is the background rate for event arrivals that are *not* caused by other events in the system we model. This background rate captures both the "natural" appearance of events (such as someone posting the URL after reading it on the original site) as well as those caused by events outside the platforms we measure (where someone posts the URL after seeing it posted elsewhere).

6.2 Results

Looking at the number of URLs in Table 11, we note that there are substantially more events for mainstream than alternative URLs. However, for Twitter, /pol/, and The_Donald, the ratios of events to URLs for alternative URLs are similar to or greater than the ratios for mainstream URLs. These high ratios explain the high background rates (also in Table 11) for alternative URLs for these platforms despite the lower number of alternative URL events.

From the Hawkes models for each URL, we obtain the weight matrix W which specifies the strength of the connections between the different platforms and subreddits. The mean weight values over all URLs for alternative and mainstream URLs, as well as the percentage difference between them are presented in Figure 12. Since the weight values can be interpreted as the expected number of additional events that will be caused a consequence of an event, we can estimate the percentage of events on each platform that were caused by each of the other platforms by multiplying the weight by the actual number of events that occurred on the source platform and dividing by the number of events that

		The_Donald	worldnews	politics	news	conspiracy	AskReddit	/pol/	Twitter
URLs	Mainstream	3,097	2,523	3,578	2,584	907	841	5,589	5,589
	Alternative	2,008	252	813	362	321	100	2,136	2,136
	Total	5,105	2,775	4,391	2,946	1,228	941	7,725	7,725
Events	Mainstream	12,312	7,517	26,160	5,794	1,995	2,302	19,746	36,250
	Alternative	7,797	458	2,484	586	497	176	7,322	23,172
	Total	20,109	7,975	28,644	6,380	2,492	2,478	27,068	59,422
Mean λ_0	Mainstream	0.001502	0.001382	0.001265	0.001392	0.000501	0.000107	0.001564	0.002330
	Alternative	0.001627	0.000619	0.000696	0.000553	0.000423	0.000034	0.001525	0.002803

Table 11: Total URLs with at least one event in Twitter, /pol/, and at least one of the subreddits; total events for mainstream and alternative URLs, and the mean background rate (λ_0) for each platform/subreddit.

occurred on the destination platform:

$$\text{Pct}_{A\to B} = \frac{\sum_{u\in\text{urls}}\left(W_{A\to B}\cdot\sum_{t=1}^{T} s_{t,A}\right)}{\sum_{u\in\text{urls}}\sum_{t=1}^{T} s_{t,B}}$$

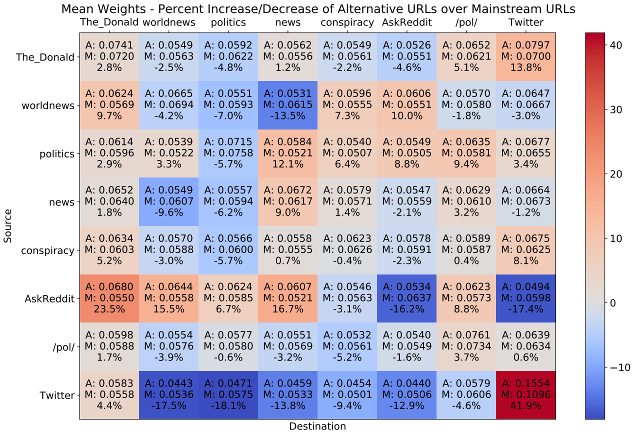
These percentages for mainstream URLs, alternative URLs and the difference between them are presented in Figure 13.

First, we look at Twitter. Background rates are high for both mainstream and alternative URLs, which is not surprising given the large number of users on the platform. The values for $W_{\text{Twitter}\to\text{Twitter}}$ are also substantially higher than all other weights: 0.1096 for mainstream URLs and 0.1554 for alternative URLs. This reflects the ease and common practice of re-tweeting: a URL in a tweet is likely to generate other events as users re-tweet it.

There are different possible explanations for why the Twitter to Twitter rate for alternative URLs is much greater than the rate for mainstream URLs. The first is bot activity—if automated Twitter bots are used to spread alternative URLs, it could result in a much higher rate of tweeting and re-tweeting. Another possible explanation is the behavior of users who read news stories from alternative sources; they might be more inclined to re-tweet the URL [11].

Looking at the weights for Twitter to the other platforms, all except The_Donald are greater for mainstream URLs, meaning that the average tweet containing a mainstream URL is more likely to cause a subsequent post on the other platforms than the average tweet containing an alternative URL. The next platforms most likely to cause events on other platforms are The_Donald and /pol/. It is worth noting that The_Donald is the only platform/subreddit that has greater alternative URL weights for all of its inputs. Assuming the population of The_Donald users also reading, say, worldnews is the same for both alternative and mainstream URLs—which is reasonable—then the difference in weights implies that the users have a stronger preference for re-posting alternative URLs back to The_Donald than for mainstream URLs. The opposite can be seen for worldnews and politics, where most of the input weights are stronger for mainstream news.

However, despite the higher weights for alternative URLs, The_Donald is also, interestingly, influenced more strongly by mainstream URLs than alternative URLs on all platforms, with the exception of Twitter. This is in part because of the greater number of mainstream URL events, but The_Donald also has a higher background rate for alternative URLs than mainstream URLs, which implies that a lot of the alternative URLs on the platform are coming from other sources.

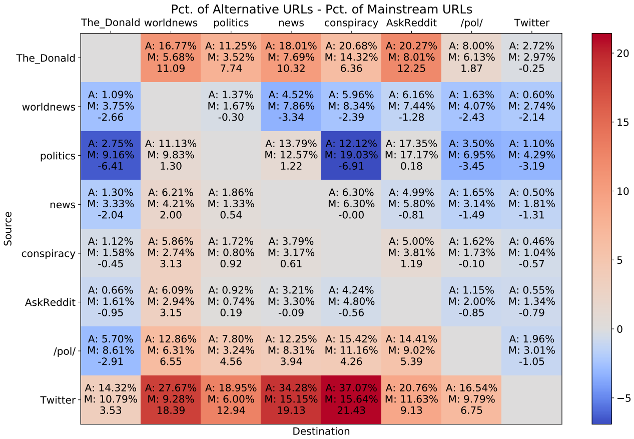

Mean Weights - Percent Increase/Decrease of Alternative URLs over Mainstream URLs

Source \ Destination	The_Donald	worldnews	politics	news	conspiracy	AskReddit	/pol/	Twitter
The_Donald	A: 0.0741 M: 0.0720 2.8%	A: 0.0549 M: 0.0563 -2.5%	A: 0.0592 M: 0.0622 -4.8%	A: 0.0562 M: 0.0556 1.2%	A: 0.0549 M: 0.0561 -2.2%	A: 0.0526 M: 0.0551 -4.6%	A: 0.0652 M: 0.0621 5.1%	A: 0.0797 M: 0.0700 13.8%
worldnews	A: 0.0624 M: 0.0569 9.7%	A: 0.0665 M: 0.0694 -4.2%	A: 0.0551 M: 0.0593 -7.0%	A: 0.0531 M: 0.0615 -13.5%	A: 0.0596 M: 0.0555 7.3%	A: 0.0606 M: 0.0551 10.0%	A: 0.0570 M: 0.0580 -1.8%	A: 0.0647 M: 0.0667 -3.0%
politics	A: 0.0614 M: 0.0596 2.9%	A: 0.0539 M: 0.0522 3.3%	A: 0.0715 M: 0.0758 -5.7%	A: 0.0584 M: 0.0521 12.1%	A: 0.0540 M: 0.0507 6.4%	A: 0.0549 M: 0.0505 8.8%	A: 0.0635 M: 0.0581 9.4%	A: 0.0677 M: 0.0655 3.4%
news	A: 0.0652 M: 0.0640 1.8%	A: 0.0549 M: 0.0607 -9.6%	A: 0.0557 M: 0.0594 -6.2%	A: 0.0672 M: 0.0617 9.0%	A: 0.0579 M: 0.0571 1.4%	A: 0.0547 M: 0.0559 -2.1%	A: 0.0629 M: 0.0610 3.2%	A: 0.0664 M: 0.0673 -1.2%
conspiracy	A: 0.0634 M: 0.0603 5.2%	A: 0.0570 M: 0.0588 -3.0%	A: 0.0566 M: 0.0600 -5.7%	A: 0.0558 M: 0.0555 0.7%	A: 0.0623 M: 0.0626 -0.4%	A: 0.0578 M: 0.0591 -2.3%	A: 0.0589 M: 0.0587 0.4%	A: 0.0675 M: 0.0625 8.1%
AskReddit	A: 0.0680 M: 0.0550 23.5%	A: 0.0644 M: 0.0558 15.5%	A: 0.0624 M: 0.0585 6.7%	A: 0.0607 M: 0.0521 16.7%	A: 0.0546 M: 0.0563 -3.1%	A: 0.0534 M: 0.0637 -16.2%	A: 0.0623 M: 0.0573 8.8%	A: 0.0494 M: 0.0598 -17.4%
/pol/	A: 0.0598 M: 0.0588 1.7%	A: 0.0554 M: 0.0576 -3.9%	A: 0.0577 M: 0.0580 -0.6%	A: 0.0551 M: 0.0569 -3.2%	A: 0.0532 M: 0.0561 -5.2%	A: 0.0540 M: 0.0549 -1.6%	A: 0.0761 M: 0.0734 3.7%	A: 0.0639 M: 0.0634 0.6%
Twitter	A: 0.0583 M: 0.0558 4.4%	A: 0.0443 M: 0.0536 -17.5%	A: 0.0471 M: 0.0575 -18.1%	A: 0.0459 M: 0.0533 -13.8%	A: 0.0454 M: 0.0501 -9.4%	A: 0.0440 M: 0.0506 -12.9%	A: 0.0579 M: 0.0606 -4.6%	A: 0.1554 M: 0.1096 41.9%

Figure 12: The mean weights for alternative URLs (A), the mean weights for mainstream URLs (M), and the percent increase/decrease between mainstream and alternative (also indicated by the coloration). Look along a row to see outputs and down a column to see inputs.

Figure 13 shows the estimated total impact of the different platforms on each other, for both mainstream URLs and alternative URLs. Twitter contributes heavily to both types of events on the other platforms—and is in fact the most influential single source for most of the other platforms. Despite Twitter's lower weights for alternative URLs, it actually has a greater influence on alternative URLs than mainstream URLs, in terms of percentage of events caused, on all the other platforms or subreddits. This is due to the fact that, even though it has lower weights, the largest proportion of alternative URL events are on the twitter platform.

After Twitter, The_Donald and /pol/ also have a strong influence on the alternative URLs that get posted on other platforms. The_Donald has a stronger effect for alternative URLs on all platforms except Twitter—although it still has the largest alternative influence on Twitter, causing an estimated 2.72% of alternative URLs tweeted. Interestingly, we observe that The_Donald causes 8% of /pol/'s alternative URLs, while /pol/'s influence on The_Donald is less, at 5.7%. For the mainstream URLs the strength of influence is reversed. Specifically, /pol/'s influence on The_Donald is



	The_Donald	worldnews	politics	news	conspiracy	AskReddit	/pol/	Twitter
The_Donald		A: 16.77% M: 5.68% 11.09	A: 11.25% M: 3.52% 7.74	A: 18.01% M: 7.69% 10.32	A: 20.68% M: 14.32% 6.36	A: 20.27% M: 8.01% 12.25	A: 8.00% M: 6.13% 1.87	A: 2.72% M: 2.97% -0.25
worldnews	A: 1.09% M: 3.75% -2.66		A: 1.37% M: 1.67% -0.30	A: 4.52% M: 7.86% -3.34	A: 5.96% M: 8.34% -2.39	A: 6.16% M: 7.44% -1.28	A: 1.63% M: 4.07% -2.43	A: 0.60% M: 2.74% -2.14
politics	A: 2.75% M: 9.16% -6.41	A: 11.13% M: 9.83% 1.30		A: 13.79% M: 12.57% 1.22	A: 12.12% M: 19.03% -6.91	A: 17.35% M: 17.17% 0.18	A: 3.50% M: 6.95% -3.45	A: 1.10% M: 4.29% -3.19
news	A: 1.30% M: 3.33% -2.04	A: 6.21% M: 4.21% 2.00	A: 1.86% M: 1.33% 0.54		A: 6.30% M: 6.30% -0.00	A: 4.99% M: 5.80% -0.81	A: 1.65% M: 3.14% -1.49	A: 0.50% M: 1.81% -1.31
conspiracy	A: 1.12% M: 1.58% -0.45	A: 5.86% M: 2.74% 3.13	A: 1.72% M: 0.80% 0.92	A: 3.79% M: 3.17% 0.61		A: 5.00% M: 3.81% 1.19	A: 1.62% M: 1.73% -0.10	A: 0.46% M: 1.04% -0.57
AskReddit	A: 0.66% M: 1.61% -0.95	A: 6.09% M: 2.94% 3.15	A: 0.92% M: 0.74% 0.19	A: 3.21% M: 3.30% -0.09	A: 4.24% M: 4.80% -0.56		A: 1.15% M: 2.00% -0.85	A: 0.55% M: 1.34% -0.79
/pol/	A: 5.70% M: 8.61% -2.91	A: 12.86% M: 6.31% 6.55	A: 7.80% M: 3.24% 4.56	A: 12.25% M: 8.31% 3.94	A: 15.42% M: 11.16% 4.26	A: 14.41% M: 9.02% 5.39		A: 1.96% M: 3.01% -1.05
Twitter	A: 14.32% M: 10.79% 3.53	A: 27.67% M: 9.28% 18.39	A: 18.95% M: 6.00% 12.94	A: 34.28% M: 15.15% 19.13	A: 37.07% M: 15.64% 21.43	A: 20.76% M: 11.63% 9.13	A: 16.54% M: 9.79% 6.75	

Figure 13: The estimated mean percentage of alternative URL events caused by alternative URL events (A), the estimated mean percentage of mainstream URL events caused by mainstream URL events (M), and the difference between alternative and mainstream (also indicated by the coloration). Look along a row to see outputs and down a column to see inputs.

8.61% whereas The_Donald's influence on /pol/ is 6.13%.

In descending order, the influences on Twitter for mainstream URLs are politics (4.29%), /pol/ (3.01%), The_Donald (2.97%), worldnews (2.74%), news (1.81%), AskReddit (1.34%), and conspiracy (1.04%). The strongest influences for alternative URLs are, unsurprisingly, The_Donald (2.72%) and /pol/ (1.96%), followed by politics (1.10%), worldnews (0.60%), AskReddit (0.55%), news (0.50%), and conspiracy (0.46%). Twitter influences the alternative URLs on other platforms to a large degree—but the largest alternative URL inputs to Twitter are The_Donald and /pol/. We are only looking at a closed system of 8 different platforms and subreddits, but Twitter is undoubtedly effective at propagating information, and the influence these two platforms have on Twitter would likely spread widely.

7. DISCUSSION & CONCLUSION

In this work, we explored how mainstream and fringe Web communities share mainstream and alternative news sources with a particular focus on how communities influence *each other*. We collected millions of posts from Twitter, Reddit, and 4chan and analyzed the occurrence and temporal dynamics of news shared from 45 mainstream and 54 alternative news sites. We found that users on the different platforms prefer different news sources, especially when it comes to alternative ones. We also explored complex temporal dynamics and we discovered, for example, that Twitter and Reddit users tend to post the same stories within a relatively short period of time, with 4chan posts lagging behind both of them. However, when a story becomes popular after a day or two, it is usually the case it was posted on 4chan first.

Using Hawkes processes, we also modeled the influence the individual platforms have on each other, finding that the interplay between platforms manifests in subtle, yet meaningful ways. For example, of all the platforms and subreddits, Twitter by far has the most influence in terms of the number of URLs it causes to be posted to other platforms, and contributes to the share of alternative news URLs on the other platforms to a much greater degree than to the share of mainstream URLs. After Twitter, The_Donald subreddit and 4chan are the next most influential when it comes to alternative URLs. For alternative URLs, The_Donald is less-influenced by the other platforms than 4chan, and has a higher background rate, meaning more of the URLs posted there come from other sources.

To the best of our knowledge, our analysis constitutes the first attempt to characterize the dissemination of mainstream and alternative news across multiple social media platforms, and to estimate a quantifiable influence between them. Naturally, our work has some limitations. Since we only look at the posting of third party URLs, we miss other modalities of diffused information. For instance, while we found that time-capsuled links to 4chan were present on Reddit, there may also be a lot of direct information transferred from 4chan occurring via screenshots, due to its ephemeral nature. Also, we did not examine the content of the news stories shared.

As part of future work, we plan to explore advanced image recognition techniques to look for screenshots shared among the different platforms, as well as Natural Language Processing to determine whether stories become a part of the platform's narrative of events – i.e., whether users continue to talk about stories without actually posting a relevant URL itself. We believe efforts into understanding how the growing phenomenon of alternative information sources affects multiple platforms can help inform detection and mitigation techniques against misinformation and disinformation campaigns, and our work constitutes a first step in that direction.

Acknowledgments. This research is supported by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie "ENCASE"project (Grant Agreement No. 691025).

8. REFERENCES

[1] H. Allcott and M. Gentzkow. Social media and fake news in the 2016 election. Technical report, National Bureau of Economic Research, 2017.

[2] C. Andrews, E. Fichet, Y. Ding, E. S. Spiro, and K. Starbird. Keeping up with the tweet-dashians: The impact of'official'accounts on online rumoring. In *CSCW*, 2016.

[3] A. Arif, K. Shanahan, F.-J. Chou, Y. Dosouto, K. Starbird, and E. S. Spiro. How information snowballs: Exploring the role of exposure in online rumor propagation. In *CSCW*, 2016.

[4] D. Beran. 4chan: The Skeleton Key to the Rise of Trump. https://medium.com/@DaleBeran/4chan-the-skeleton-key-to-the-rise-of-trump-624e7cb798cb, 2017.

[5] C. Budak, D. Agrawal, and A. El Abbadi. Limiting the spread of misinformation in social networks. In *WWW*, 2011.

[6] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali. Hate is not binary: Studying abusive behavior of #gamergate on twitter. In *WebSci*, 2017.

[7] L. Dearden. Nato accuses Sputnik News of distributing misinformation as part of 'Kremlin propaganda machine'. http://ind.pn/2luLjs0, 2016.

[8] M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi. The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113(3), 2016.

[9] S. Finn, P. T. Metaxas, and E. Mustafaraj. Investigating Rumor Propagation with TwitterTrails. *arXiv preprint 1411.3550*, 2014.

[10] A. Friggeri, L. A. Adamic, D. Eckles, and J. Cheng. Rumor cascades. In *ICWSM*, 2014.

[11] A. Gupta, H. Lamba, P. Kumaraguru, and A. Joshi. Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In *WWW*, 2013.

[12] G. E. Hine, J. Onaolapo, E. De Cristofaro, N. Kourtellis, I. Leontiadis, R. Samaras, G. Stringhini, and J. Blackburn. Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan's Politically Incorrect Forum and Its Effects on the Web. In *ICWSM*, 2017.

[13] S. Kumar, R. West, and J. Leskovec. Disinformation on the Web: Impact, Characteristics, and Detection of Wikipedia Hoaxes. In *WWW*, 2016.

[14] S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang. Prominent features of rumor propagation in online social media. In *ICDM*, 2013.

[15] S. W. Linderman and R. P. Adams. Discovering Latent Network Structure in Point Process Data. In *ICML*, 2014.

[16] S. W. Linderman and R. P. Adams. Scalable Bayesian Inference for Excitatory Point Process Networks. *ArXiv pre-print 1507.03228*, 2015.

[17] M. Mendoza, B. Poblete, and C. Castillo. Twitter Under Crisis: Can we trust what we RT? In *SOMA*, 2010.

[18] D. T. Nguyen, N. P. Nguyen, and M. T. Thai. Sources of misinformation in online social networks: Who to suspect? In *MILCOM*, 2012.

[19] E. Seo, P. Mohapatra, and T. Abdelzaher. Identifying rumors and their sources in social networks. In *SPIE*, 2012.

[20] D. Shah and T. Zaman. Rumors in a network: Who's the culprit? *IEEE Transactions on information theory*, 57(8), 2011.

[21] C. Shao, G. L. Ciampaglia, A. Flammini, and F. Menczer. Hoaxy: A Platform for Tracking Online Misinformation. In *WWW Companion*, 2016.

[22] H. Situngkir. Spread of Hoax in Social Media. BFI Working Paper No. WP-4-2011, 2011.

[23] K. Starbird. Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on twitter. In *ICWSM*, 2017.

[24] K. Starbird, J. Maddock, M. Orand, P. Achterman, and R. M. Mason. Rumors, False Flags, and Digital Vigilantes: Misinformation on Twitter After the 2013 Boston Marathon Bombing. In *iConference*, 2014.

[25] O. Varol, E. Ferrara, C. A. Davis, F. Menczer, and A. Flammini. Online human-bot interactions: Detection, estimation, and characterization. In *ICWSM*, 2017.

[26] Z. Wang, W. Dong, W. Zhang, and C. W. Tan. Rumor source detection with multiple observations: Fundamental limits and algorithms. In *ACM SIGMETRICS Performance Evaluation Review*, volume 42, 2014.

[27] Wikipedia. Pizzagate conspiracy theory. https://en.wikipedia.org/wiki/Pizzagate_conspiracy_theory, 2017.