A Multi-Platform Analysis of Political News Discussion and Sharing on Web Communities

Yuping Wang[1], Savvas Zannettou[2], Jeremy Blackburn[3], Barry Bradlyn[4], Emiliano De Cristofaro[5], and Gianluca Stringhini[1]

[1]Boston University, [2]Max Planck Institute for Informatics, [3]Binghamton University, [4]University of Illinois at Urbana-Champaign, [5]University College London

– iDRAMA Lab, https://idrama.science –

arXiv:2103.03631v1 [cs.CY] 5 Mar 2021

Abstract

The news ecosystem has become increasingly complex, encompassing a wide range of sources with varying levels of trustworthiness, and with public commentary giving different spins to the same stories. In this paper, we present a multi-platform measurement of this ecosystem. We compile a list of 1,073 news websites and extract posts from four Web communities (Twitter, Reddit, 4chan, and Gab) that contain URLs from these sources. This yields a dataset of 38M posts containing 15M news URLs, spanning almost three years.

We study the data along several axes, assessing the trustworthiness of shared news, designing a method to group news articles into *stories*, analyzing these stories are discussed, and measuring the influence various Web communities have in that. Our analysis shows that different communities discuss different types of news, with polarized communities like Gab and /r/The_Donald subreddit disproportionately referencing untrustworthy sources. We also find that fringe communities often have a disproportionate influence on other platforms w.r.t. pushing narratives around certain news, for example about political elections, immigration, or foreign policy.

1 Introduction

The Web has facilitated the growth of fast-paced, online-first news sources. It has also allowed users to actively contribute to and shape the discussion around news. This creates an environment where journalists are not necessarily the arbiters of how a news story develops and spreads. In today's "hybrid" media system [15], the popularity of a news story is also influenced by how users discuss it. Although such discussions usually happen organically, various actors from polarized online communities or state-sponsored troll farms might also attempt to manipulate them, e.g., by pushing [20] or weaponizing [96] certain narratives.

Previous work studying the news ecosystem on the Web has mostly focused on Twitter, looking at single news articles [65, 73, 101], or on the discussion surrounding a small set of events [18, 79, 92]. Moreover, efforts to study the intertwined relationship between news coverage and social media discussions have been limited to direct quotes from news articles posted on Twitter [44], or on how Web communities influence each other in spreading *single* news URLs [96]. Overall, despite the crucial role played by online news in our society, we still lack computational tools to monitor how news stories unfold and are discussed across multiple online services.

In this paper, we present a longitudinal study of how news is disseminated across *multiple* Web communities. We introduce an analysis pipeline (which is independent of the data sources and thus reusable), consisting of different components to: 1) collect data, 2) extract named entities, 3) group articles belonging to the same story, and 4) estimate influence of a Web community on other ones.

We instantiate the pipeline by focusing on a mix of mainstream and fringe communities – namely, Twitter, Reddit, 4chan's Politically Incorrect board (/pol/), and Gab – and extract 38M posts including 15M news URLs, spanning a period of almost three years. We use named entity extraction to analyze what types of news these communities discuss. We also study the interplay between news discussion and the trustworthiness of the news sources cited using NewsGuard [58], a trustworthiness assessment site compiled by professional fact checkers.

Next, we perform community detection to group together related articles into news stories, and study how they are discussed on different Web communities. To this end, we use GDELT, a dataset that labels global news events [42]. Because of GDELT's focus on politics, our measurement also concentrates on political news stories. Finally, to study the influence that different Web communities have on each other in spreading news stories, we use Hawkes Processes [31]. These allow us to estimate which news stories are organically discussed on the Web, and for which ones certain communities exercise a significant influence in spreading them.

Our analysis yields the following main findings:

- When discussing news, different Web communities post URLs to news outlets with varying levels of trustworthiness. In particular, Gab and /r/The_Donald (subreddit) prefer untrustworthy ones compared to Reddit, Twitter, and 4chan.
- Some communities are particularly influential in the dissemination of news. While large ones like Twitter and Reddit have a consistent influence on the others, relatively small/fringe communities like /r/The_Donald have

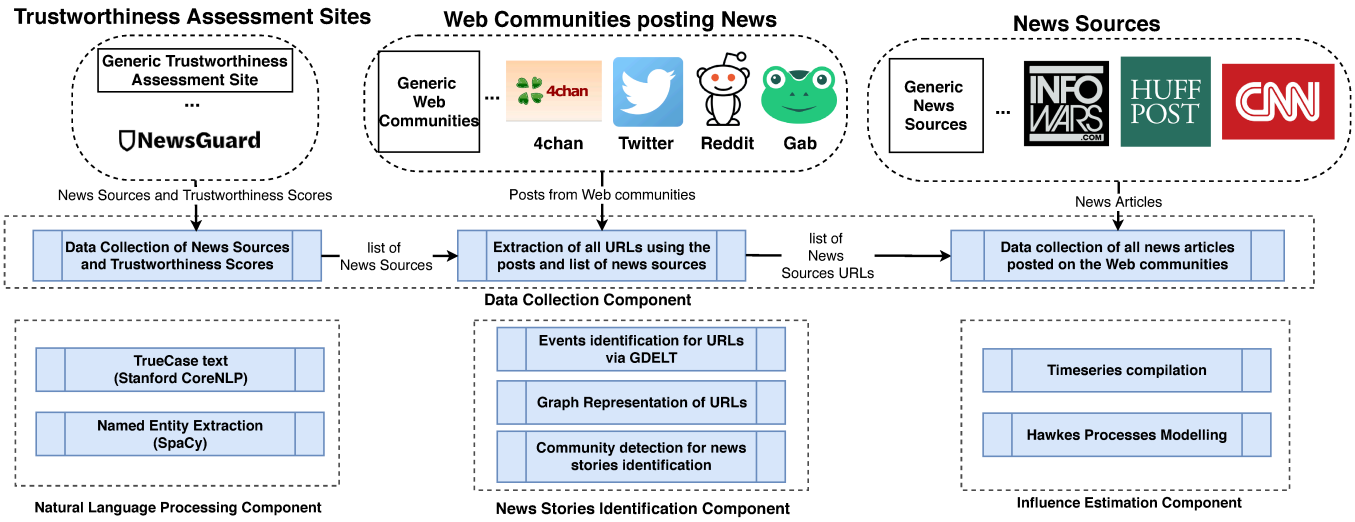


Figure 1: High-level overview of our processing pipeline.

a much larger external influence, affecting the posting of stories at a rate that is much larger than their relative size would suggest.

- Some topics are very popular across the board, however, different communities focus on different narratives about the same story. For instance, /r/The_Donald and /pol/ are particularly influential in spreading anti-immigration rhetoric and conspiracy theories.

2 Methodology & Datasets

In this section, we present the methodology used to analyze the appearance and discussion of news articles across multiple Web communities, as well as the datasets we collect in the process. Figure 1 presents a high-level overview of our pipeline, which consists of four components:

1. *Data Collection*: selecting news sources, collecting related posts from social media, and gathering the news content.
2. *Natural Language Processing*: identifying the "entities" mentioned in news articles and discussions in Web communities.
3. *News Stories Identification*: grouping articles belonging to the same story.
4. *Influence Estimation*: assessing the influence of each Web community on others with respect to news stories.

2.1 Data Collection

This component is used to select a set of suitable news sources, determine their trustworthiness, and retrieve posts on different social networks including URLs to these sources. While our methodology is general and can be used with any data source, in the following, we describe it along with the specific services selected for this work.

News Sources. Previous work has mostly relied on predetermined lists of websites [2, 13, 29, 96]. However, this incurs important limitations, as the popularity of news sites varies over time [72], and low-reputation sites are often ephemeral. For instance, out of the 54 "alternative" news sites studied in [96], only 23 are still active as of October 2020. Thus, we opt to take a more systematic approach. First, we gather popular domains using the top 30K websites from the Majestic list [48] as of February 2019. Then, we use the VirusTotal API [1], a service that provides domain categorization, to select only the domains categorized as *news and media* or *news*. Note that VirusTotal's categorization is not exempt from mistakes; e.g., domains like ananova.com, adbusters.org, and cagle.com are misclassified as news sites. To further refine this list, we use the NewsGuard API [58], a service that ranks news sources based on their trustworthiness, and restrict our analysis to news sites that are rated by NewsGuard as of February 2019, i.e., before it became a paid service. As a result, we obtain a total of 1,073 news websites.

Source Trustworthiness. We use NewsGuard [58] also to characterize the trustworthiness of a news Website, as it provides credibility/transparency scores. The scores are based on nine journalistic criteria, focused on different aspects (e.g., whether a news site consistently publishes false content, uses deceptive headlines, etc.) that do not take into account political leanings, and range from 0 to 100. If the score is no less than 60, the news source is labeled as trustworthy, and untrustworthy otherwise [56]. NewsGuard's evaluation is conducted by a team of experts [56], and this manual vetting provides us with reasonable confidence in its accuracy. The threshold of 60 is pre-defined by NewsGuard, and its assessment evaluation is designed to fix the threshold first and then assign the points for each criterion according to this threshold [56]. Therefore, we stick to this threshold.

Overall, we are confident in the meaningfulness of NewsGuard scores, as a recent Gallup study [23] has confirmed it as a useful tool to help readers identify untrustworthy news outlets. Approximately 600 news outlets in the US and Europe have refined their editorial practices to get higher scores [55]. Also, NewsGuard has been working with researchers [60, 67, 103], libraries, Web browsers, and service

Community	#Posts Trust.	#Posts Untrust.	#Unique URLs Trust.	#Unique URLs Untrust.
Twitter	7,123,715	686,497	3,893,357	291,354
Reddit	23,605,406	1,342,429	11,170,005	612,213
/r/The_Donald	528,142	190,742	385,384	122,204
4chan	458,431	75,705	275,422	37,472
Gab	2,369,149	2,265,336	749,547	385,317
Total	33,556,092	4,369,923	14,636,451	984,812

Table 1: Overview of our datasets. For each community, we report the number of posts that include URLs to trustworthy and untrustworthy news sources.

providers to increase the transparency of news credibility assessments [55]. News outlets are evaluated in a transparent way, as detailed information is published in the corresponding "nutrition label" page where readers can find the reasons for the judgment [57]. *NB: The full list of news sources used in this paper along with their NewsGuard scores is available, anonymously, from [4].*

Web Communities. We retrieve social media posts that include URLs from the 1,073 news sources in our dataset. Our selection is based on highlighting the interplay and the influence between different online communities instead of political leanings. While our pipeline can include any platforms, in this paper, we focus on a few Web communities: Twitter, Reddit, 4chan, and Gab. That is, we study both mainstream communities like Twitter as well as "fringe" ones like 4chan. In particular, we turn to 4chan's Politically Incorrect board (/pol/) as prior work shows it is an influential actor with respect to the dissemination of news [96] and memes [95]. We also include Gab because it is an emerging community marketed as a "free speech haven," which is heavily used by alt- and far-right users [94]. As for Reddit, we also choose to study /r/The_Donald as a separate community, since previous research has highlighted its influence in spreading information on the Web [22, 96].[1]

Table 1 provides a summary of the number of posts and unique news URLs for each community. Next, we describe the data that we collect for each Web community in detail.

Twitter. We collect tweets made available through the 1% Streaming API between Jan 1, 2016 and Oct 31, 2018. Note that, due to failure on our data collection infrastructure, we have some gaps in our dataset; specifically, 1) Dec 4-11, 2016; 2) Dec 25, 2016 to Jan 08, 2017; 3) Dec 17, 2017 to Jan 28, 2018; and 4) Sep 20 to Oct 31, 2018. We extract all tweets containing URLs to one of the news sources we study, collecting 7M tweets containing URLs to trustworthy news sources and 686K tweets containing URLs to untrustworthy news sources. The total number of unique (news) URLs is 3.9M and 291K for, respectively, trustworthy and untrustworthy news.

Reddit and /r/The_Donald. For Reddit, we use the monthly dumps available from Pushshift [7]. We collect all submissions and comments from Jan 1, 2016 to Oct 31, 2018, and extract all submissions and comments that include a URL to the news sources in our data. For the whole Reddit dataset, we find 23M and 1.3M posts with trustworthy and untrustworthy news URLs, respectively, while for /r/The_Donald, 528K posts with trustworthy and 190K with untrustworthy news URLs. The number of unique URLs is 11.2M and 612K, respectively, for trustworthy and untrustworthy news for the entirety of Reddit and 385K and 122K for /r/The_Donald. Note that the Reddit dataset also includes /r/The_Donald, as we aim to study the dynamics of Reddit as a whole. Nevertheless, even though /r/The_Donald is a very small subset, and as such it has negligible effect on the analysis, we remove it from Reddit for our influence estimation presented in Section 4.2 to guarantee the quality of the results of the underlying statistical model.

4chan's /pol/. We obtain all posts on 4chan's Politically Incorrect board (/pol/) between Jun 30, 2016 and Oct 31, 2018 from [61]. We extract all posts containing URLs to one of the news sources, collecting 458K and 75K posts with URLs to trustworthy and untrustworthy news sources, respectively, for a total of 275K and 37K, respectively, unique URLs.

Gab. We use the data collection methodology presented in [94] to collect Gab posts from Aug 10, 2016, and Oct 31, 2018. Once again, we extract posts that include URLs to trustworthy and untrustworthy news sources, collecting 2.3M posts containing trustworthy news URLs and 2.2M posts containing untrustworthy news URLs.

In total, we extract 15.6M URLs, 14.6M pointing to trustworthy and 984K to untrustworthy news sources, posted on the five Web communities. Note that the Twitter and Reddit datasets start a few months earlier (January 2016) than Gab and 4chan. This is because the authors of [61] began collecting 4chan data in June 2016, and 4chan data is ephemeral, therefore it is not possible to retrieve older posts. Gab, on the other hand, was launched in August 2016.

News Content. Next, we collect the *content* of the 15.6M news articles using the Newspaper library for Python3 [59], which, given a URL, retrieves the text from an article. For sanity check, we have one author randomly select 20 URLs from the Web communities dataset, download the text of these articles, and then manually compare the text with the content on the Web page. For 18 articles, the library downloads the full content, while, for 2, the text is partially downloaded—specifically, one misses the first two paragraphs, and the other only has the first paragraph. This provides us with reasonable confidence of the effectiveness of Newspaper to download the text of online news articles. Although the text of a small number of articles might be downloaded partially, this has a limited effect on our analysis since this text is only used when performing named entity recognition. Since the first paragraph of an article usually provides the most important information about the covered story followed by details, referred to as "inverted pyramid"style [91], we expect the overall effect of this issue to be small. See Section 3.2 for more details.

Note that we are unable to retrieve the content of about 1.4M articles due to server-side problems (e.g., the article was deleted from the server or the server was down at that time) and about 1M articles because of paywalls. For the latter, manual

[1] Note that /r/The_Donald was banned by Reddit in 2020 [85]. However, our research was done before this ban.

inspection shows that paywalls typically trigger a set of standard sentences being displayed instead of the actual news content (e.g., "sign up for a new account and purchase a subscription"). Thus, we parse results to exclude articles containing these sentences. At the end, we gather the text of 13M articles, 12M from trustworthy sources and almost 1M from untrustworthy ones.

Ethical Considerations. Our datasets only include publicly available data and we do not interact with users. As such, this is not considered human subjects research by our IRB. Also, we follow standard ethical guidelines [69], encrypt data at rest, and make no attempt to de-anonymize users.

2.2 Natural Language Processing

We now describe the NLP component, which we use to extract meaningful named entities that are referenced both on news articles and on discussions on several Web communities. Our NLP component involves two models: 1) a true case model that predicts and converts text into its correct case (e.g., "donald trump is the president" is converted to "Donald Trump is the president"); and 2) a named entity detection model that extracts known named entities from text along with an associated category (i.e., whether the extracted entity is a person, an organization, etc.). The former is necessary since the latter is case sensitive.

True Case Model. We use *TrueCaseAnnotator* from the Stanford CoreNLP toolkit [49]. This converts the case of the text to match as it should appear in a well-edited format (e.g., "united states" becomes "United States"), using a discriminative model built on Conditional Random Fields (CRFs) [39].

Named Entity Extraction Model. To obtain named entities, we rely on the SpaCy library [77] and the *en_core_web_lg* model. We choose this model since it is trained on the largest available dataset. The named entity detection model leverages millions of Web entries consisting of blogs, news articles, and comments to detect and extract a wide variety of entities from text, ranging from people to countries and events (see [78] for a list of all the supported types of entities). The model relies on Convolutional Neural Networks (CNNs), trained on the OntoNotes dataset [90], as well as Glove vectors [62] trained on the Common Crawl dataset [17].

2.3 News Stories Identification

The news articles in our dataset cover various aspects ranging from politics to entertainment. Among all categories, we focus on politics because previous work showed that these stories are often discussed differently on different online communities [96] and are often used to spread disinformation narratives [79, 92, 98]. Therefore, we design a news stories identification component to group political news articles covering the same "story."

We use the definition by Marcelino et al. [51], whereby every news "story" is composed of several story "segments." In a nutshell, we perform three tasks: 1) we identify segments using the GDELT dataset [42]; 2) we build a graph where news articles are nodes and edges are common segments discussed in them; and 3) we perform community detection on the graph to identify articles that discuss the same story.

Event Identification with GDELT. In this study, we use events identified by GDELT [42] in an article as a news story "segment." GDELT is a dataset containing event information for articles (published between Oct 30, 2015 and Nov 3, 2018) covering political news stories. GDELT's focus on politics makes it the ideal candidate for our analysis pipeline. We find 31M unique news URLs belonging to the 1,073 domains that we study in the GDELT dataset, composed of 30M unique trustworthy news URLs and 712K unique untrustworthy news URLs. For each URL, GDELT lists *events* (e.g., "Egyptian Minister of Foreign Affairs Mohamed Orabi attended the summit yesterday" [42]), which are each assigned a globally unique identifier ("Event ID"). The event extraction is performed at the sentence level by an automated coding engine called TABARI [42], which identifies the actors involved in the event, the action performed, and where the event happened. The result of this is that two different sentences referring to the same event are given an identical event ID, and each sentence is an *event mention* of this event ID [24]. When identifying an event from a sentence, GDELT also gives a confidence score to the event mention, ranging from 10 to 100% (in 10% increments), representing how sure the system is that this sentence indeed corresponds to that event ID [24].

To extract the segments associated to the news articles in our dataset, we first look up which of the URLs in our dataset are present in GDELT after a number of pre-processing steps, such as expanding shortened URLs, removing the query string as well as the `www` prefix or slash suffixes from the URLs. Then, we extract the list of corresponding events for each matched URL. We find 3M URLs in the dataset, comprising 24.6M event mentions (i.e., story segments). As we mentioned, the reason why GDELT focuses on political news, and therefore does not cover all news in our dataset, which are often about other topics like sports or entertainment.

Graph Representation of Stories. After labeling news URLs with events that are relevant to them, we build a graph linking single articles with common events they cover. In other words, if two articles share one event, these two articles are "related" and the more events two articles share, the closer their content is. The graph is built as follows: 1) We treat each URL (stripped of its parameters) as a node. 2) We remove all event mentions for which GDELT has a confidence lower than 60%. We select the 60% threshold as a tradeoff between removing too many events and ensuring high precision in event identification. As discussed later, event mentions with low confidence are not reliable and keeping them in the graph ends up producing poor results; as a result we remove 18.2M out of the 24.6M event mentions. 3) If two URLs share at least one event, we build an edge between the two nodes. 4) The edge weight is computed as the number of unique event IDs that two URLs share.

Community Detection. Two URLs that share one or more events are not guaranteed to cover the same story. To further refine the association between common events and news stories, we apply community detection on the graph. We then

consider URLs to belong to the same news story if they are part of the same community. We apply the Louvain Method [11], which allows to efficiently find communities in sparse graphs like the one we are dealing with: our graph is composed of 3M nodes and 1.6M edges. First, though, we prune edges with weight lower than a threshold d, since, upon manual inspection, we find that the GDELT events include some noise, possibly due to crawler or event extraction faults. In the following, we discuss how we select the value of the threshold d for our experiments.

Selecting the Story Edge Weight Threshold. To select the threshold d, we first discard URLs with more than 60 unique event IDs (220 URLs in total). We find that these results are due to errors in GDELT's crawling process; when manually inspecting these URLs, we find that the content is mostly homepages of news outlets, including numerous headlines and therefore flagged with multiple spurious event IDs. Further, we remove communities whose URLs are from a single domain only, since by manually looking at the clusters, we find URLs within such a cluster are published on the same day and share several events even though their texts are totally different.

Then, to select the threshold d, we perform the following steps with $d \in \{1, 2, 3, 4\}$. We apply the Louvain Method, obtain the corresponding communities of URLs, and randomly select 20 communities with size larger than 10. These samples are independently inspected by two authors to determine if the articles in them belong to the same story. After comparing their results, the two authors agreed that all samples with both $d = 3$ and $d = 4$ have a precision (i.e., the number of correctly grouped articles vs. all articles in the community) above 90%. Since $d = 3$ produces more communities than $d = 4$ (43K vs 26K), we settle on $d = 3$.

To further verify the appropriateness of our parameter choice, we also run an experiment in which we keep the threshold at $d = 3$, but this time we do not prune events with confidence lower than 60%, keeping them in the graph. As a result of this experiment we obtain 115k communities in total, which is much higher than on the pruned graph (43K). However, upon manual inspection, we observe that the quality of the identified communities is not satisfactory in this case. (Upon checking a sample of 20 communities, only 13 of them had a precision higher than 90%).

Alternatives to GDELT. Note that we also tested alternatives to GDELT as external "ground truth." More specifically, we group articles based on the TF-IDF [50] of their text and DBSCAN clustering [19]. However, manual analysis reveals that the performance of these methods is substantially worse. An alternative approach would have been to use topic modeling, e.g., LDA [10]. However, these methods are most effective when modeling topics that are broader than fine-grained news stories, and are therefore less appropriate than our approach in this case. The reason is that features from LDA, as well as TF-IDF, are obtained at the *word* level. So keywords shared by two different stories can interfere with the clustering result while features from our method are obtained from the *sentence* level (i.e., events in the stories), which avoids this issue. One example is a pair of stories found in our result: "Donald Trump's call to punish flag burners caters to voters in his base" and "air conditioning company Carrier says it has deal with Trump to keep jobs in Indiana." Both stories are from Nov 29, 2016 (and thus cannot be distinguished by date), and their text share some key word candidates: donald, trump, president-elect, tweet, which makes it difficult for LDA and TF-IDF to distinguish between them.

2.4 Influence Estimation

After grouping articles into news stories, we are interested in studying the temporal characteristics of how these stories are discussed in various communities of interest. More specifically, we aim to understand and measure the interplay between multiple Web communities with respect to the news stories they share. To do so, we create a timeseries that captures the cascades of each news story per Web community. After obtaining the timeseries, we model the interplay between Web communities using a statistical framework known as Hawkes Processes [31], which lets us quantify the influence that each Web community has on the others with respect to the dissemination of news stories.

Timeseries compilation. As a first step, we organize our data into timeseries. For each community of interest, we focus on the news stories that appear at least 100 times in our datasets (0.84% of all stories). This restriction helps to ensure the quality of the data under analysis. Next, for each news story i, on each Web community k, we build a timeseries $u_{ik}(t)$, whose value is the number of occurrences of news URL related to a specific news story per t hours.

Hawkes Processes are self-exciting temporal point processes [31] that describe how events (in our case, posts including news URLs pertaining to a news story) occur on a set of processes (in our case, Web communities). A Hawkes model consists of K point processes, each with a *background rate* of events $\lambda_{0,ik}$. Note that the events considered for Hawkes processes are a set of posts made on Web communities, and do not have to be confused with the event IDs from the GDELT dataset, that we used to identify the news stories. For us, the point processes will be the timeseries $\{u_{ik}|k = 1, \ldots, K\}$ for a given story i. The background rate is the expected rate at which events referring to a story will occur on a process *without* influence from the processes modeled or previous events; this captures stories mentioned for the first time, or those seen on a process we do not model and then occur on a process we do model. An event on one process can cause an *impulse response* on other processes, which increases the probability of an event occurring above the processes' background rates. The shape of the impulse response determines how the probability of these events occurring is distributed over time. Hawkes Processes are used for various tasks like modeling the influence of specific accounts [3, 97, 98], quantifying the influence between Web communities [95, 96, 102], and modeling information diffusion [30, 34, 47, 75]. Here, we use them to quantify the influence between multiple Web communities with respect to the dissemination of news.

Model fitting. We assume a Hawkes model that is fully con-

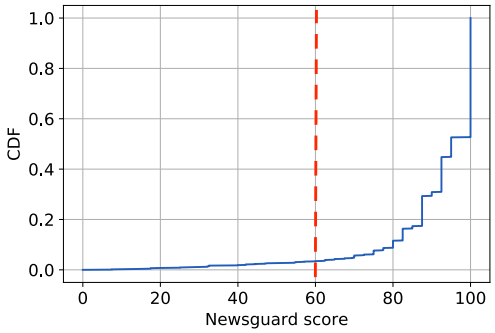


Figure 2: CDF of the NewsGuard scores of the news sources.

nected: each process can influence all the others, as well as itself, which describes behavior where a user on a Web community sees a news story and re-posts it on the same platform. Fitting a Hawkes model to a series of events on a set of processes provides us with the values for the background rates for each process, along with the probability of an event on one process causing events on other processes. Background rates also let us account for the probability of an event caused by external sources of information—i.e., a Web community that we do not model. Thus, while we only model the influence for a limited number of Web communities, the resulting probabilities are affirmatively attributable to each of them as the influence of the greater Web is captured by the background rates. This also helps addressing limitations of our datasets. In particular, the tweets that are not included in the Twitter 1% Streaming API is absorbed into the background rate of each community, avoiding to erroneously attribute an event to a different community. To fit a Hawkes model for each news story, we use the approach described in [45, 46], which uses Gibbs sampling to infer the parameters of the model from the data, including the background rates and the shape of the impulse responses between the processes.

Influence. Overall, this enables us to capture the interplay between the posting of news stories across multiple Web communities and quantify the influence that each Web community has on each other. More precisely, we use two different metrics: 1) ***raw influence***, which can be interpreted as the percentage of news story appearances that are created on a *destination* Web community in response to previously occurring appearances on a *source* one; and 2) ***normalized influence*** (or efficiency), which normalizes the raw influence with respect to the number of news story appearances on the source Web community, hence denoting how efficient a community is in spreading news stories to other Web communities.

3 General Characterization

3.1 News Sources

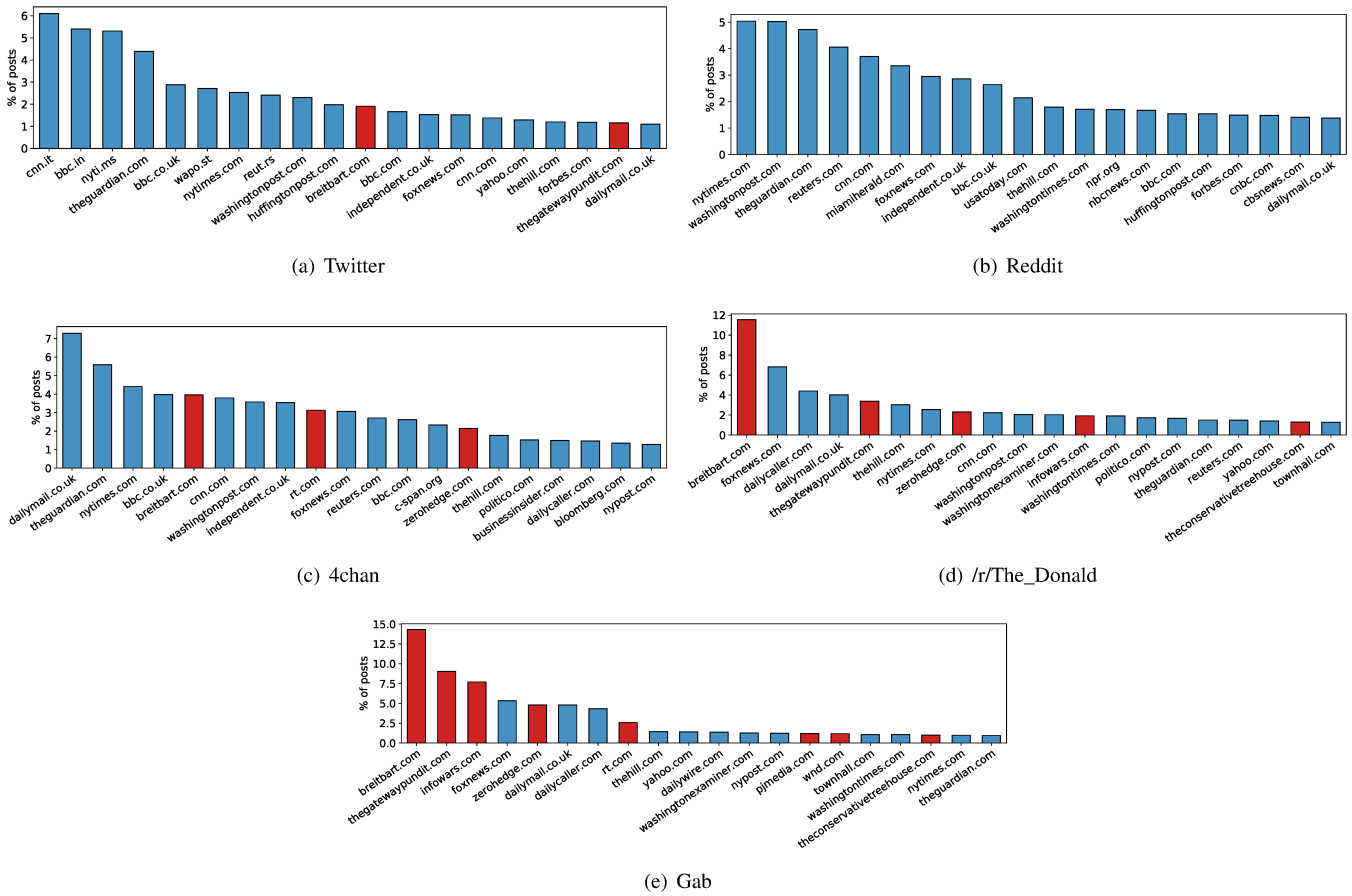
Using the NewsGuard scores, we find that out of the 1,073 news sources in our dataset, 1,036 (96.6%) are labeled as trustworthy (e.g., the New York Times, the Washington Post) and only 37 (3.4%) untrustworthy (e.g., Infowars, Breitbart), i.e., they have a score of less than 60/100. Figure 2 plots the CDF of the trustworthiness scores: most sources obtain relatively high scores, with 69% of outlets scoring above 90, and almost half (47%) receiving 100. However, out of the 14M URLs in our dataset, 996K are to untrustworthy and 13M trustworthy news sources. That is, over 7% of posted URLs are from untrustworthy news even though these only account for 3.4% of the sources. Recall that the threshold of 60 is pre-defined by NewsGuard and is used as a guideline by their experts to rank news organizations. The threshold value is an important factor when designers assign the points for each criterion. For instance, even if a news outlet meets all transparency criteria (e.g., clearly lists funders) but fails all credibility criteria (e.g., repeatedly published false content and does not properly publish retractions), it would still receive a NewsGuard score of 25 and be therefore considered untrustworthy. Any change of threshold need to reevaluate the points for each criterion at the same time, which is out the scope of this paper. For this reason, it would not make sense for us to select a different threshold in this study.

Trustworthy		**Untrustworthy**	
Entity	**(%, out of 12M)**	**Entity**	**(%, out of 920K)**
Trump	17.94%	Trump	27.52%
U.S.	15.53%	US	18.74%
American	11.15%	Donald Trump	18.30%
Donald Trump	11.08%	American	15.54%
United States	10.13%	U.S.	13.76%
Republican	9.03%	Russia	13.39%
Washington	8.44%	United States	13.14%
America	7.24%	America	11.54%
New York	6.87%	Russian	11.14%
Americans	6.75%	Obama	10.20%
Reuters	6.52%	Americans	9.77%
Congress	6.18%	Republican	9.50%
Republicans	6.11%	Washington	9.26%
Obama	5.98%	Democrats	8.81%
US	5.96%	Facebook	8.44%
Democratic	5.90%	Hillary Clinton	7.48%
Democrats	5.78%	Congress	7.46%
Russia	5.73%	Republicans	6.95%
Facebook	5.71%	Syria	6.73%
Twitter	5.53%	Twitter	6.66%
White House	5.42%	Clinton	6.47%

Table 2: Top 20 named entities extracted from news articles originating from trustworthy and untrustworthy sources.

3.2 Named Entities

Next, we describe the named entities extracted as per the methodology described in Section 2.2. Note that although GDELT does offer extracted entities in their metadata, we find that their labeling is not suitable for our purposes. More specifically, GDELT relies on two databases of public figures which were last updated in 2010 [82]. So, for example, "Trump" does not appear in any of the entity metadata. Instead, we use TrueCaseAnnotator, SpaCy, and *en_core_web_lg*. Next, we describe the named entities extracted from the news articles in our dataset, and then move to the one extracted from the posts on Web communities containing news URLs.



(a) Twitter

(b) Reddit

(c) 4chan

(d) /r/The_Donald

(e) Gab

Figure 3: Top 20 domains according to their popularity. Blue and red bars denote, respectively, trustworthy and untrustworthy news sources.

Articles. Using the article text and our pipeline's NLP component (Section 2.2), we extract the named entities for each article in our dataset. Table 2 reports the top 20 named entities extracted from both trustworthy and untrustworthy news articles. We observe that the most popular entities referenced in both trustworthy and untrustworthy news are related to US politics. For example, Donald Trump is referenced in 18% and 27% of the trustworthy and untrustworthy news articles, respectively. We also find that both trustworthy and untrustworthy news articles mention *several* entities; about 90% of them reference between 2 and 100 (the corresponding CDF plot is not included due to space limitations). Finally, we note some differences between the top entities of trustworthy and untrustworthy news articles: for instance, "Hillary Clinton" appears in the latter but not in the former.

Web Communities. We also study the named entities that appear in posts including news URLs. Note that these entities are *not* related to the text of the news article pointed by the URL, but rather the comment it was posted with. Table 3 reports the top 20 named entities detected for each of the five Web communities. Similar to the named entities detected on the news articles, most of the entities appearing in the posts are related to world events and politics, and in particular US politics. For instance, one of the most popular entities is "Trump," with 7.6%, 9%, 9.4%, 13.2%, and 2.9%, for Twitter, Reddit, /r/The_Donald, 4chan, and Gab, respectively.

There are also some interesting differences between top entities across the communities: e.g., on 4chan, several entities are Jewish and Israel related (Israel with 8%, Jews 5%, Jewish 4.8%, and Israeli with 3.7%), while, on /r/The_Donald and Gab, we find Islam-related entities ("Muslim" with 4.6% on /r/The_Donald and 1.1% on Gab).

3.3 News Domains in Web Communities

Next, we study the popularity of the news sources on each Web community. Overall, we find that 8.7%, 5.0%, 25.7%, 12.6%, and 48.7% of the total occurrences point to untrustworthy URLs for Twitter, Reddit, /r/The_Donald, 4chan, and Gab, respectively, while the rest point to trustworthy URLs. In Figure 3, we report the top 20 news sources, in terms of their appearance, on each community. Untrustworthy news sources are highlighted in red.

On Twitter (Figure 3(a)) and Reddit (Figure 3(b)), the most popular domains are mainstream, trustworthy news sources like The New York Times, The Washington Post, and CNN. On /r/The_Donald(Figure 3(d)), the most popular news source is Breitbart, which was considered an untrustworthy news source from NewsGuard at the time of our experiments (57 score). We also find other untrustworthy news sources, e.g., the Gateway Pundit (20 score), Zerohedge (0 score), and Infowars (25 score), among the top 12 most popular news sources. We observe this phenomenon to an even greater extent on Gab (Fig-

Twitter	(%, out of 7.8M)	Reddit	(%, out of 24M)	/r/The_Donald	(%, out of 712K)	4chan	(%, out of 521K)	Gab	(%, out of 4.6M)
Trump	7.67%	US	11.04%	Trump	9.46%	Trump	13.28%	Trump	2.94%
US	1.46%	Trump	9.03%	Clinton	8.70%	US	10.45%	Obama	2.08%
U.S.	1.26%	Russia	8.28%	Obama	8.05%	UK	8.36%	US	1.97%
Donald Trump	1.19%	Russian	6.27%	Hillary	7.29%	Israel	8.13%	FBI	1.85%
Russia	1.02%	U.S.	5.47%	US	6.66%	Russia	7.75%	Democrats	1.42%
Obama	1.02%	China	4.03%	CNN	6.50%	EU	6.62%	America	1.32%
Clinton	0.95%	Clinton	3.88%	FBI	5.09%	U.S.	5.76%	CNN	1.30%
GOP	0.91%	Obama	3.47%	Russia	5.01%	Russian	5.60%	Russia	1.16%
UK	0.79%	FBI	3.44%	Muslim	4.62%	Donald J	5.42%	U.S.	1.12%
CNN	0.68%	CNN	3.30%	Muslims	4.55%	TRUMPTV	5.39%	Muslim	1.10%
Hillary Clinton	0.66%	American	2.93%	Hillary Clinton	4.54%	American	5.19%	American	1.04%
China	0.64%	Democrats	2.82%	U.S.	4.36%	Jews	5.16%	UK	1.02%
America	0.63%	UK	2.72%	American	4.23%	Syria	5.04%	Russian	0.84%
Russian	0.60%	GOP	2.59%	America	4.16%	Jewish	4.89%	Hillary	0.75%
Hillary	0.57%	Republicans	2.45%	Islam	3.61%	China	4.66%	Americans	0.73%
FBI	0.56%	White House	2.37%	Soros	3.25%	Clinton	4.44%	Clinton	0.71%
Republicans	0.52%	America	2.35%	Democrats	3.17%	Brexit	4.19%	EU	0.70%
BBC News	0.51%	Putin	2.33%	Donald Trump	2.85%	Britain	3.78%	Google	0.70%
Democrats	0.51%	Syria	2.28%	Russian	2.81%	Obama	3.71%	Democrat	0.68%
Brexit	0.50%	Washington Post	2.23%	Facebook	2.50%	Israeli	3.70%	POTUS	0.68%

Table 3: Top 20 entities in posts that contain URLs to news articles.

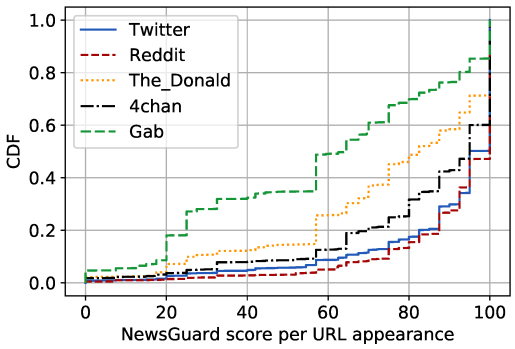


Figure 4: CDF of the NewsGuard score for each news URL.

ure 3(e)): the three most popular news sources are untrustworthy, with Breitbart being included in almost 15% of Gab posts with news URLs. For 4chan, we find mostly trustworthy news sources in the top 20, with the exception of Breitbart, the Russian state-sponsored RT (32.5 score), and Zerohedge. The most popular news source is actually the Daily Mail, which has a 64.5 NewsGuard score.

Overall, these figures show that /r/The_Donald and Gab are particularly polarized communities that extensively share news from untrustworthy sources, while Reddit and Twitter, which are more mainstream, do so to a much lesser extent. 4chan seems to be somewhat in the middle of the two: we find a substantial number of URLs from both trustworthy and untrustworthy news sources, which is perhaps surprising considering that 4chan is one of the most "extreme" communities on the Web [32].

We also study trustworthiness at the granularity of specific URL appearances. For each URL appearance, we extract the news source and assign its NewsGuard score. In Figure 4, we plot the resulting CDF. Note that Gab shares substantially more URLs from less trustworthy sources (with a median score of 64.5), followed by /r/The_Donald (median 82.5). 4chan shares substantially more URLs from trustworthy sources (with a median of 95) compared to Gab and /r/The_Donald, and its median matches the one from Twitter (95). Finally, Reddit users share predominantly URLs from trustworthy sources (with a median of 100). To confirm these observations, we perform χ^2 tests of independence on the proportion of trustworthy and untrustworthy news shared on each community. The results allow us to reject the null hypothesis that there is no difference in the rate of trustworthy and untrustworthy shared by Web communities ($p < 0.01$, with statistics values higher than 10,000 for all experiments).

3.4 Main Takeaways

Overall, different Web communities discuss different types of news; e.g., 4chan focus more than others on Jewish and Israel related news, and /r/The_Donald and Gab on news about Muslims. Also, users on /r/The_Donald and Gab prefer to cite untrustworthy news outlets to support their discussion.

We also perform a temporal analysis, aiming to capture the evolution of the trustworthiness scores as well as the interplay between trustworthy and untrustworthy news URLs on each platform. Although we do not include the details in the paper due to space limitations, we find that the use of untrustworthy news outlets on Gab and /r/The_Donald has been increasing over time, and on 4chan slightly decreasing, while remaining relatively stable on the other Web communities.

4 Analyzing News Stories

In this section, we set to understand how news stories, rather than single URLs, are discussed on different Web communities. First, we describe the news stories identified, focusing on

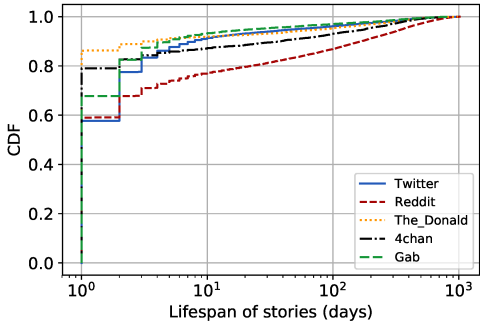


Figure 5: CDF of lifespan of stories on Web communities.

the differences across platforms. Then, we study the influence that different communities have on each other, using Hawkes Processes, and discuss a few interesting case studies.

4.1 News Stories

Out of the 14M news URLs we extract from Twitter, Reddit, /r/The_Donald, 4chan, and Gab, 3.2M of them appear in the GDELT dataset. After the story identification, we extract 43,312 unique stories: 21,878 on Twitter, 42,783 on Reddit, 4,943 on /r/The_Donald, 5,007 on 4chan, and 9,929 on Gab. These correspond to 109,153 unique URLs, 105,143 trustworthy and 4,010 untrustworthy. Trustworthy URLs occur 130,235 times on Twitter, 469,058 on Reddit, 9,249 on /r/The_Donald, 14,184 on 4chan, and 46,559 on Gab. Untrustworthy URLs occur 2,759 times on Twitter, 9,221 on Reddit, 990 on /r/The_Donald, 656 on 4chan, and 11,469 on Gab. Recall that GDELT is focused on political stories with a national and international relevance. As such, we expect news about local matters as well as sports or entertainment not to be included in this dataset.

Comparing this to the results in Section 3.3, we note that untrustworthy URLs appear in 2.1% of all news story posts on Twitter (compared to 8.7% for all news URL occurrences), while for Reddit this is 1.9% (compared to 5.0% overall), 4.4% for /r/The_Donald (26% overall), 4.4% for 4chan (13% overall), and 21% for Gab (49% overall). This might indicate that although some communities prefer to use untrustworthy news URLs to support their discussion, when discussing political news stories they still prefer to quote trustworthy ones. Our analysis in Section 4.3 suggests that this is sometimes done to give trustworthiness to a claim.

We then look at the *lifespan* of news stories on different Web communities, i.e., the time between the first and the last time a URL from a given story is posted on a platform. Figure 5 plots the CDF of the news story lifespans on the five Web communities. The vast majority of stories on all platforms are short-lived, with a small number of notable exceptions; for instance, the story with the longest lifespan on Twitter (984 days) is about "Saudi Arabia beheadings reaching the highest level in decades."

Overall, Reddit users discuss news stories the longest, as almost one out of ten (8.8%) last for 200 days or more. Interestingly, 4chan comes next, with 4.6% of the stories being discussed on that platform for 100 days or more. This is interesting, considering that posts on 4chan are ephemeral (i.e., they are deleted after a few days), with news content disappearing from the platform on a regular basis; hence, the fact that the 4chan community keeps discussing the same story for long periods of time indicates that new threads about it are constantly created.

Twitter	Reddit	/r/The_Donald	4chan	Gab	Total
24,987	65,610	1,926	3,163	11,800	107,486

Table 4: Number of events modeled via Hawkes Processes.

4.2 Influence Estimation

We now study the influence of Web communities w.r.t. news stories. As discussed in Section 2.4, we create a Hawkes model for each news story that appears at least 100 times on any platform, and more precisely 364 stories. Note that each model consists of five processes, one per community. Then, for each story, we fit a Hawkes model using Gibbs sampling.

Table 4 reports the overall number of events (i.e., appearances of news stories) modeled with Hawkes Processes for each Web community. Looking at the raw number of events, unsurprisingly, Reddit and Twitter are the communities with the most events in the selected 364 news stories.[2]

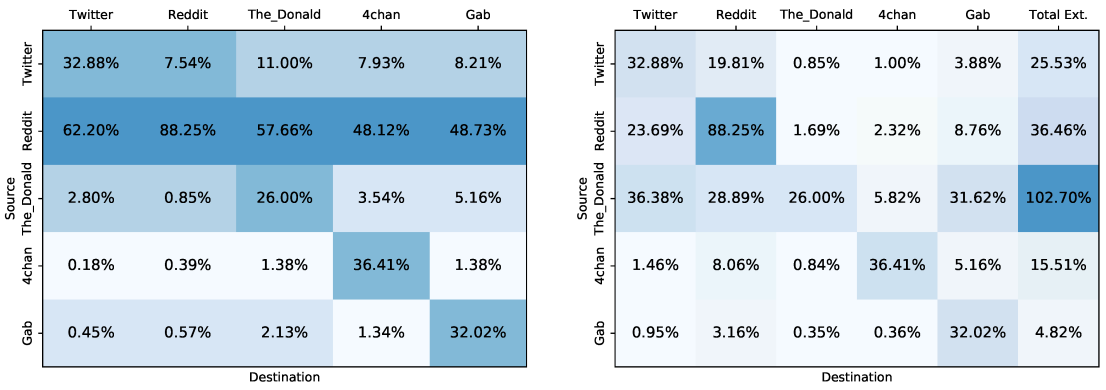
Fitting a Hawkes model provides us with the parameters for the background rates and impulse responses of each process, thus, we are able to quantify the influence that each Web community has on each other. Figure 6 shows our influence estimation results, which capture how influential and efficient Web communities are in spreading news stories.

Overall, we make several observations. First, in terms of raw influence (see Figure 6(a)), Twitter and Reddit are the most influential Web communities, mainly because of the large number of news story appearances that they produce. Moreover, out of the three smaller communities (4chan, Gab, and /r/The_Donald), /r/The_Donald is the most influential Web community for news stories that appear on Twitter and Reddit. This is particularly interesting since the overall number of news story appearances on /r/The_Donald is substantially smaller compared to the ones on 4chan and Gab (see Table 4). Finally, in terms of efficiency (see Figure 6(b)), /r/The_Donald is by far the most efficient Web community in making news stories appear on other Web communities.

The last column in Figure 6(b) is the sum of normalized influence from the specific source community to the rest of the platforms and as such can be over 100%. Generally, the bigger the percentage, the bigger the overall external influence from the source community to all the others; for instance, /r/The_Donald has a high influence on all the other platforms, which adds up to over 100%.

Note that statistical tests, e.g., to elicit confidence intervals

[2]Note that the start date of 4chan and Gab is behind the other Web communities (see Section 2). We find that 12.9% of the analyzed stories occur before the collection period for these platforms. This accounts 8.3k out of 110k (7.5%) events. Since only a minority of events are affected by this discrepancy, we consider the influence estimation experiments in this section to be an accurate reflection of real world trends.



(a) (b)

Figure 6: Influence estimation results: a) Raw influence between source and destination Web communities, which can be interpreted as the expected percentage of events created on the destination community because of previously occurring events on the source community; and b) Normalized influence (efficiency) of each Web community, which can be interpreted as the influence per news story appearance.

for the influence probabilities are hard to compute for Hawkes processes as, to the best of our knowledge, there is no statistical tool that is both meaningful and tractable. More specifically, goodness of fit for Hawkes process exists but it has not been implemented or tested at scale and therefore we leave it as part of future work.

4.3 Selected Case Studies

Since our influence estimation is done on a per-story basis, we can identify stories for which each community is the most influential. This enables us to identify a few case studies we believe to be particularly interesting. To do so, we calculate the overall external influence of each community and extract the top 20 externally influential stories (i.e., we do not include a community's self influence). Given that a number of stories are influential on multiple Web communities, we obtain a set of 88 unique stories, which are the ones where either Twitter, Reddit, /r/The_Donald, 4chan, or Gab is the most influential community on the other communities. Below, we present some case studies from these 88 stories.

Presidential election. Out of these 88 news stories, 37 are related to the 2016 US Presidential Election. Although the five Web communities under study all extensively discuss the election, we find that different communities push different narratives and topics onto other platforms. For instance, Twitter influences the other platforms to discuss a story related to then-House Majority Leader Kevin McCarthy believing that Vladimir Putin was paying President Trump (e.g., [52]), while Reddit does so with respect to President Trump revealing classified information to the Russian Foreign Minister (e.g., [16]). /r/The_Donald is influential in spreading a story suggesting that China hacked Hillary Clinton's email server (e.g., [83]), while 4chan for a story about an Iowa woman arrested on suspicion of voting twice (e.g., [87]). For Gab we find that the community was influential for a story reporting that Trump won the vote in Wisconsin and Pennsylvania after a recount (e.g., [53]).

Immigration. Over the past few years, refugee crises have been often covered in the news, as also confirmed by previous research [12]. In our dataset, we find 11 news stories for which one Web community has influenced its spread on other platforms. On Twitter, we find a story with articles about migrants stuck in US airports as a consequence of President Trump's immigration ban in 2017 (e.g., [86]). For 4chan, we find misrepresentations of remarks made by the Mexican government at the NAFTA summit, interpreted as an acceptance to pay for the border wall (e.g., [33]), and a story alleging that being too lenient in accepting refugees made Sweden the "rape capital of Europe" (e.g., [8]). Among the stories for which /r/The_Donald has influence on other platforms, there is one claiming that the influx of immigrants is the cause of the rise in violent crime in Germany (e.g., [9]).

Syrian conflict. Another interesting set of (six) stories is centered around the Syrian conflict. Again, the narratives differ greatly. Twitter pushes a story on French officials confirming a chemical attack carried out by the Syrian government (e.g. [68]), while /r/The_Donald reports that the US dropped 26,141 bombs over Syria during the Obama Administration (e.g., [84]). Gab has influence regarding a story on the US freezing funding to the White Helmets movement after false allegations of chemical attacks in Syria (e.g., [70]). This last example confirms the observations made by previous research on alternative narratives being assembled surrounding the White Helmets on social media [80].

4.4 Main Takeaways

In summary, we find that Twitter and Reddit are the most influential Web communities w.r.t. discussing news stories. However, /r/The_Donald is the most efficient considered its small(er) size. Our analysis also shows that different communities influence discussion on different stories. In particular, while Twitter and Reddit do so for major events reported by mainstream news along "neutral" narratives, more polarized communities are influential in discussing stories with specific narratives, from celebrating or criticizing political figures [53, 84], to promoting anti-immigration rhetoric [8, 9] or

distributing false news and conspiracy theories [33, 70, 83].

5 Related Work

News spread on social networks. Zhao et al. [101] compare news topics on Twitter to traditional media, while others [43, 96] investigate how news from mainstream and alternative news sources spread on different Web communities. Ratkiewicz et al. [65, 66] present a service that aims to track the spread of political astroturfing on Twitter. Vosoughi et al. [88] study the spread of true and false news on Twitter, finding that false news spread wider and faster than real news. Shao et al [73] analyze the role of bots in spreading false news on the Web, while Leskovec et al. [44] characterize news articles by identifying textual memes in them. Tan et al [81] build a four-layer structure model to study how information spreads. Zannettou et al. [99] study whether the appearance of news on Web communities like Reddit and 4chan affect the commenting activity of news articles with regards to hate speech. A comprehensive survey on this line of work is available from [35].

Other researchers have also focused on specific events, and in particular on disinformation. Wilson et al. [92] analyze comments on Twitter around the Aleppo boy conspiracy theory, while Backfried et al. [6] investigate attitudes towards refugees in Europe on German media. Starbird [79] studies how disinformation related to massive shooting events spreads on Twitter, and Conover et al. [18] characterize Twitter users' political orientations based on tweets related to the 2010 U.S. Congress midterm elections.

Finally, Zannettou et al. [96] look at the occurrence of *single* URLs. By contrast, we focus on organic discussion of news *stories* rather than single URLs, which lets us provide a comprehensive view of how news are discussed online. We also cover significantly more news outlets (1073 vs 99), and more Web communities (Gab was not included in [96]).

Overall, our work is the first, to the best of our knowledge, to study how different Web communities discuss political news stories and how they influence each other in doing so.

News credibility. Researchers have studied how to detect false news, focusing on the news story level, which is orthogonal to ours. Typically, they formulate the problem as a classification task and use machine learning to solve it [14, 74, 89, 93]. Another direction is to focus on the news outlet level; these two directions are often closely related. As shown in [21], the source of news plays a key role when people judge the authenticity of the story. Several papers, e.g, [40], attempt to determine what are the untrustworthy news outlets. Pennycook and Rand [63] assess the trustworthiness of a news outlet based on laymen's evaluations, showing that crowdsourced judgements are successful in assessing trustworthy news sources, although not as much as professional fact-checkers.

Gentzhow et al. [25, 26] show that news outlets can report news in a biased way, which could mislead the readers; in fact, Soraka et al. [76] demonstrate that people tend to be more "attracted" by negative news stories. To reduce bias, Babaei et al. [5] propose a method to identify "purple news," i.e., news that can be unanimously accepted by readers who have opposite political leanings.

Resnick et al. [67] propose a metric, called the "Iffy Quotient", to evaluate the spread of untrustworthy news sources on Twitter and Facebook, also relying on NewsGuard. In [13, 29], researchers collect false news outlets from various sources; Grinberg et al. [29] find that, among 2016 Presidential Election voters on Twitter, only a small portion are exposed to and share news from untrustworthy news outlets. Budak [13] compares the prevalence of news from trustworthy and untrustworthy outlets, obtained from [2], during the 2016 election campaign. Although news from trustworthy sources are shared the most, a growing number from untrustworthy outlets spreads over time.

Overall, our work differs from this line of work not only in terms of methodology, but also because, besides Twitter, we also study fringe, impactful communities like Gab and 4chan. Moreover, using Hawkes processes, we are able to analyze the influence of fringe communities on mainstream ones, which may help to better understand the influence dynamics of false news sharing.

Events recording databases. Our work relies on NewsGuard [58] to assess trustworthiness, GDELT [24] to find story events, etc. Other sources in this context include the News API [54] and Google news [28]. Event Registry [41, 71] is another service that aggregates news and provides insights to its users. Kwak and An [38] compare GDELT to Event Registry, showing that the former contains a larger set of articles and is therefore more suitable for research. Overall, GDELT has been extensively used by researchers to study topics related to refugees [12], protests [100], the media landscape [64], objects in news pictures [37], and so on [27, 36].

6 Conclusion

This paper analyzed the sharing and the spreading of online news. We showed that different communities present fundamental differences; for instance, Gab and /r/The_Donald "prefer" untrustworthy news sources (e.g., on Gab, 48.7% of all news URLs are from untrustworthy sources, compared to the 8.7% for Twitter). We also found that smaller Web communities can appreciably influence the news discussion on larger ones, with /r/The_Donald being very effective in pushing news stories on Twitter and the rest of Reddit.

Naturally, our work is not without limitations. First of all, while we did our best to gather a view of online news discussion that was as comprehensive as possible, our dataset of news websites only includes English news websites as identified by the Majestic list and NewsGuard. Moreover, we focused our analysis to four social networks, i.e., leaving out others like Facebook, due to the difficulty of collecting data. Finally, we relied on the GDELT dataset, which, as discussed, presented noise and crawling errors and on a named entity recognition model that is mostly trained on well-edited text like news articles. However, as discussed, we took several steps to mitigate these issues, by performing a sensitivity analysis that allowed us to build accurate communities of news articles to form the news stories that we analyzed.

Overall, our analysis builds on a novel, re-usable computational pipeline relying on tools from natural language processing, graph analysis, and statistics. As such, our approach to group related news together, track their discussion on multiple networks, and assess influence between Web communities in discussing them could serve as the foundation for a wealth of research not only in computer science, but also in journalism and political science.

As part of future work, we plan to expand our methodology to look at what language is used to discuss the same news story on different Web communities, and at whether or not using certain types of language (e.g., hate speech) has a particular influence in news discussion.

References

[1] VirusTotal API. https://www.virustotal.com, 2019.

[2] H. Allcott, M. Gentzkow, and C. Yu. Trends in the diffusion of misinformation on social media. *Research & Politics*, 6(2), 2019.

[3] H. Alvari and P. Shakarian. Hawkes Process for Understanding the Influence of Pathogenic Social Media Accounts. In *arXiv:1902.01970*, 2019.

[4] Anonymous. List of news sources along with Newsguard credibility score. https://drive.google.com/file/d/1Iysd5uYv9GZ5e0dIV590x8nQabR49Asc/view, 2019.

[5] M. Babaei, J. Kulshrestha, A. Chakraborty, F. Benevenuto, K. P. Gummadi, and A. Weller. Purple feed: Identifying high consensus news posts on social media. In *AIES*, 2018.

[6] G. Backfried and G. Shalunts. Sentiment analysis of media in german on the refugee crisis in europe. In *International Conference on Information Systems for Crisis Response and Management in Mediterranean Countries*, 2016.

[7] J. Baumgartner, S. Zannettou, B. Keegan, M. Squire, and J. Blackburn. The Pushshift Reddit Dataset. In *ICWSM*, 2020.

[8] BBC. Reality Check: Is Malmo the 'rape capital' of Europe? https://www.bbc.com/news/uk-politics-39056786, 2017.

[9] BBC. Germany: Migrants 'may have fuelled violent crime rise'. https://www.bbc.com/news/world-europe-42557828, 2018.

[10] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 2003.

[11] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), 2008.

[12] E. Boudemagh and I. Moise. News media coverage of refugees in 2016: A gdelt case study. In *ICWSM*, 2017.

[13] C. Budak. What happened? the spread of fake news publisher content during the 2016 us presidential election. In *The WebConf*, 2019.

[14] C. Castillo, M. Mendoza, and B. Poblete. Information credibility on twitter. In *WWW*, 2011.

[15] A. Chadwick. The hybrid media system. In *ECPR*, 2011.

[16] Chicago Tribune. Trump revealed highly classified information to Russian diplomats, U.S. officials say. http://www.chicagotribune.com/news/nationworld/ct-trump-revealed-classified-information-russians-20170515-story.html, 2016.

[17] Common Crawl Repository. http://commoncrawl.org/, 2019.

[18] M. D. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, F. Menczer, and A. Flammini. Political polarization on twitter. In *ICWSM*, 2011.

[19] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et al. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *KDD*, 1996.

[20] E. Ferrara. Disinformation and social bot operations in the run up to the 2017 French presidential election. *First Monday*, 22(8), 2017.

[21] M. Flintham, C. Karner, K. Bachour, H. Creswick, N. Gupta, and S. Moran. Falling for fake news: investigating the consumption of news via social media. In *ACM CHI*, 2018.

[22] C. I. Flores-Saviaga, B. C. Keegan, and S. Savage. Mobilizing the trump train: Understanding collective action in a political trolling community. In *ICWSM*, 2018.

[23] Gallup. NewsGuard's Online Source Rating Tool: User Experience. https://www.newsguardtech.com/wp-content/uploads/2019/01/Gallup-NewsGuards-Online-Source-Rating-Tool-User-Experience-1.pdf, 2019.

[24] GDELT. The GDELT Event Database Data Format Codebook V2.0. http://data.gdeltproject.org/documentation/GDELT-Event_Codebook-V2.0.pdf, 2015.

[25] M. Gentzkow and J. M. Shapiro. Media bias and reputation. *Journal of Political Economy*, 114(2), 2006.

[26] M. Gentzkow, J. M. Shapiro, and D. F. Stone. Media bias in the marketplace: Theory. In *Handbook of media economics*, volume 1. 2015.

[27] K. S. Gleditsch, N. W. Metternich, and A. Ruggeri. Data and progress in peace and conflict research. *Journal of Peace Research*, 51(2), 2014.

[28] Google. Google News. https://news.google.com/, 2019.

[29] N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, and D. Lazer. Fake news on twitter during the 2016 us presidential election. *Science*, 363(6425), 2019.

[30] F. Guo, C. Blundell, H. Wallach, and K. Heller. The bayesian echo chamber: Modeling social influence via linguistic accommodation. In *Artificial Intelligence and Statistics*, 2015.

[31] A. G. Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 1971.

[32] 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.

[33] Infowars. Mexico Agrees to Pay for Wall. https://www.infowars.com/mexico-agrees-to-pay-for-wall/, 2018.

[34] Q. Kong. Linking Epidemic Models and Hawkes Point Processes for Modeling Information Diffusion. In *WSDM*, 2019.

[35] S. Kumar and N. Shah. False information on web and social media: A survey. In *arXiv:1804.08559*, 2018.

[36] H. Kwak and J. An. A first look at global news coverage of disasters by using the gdelt dataset. In *International Conference on Social Informatics*, 2014.

[37] H. Kwak and J. An. Revealing the hidden patterns of news photos: Analysis of millions of news photos using gdelt and deep learning-based vision apis. In *arXiv:1603.04531*, 2016.

[38] H. Kwak and J. An. Two tales of the world: Comparison of widely used world news datasets gdelt and eventregistry. In

ICWSM, 2016.

[39] J. Lafferty, A. McCallum, and F. C. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, 2001.

[40] D. M. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, et al. The science of fake news. *Science*, 359(6380), 2018.

[41] G. Leban, B. Fortuna, J. Brank, and M. Grobelnik. Event registry: learning about world events from news. In *WWW*, 2014.

[42] K. Leetaru and P. A. Schrodt. Gdelt: Global data on events, location, and tone, 1979-2012. In *ISA Annual Convention*, 2013.

[43] K. Lerman and R. Ghosh. Information contagion: An empirical study of the spread of news on digg and twitter social networks. In *ICWSM*, 2010.

[44] J. Leskovec, L. Backstrom, and J. M. Kleinberg. Meme-tracking and the Dynamics of the News Cycle. In *KDD*, 2009.

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

[46] S. W. Linderman and R. P. Adams. Scalable Bayesian Inference for Excitatory Point Process Networks. In *arXiv:1507.03228*, 2015.

[47] M. Lukasik, P. Srijith, D. Vu, K. Bontcheva, A. Zubiaga, and T. Cohn. Hawkes processes for continuous time sequence classification: an application to rumour stance classification in twitter. In *ACL*, 2016.

[48] Majestic. The Majestic Million List. https://majestic.com/reports/majestic-million, 2019.

[49] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky. The Stanford CoreNLP natural language processing toolkit. In *ACL*, 2014.

[50] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*. 2008.

[51] G. Marcelino, D. Semedo, A. Mourão, S. Blasi, M. Mrak, and J. Magalhaes. A benchmark of visual storytelling in social media. In *ICMR*, 2019.

[52] McClatchy DC Bureau. House majority leader told his colleagues in 2016: 'I think Putin pays' Trump. https://www.mcclatchydc.com/news/politics-government/article151133157.html, 2017.

[53] New York Post. Trump's win in Wisconsin confirmed after vote recount. https://nypost.com/2016/12/12/trumps-win-in-wisconsin-confirmed-after-vote-recount/, 2017.

[54] News API. https://newsapi.org/, 2019.

[55] NewsGuard. Inside NewsGuard's First Year Fighting Misinformation. https://www.newsguardtech.com/press/newsguards-first-year/, 2019.

[56] NewsGuard. Rating Process and Criteria. https://www.newsguardtech.com/ratings/rating-process-criteria/, 2019.

[57] NewsGuard. Sample nutrition labels. https://www.newsguardtech.com/ratings/sample-nutrition-labels/, 2019.

[58] NewsGuard. The Internet Trust Tool. https://www.newsguardtech.com/, 2019.

[59] Newspaper3k. Article scraping & curation. https://newspaper.readthedocs.io/en/latest/, 2013.

[60] J. Nørregaard, B. D. Horne, and S. Adalı. Nela-gt-2018: A large multi-labelled news dataset for the study of misinformation in news articles. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 630–638, 2019.

[61] A. Papasavva, S. Zannettou, E. De Cristofaro, G. Stringhini, and J. Blackburn. Raiders of the Lost Kek: 3.5 Years of Augmented 4chan Posts from the Politically Incorrect Board. In *ICWSM*, 2020.

[62] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014.

[63] G. Pennycook and D. G. Rand. Fighting misinformation on social media using crowdsourced judgments of news source quality. *PNAS*, 116(7), 2019.

[64] J. Rappaz, D. Bourgeois, and K. Aberer. A dynamic embedding model of the media landscape. In *The WebConf*, 2019.

[65] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, S. Patil, A. Flammini, and F. Menczer. Truthy: mapping the spread of astroturf in microblog streams. In *WWW Companion*, 2011.

[66] J. Ratkiewicz, M. Conover, M. R. Meiss, B. Goncalves, S. Patil, A. Flammini, and F. Menczer. Detecting and Tracking the Spread of Astroturf Memes in Microblog Streams. In *arXiv:1011.3768*, 2010.

[67] P. Resnick, A. Ovadya, and G. Gilchrist. Iffy quotient: A platform health metric for misinformation. https://csmr.umich.edu/wp-content/uploads/2018/10/UMSI-CSMR-Iffy-Quotient-Whitepaper-810084.pdf, 2018.

[68] Reuters. French declassified intelligence report on Syria gas attacks. https://www.reuters.com/article/us-mideast-crisis-syria-france-intellige/french-declassified-intelligence-report-on-syria-gas-attacks-idUSKBN1HL0N1, 2018.

[69] C. M. Rivers and B. L. Lewis. Ethical research standards in a world of big data. *F1000Research*, 2014.

[70] RT. US 'freezes funding' for White Helmets as group's Douma chem attack claim falls apart. https://www.rt.com/news/425810-white-helmets-us-funding-freeze/, 2017.

[71] J. Rupnik, A. Muhic, G. Leban, P. Skraba, B. Fortuna, and M. Grobelnik. News across languages-cross-lingual document similarity and event tracking. *Journal of Artificial Intelligence Research*, 55, 2016.

[72] Q. Scheitle, O. Hohlfeld, J. Gamba, J. Jelten, T. Zimmermann, S. D. Strowes, and N. Vallina-Rodriguez. A long way to the top: significance, structure, and stability of internet top lists. In *ACM IMC*, 2018.

[73] C. Shao, G. L. Ciampaglia, O. Varol, K.-C. Yang, A. Flammini, and F. Menczer. The spread of low-credibility content by social bots. *Nature communications*, 9(1), 2018.

[74] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1), 2017.

[75] S. Soni, S. L. Ramirez, and J. J. Eisenstein. Detecting Social Influence in Event Cascades by Comparing Discriminative Rankers. In *SIGKDD Workshop on Causal Discovery*, 2019.

[76] S. Soroka, P. Fournier, and L. Nir. Cross-national evidence of a negativity bias in psychophysiological reactions to news. *PNAS*, 116(38), 2019.

[77] spaCy. Industrial-Strength Natural Language Processing. https://spacy.io/, 2019.

[78] spaCy. Named Entity Recognition. https://spacy.io/api/annotation#named-entities, 2019.

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

[80] K. Starbird, A. Arif, T. Wilson, K. Van Koevering, K. Yefimova, and D. Scarnecchia. Ecosystem or echo-system? Exploring content sharing across alternative media domains. In *ICWSM*, 2018.

[81] C. Tan, A. Friggeri, and L. A. Adamic. Lost in propagation? unfolding news cycles from the source. In *ICWSM*, pages 378–387, 2016.

[82] The Computational Event Data System. Dictionaries. http://eventdata.parusanalytics.com/software.dir/dictionaries.html, 2014.

[83] The Daily Caller. SOURCES: China Hacked Hillary Clinton's Private Email Server. https://dailycaller.com/2018/08/27/china-hacked-clinton-server, 2016.

[84] The Guardian. America dropped 26,171 bombs in 2016. What a bloody end to Obama's reign. https://www.theguardian.com/commentisfree/2017/jan/09/america-dropped-26171-bombs-2016-obama-legacy, 2017.

[85] The New York Times. Reddit, acting against hate speech, bans 'the_donald' subreddit. https://www.nytimes.com/2020/06/29/technology/reddit-hate-speech.html.

[86] The New York Times. Judge Blocks Trump Order on Refugees Amid Chaos and Outcry Worldwide. https://www.nytimes.com/2017/01/28/us/refugees-detained-at-us-airports-prompting-legal-challenges-to-trumps-immigration-order.html, 2017.

[87] The Washington Post. Trump supporter charged with voting twice in Iowa. https://www.washingtonpost.com/news/post-nation/wp/2016/10/29/trump-supporter-charged-with-voting-twice-in-iowa/, 2016.

[88] S. Vosoughi, D. Roy, and S. Aral. The spread of true and false news online. *Science*, 359(6380), 2018.

[89] W. Y. Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *arXiv:1705.00648*, 2017.

[90] R. Weischedel, M. Palmer, M. Marcus, E. Hovy, L. R. Sameer Pradhan, N. Xue, A. Taylor, J. Kaufman, M. Franchini, M. El-Bachouti, R. Belvin, and A. Houston. OntoNotes Release 5.0. https://catalog.ldc.upenn.edu/LDC2013T19, 2019.

[91] Wikipedia. https://en.wikipedia.org/wiki/Inverted_pyramid_(journalism).

[92] T. Wilson, K. Zhou, and K. Starbird. Assembling strategic narratives: Information operations as collaborative work within an online community. In *CSCW*, 2018.

[93] L. Wu and H. Liu. Tracing fake-news footprints: Characterizing social media messages by how they propagate. In *WSDM*, 2018.

[94] S. Zannettou, B. Bradlyn, E. De Cristofaro, H. Kwak, M. Sirivianos, G. Stringini, and J. Blackburn. What is gab: A bastion of free speech or an alt-right echo chamber. In *The WebConf Companion*, 2018.

[95] S. Zannettou, T. Caulfield, J. Blackburn, E. De Cristofaro, M. Sirivianos, G. Stringhini, and G. Suarez-Tangil. On the origins of memes by means of fringe web communities. In *ACM IMC*, 2018.

[96] S. Zannettou, T. Caulfield, E. De Cristofaro, N. Kourtellis, I. Leontiadis, M. Sirivianos, G. Stringhini, and J. Blackburn. The Web Centipede: Understanding How Web Communities Influence Each Other Through the Lens of Mainstream and Alternative News Sources. In *ACM IMC*, 2017.

[97] S. Zannettou, T. Caulfield, E. De Cristofaro, M. Sirivianos, G. Stringhini, and J. Blackburn. Disinformation warfare: Understanding state-sponsored trolls on Twitter and their influence on the web. In *The WebConf Companion*, 2019.

[98] S. Zannettou, T. Caulfield, W. Setzer, M. Sirivianos, G. Stringhini, and J. Blackburn. Who Let The Trolls Out?: Towards Understanding State-Sponsored Trolls. In *WebSci*, 2019.

[99] S. Zannettou, M. Elsherief, E. Belding, N. Shirin, and G. Stringhini. Measuring and Characterizing Hate Speech on News Websites. In *WebSci*, 2020.

[100] H. Zhang and J. Pan. Casm: A deep-learning approach for identifying collective action events with text and image data from social media. *Sociological Methodology*, 49(1), 2019.

[101] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li. Comparing twitter and traditional media using topic models. In *European Conference on Information Retrieval*, 2011.

[102] K. Zhou, H. Zha, and L. Song. Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes. In *Artificial Intelligence and Statistics*, 2013.

[103] X. Zhou, A. Mulay, E. Ferrara, and R. Zafarani. Recovery: A multimodal repository for covid-19 news credibility research. *arXiv preprint arXiv:2006.05557*, 2020.