The Rise and Fall of Fake News sites: A Traffic Analysis

Manolis Chalkiadakis
FORTH/University of Crete
Greece

Alexandros Kornilakis
FORTH/University of Crete
Greece

Panagiotis Papadopoulos
Telefonica Research
Spain

Evangelos P. Markatos
FORTH/University of Crete
Greece

Nicolas Kourtellis
Telefonica Research
Spain

ABSTRACT

Over the past decade, we have witnessed the rise of misinformation on the Internet, with online users constantly falling victims of fake news. A multitude of past studies have analyzed fake news diffusion mechanics and detection and mitigation techniques. However, there are still open questions about their operational behavior such as: How old are fake news websites? Do they typically stay online for long periods of time? Do such websites synchronize with each other their up and down time? Do they share similar content through time? Which third-parties support their operations? How much user traffic do they attract, in comparison to mainstream or real news websites? In this paper, we perform a first of its kind investigation to answer such questions regarding the online presence of fake news websites and characterize their behavior in comparison to real news websites. Based on our findings, we build a content-agnostic ML classifier for automatic detection of fake news websites (i.e., F1 score up to 0.942 and AUC of ROC up to 0.976) that are not yet included in manually curated blacklists.

CCS CONCEPTS

• **Information systems** → *Traffic analysis*; **World Wide Web**; • **Social and professional topics**;

ACM Reference Format:
Manolis Chalkiadakis, Alexandros Kornilakis, Panagiotis Papadopoulos, Evangelos P. Markatos, and Nicolas Kourtellis. 2021. The Rise and Fall of Fake News sites: A Traffic Analysis. In *13th ACM Web Science Conference 2021 (WebSci '21), June 21–25, 2021, Virtual Event, United Kingdom.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3447535.3462510

1 INTRODUCTION

Falsehoods about US 2020 election fraud, QAnon conspiracy theories and COVID anti-vaccine lies have recently taken over the online news by storm. False information circulated on the Web includes news, stories or hoaxes created to deliberately misinform or deceive readers. Usually, these stories are created to (i) either lure users and become a profitable business for the publishers (i.e., clickbait) or (ii) influence people's views, push a political agenda or cause confusion. False information can deceive people when published on websites that have a look-and-feel of other trusted websites, or using similar names and Web addresses to reputable news organisations.

In a recent poll of 1,115 adults in US [24, 36], 83% said they were concerned about the spread of false information, when fewer than half were able to identify as false a QAnon conspiracy theory about pedophilic Satan worshippers trying to control politics and the media [17]. An analysis [18] found that on Facebook, the top 20 fake news stories about the 2016 U.S. presidential election received more engagement than the top 20 election stories from 19 major media outlets. According to a different study [10], US citizens rate fake news as a larger problem than racism, climate change, or terrorism. According to the study, more than making people believe false things, the rise of fake news is making it harder for people to recognize the truth, thus, making them especially conservative and less well-informed.

Considering the importance of this emerging threat to society, there is a significant body of research in the last years, aiming to analyze the content, the methodologies, the possible detection and mitigation techniques or the way fake news spread (e.g., [9, 15, 16, 38, 46, 47]). In fact, fake news can be categorised into [48]: (a) *Clickbait*: stories that are carefully fabricated to gain more website visits and drive advertising revenues for publishers. Such stories use sensationalist headlines to grab attention and increase Click-through rates normally at the expense of truth or accuracy. (b) *Propaganda*: stories that are created to deliberately mislead audiences, promote a biased point of view or particular political agenda. (c) *Sloppy Journalism*: stories with unreliable information or without verified facts which can mislead audiences. (d) *Misleading Headings*: stories that are not completely false but distorted using misleading or sensationalist headlines, in such a way that can spread quickly via social media sites where only headlines and small snippets of the full article are displayed on audience newsfeeds. (e) *Satire*: stories for entertainment and parody (e.g., the Onion, The Daily Mash, etc.).

In fact, fake news involved in the 2016 elections has received significant attention and well studied (e.g., [11]), although several follow a more generic approach of analysis [21] . There have also been works on the spread of fake news on social networks. For example, Shao et al. [37] studied the spread of fake news by social bots. Also, Fourney et al. [7] conducted a traffic analysis of websites known for publishing fake news in the months preceding the 2016 US presidential election.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

WebSci '21, June 21–25, 2021, Virtual Event, United Kingdom
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8330-1/21/06…$15.00
https://doi.org/10.1145/3447535.3462510

Regardless of the recent literature in this area, we still do not know much about the network characteristics of fake news distributing websites: What is the lifetime of these websites? What is the volume of traffic they receive and how engaged is their audience? How do they connect with other marked-as-fake news sites? All these questions would help us not only understand the network characteristics of the websites that deliver such content, but also extract important features that would help us detect and flag such websites in a content-agnostic way (i.e., without relying on their content itself). Such a detection strategy would (i) allow the society to surpass the language barriers of manual blacklist solutions and (ii) provide an automated way of fake news detection.

In this study, we take the first step towards this exact direction. We collect a dataset of 283 websites tagged from known fact-checking lists as delivering fake news and perform traffic and network analysis of such websites. In particular, and contrary to related work (e.g., [7]), we study and compare the user engagement of fake and real news sites by analyzing traffic-related metrics. Additionally, we explore the lifetime of fake news sites and their typical uptime periods over a time range of more than 20 years, and propose a methodology to detect websites that are synchronizing not only their uptime periods but also the content they serve during these periods. Based on our findings, we design a content-agnostic ML classifier for the automatic detection of fake news websites.
Contributions. In summary, this paper makes the following main contributions:

(i) We conduct the first of its kind temporal and traffic analysis of the network characteristics of fake new sites aiming to shed light on the user engagement, lifetime and operation of these special purpose websites. We compose an annotated dataset of 283 fake news sites indicating when such websites are alive or dormant, which we provide open sourced[1].
(ii) We propose a methodology to study how websites may be synchronizing their alive periods, and even serving the exact same content for months at a time. We detect numerous clusters of websites synchronizing their uptime and content for long periods of time within the USA election years 2016-2017.
(iii) We study the third-party websites embedded in different types of news sites (real and fake) and find that during the aforementioned election years there is a significant increase in the use of analytics in fake news sites but not an increase in the use of ad-related third-parties. Also, domains like *doubleblick*, *googleadservices* and *scorecardresearch* have higher presence in real than in fake news sites. On the contrary, *facebook* and *quantserve* have higher presence in fake news sites.
(iv) We build a novel, content-agnostic ML classifier for automatic detection of fake news websites that are not yet included in manually curated blacklists. We tested various supervised and unsupervised ML methods for our classifier which achieved F1 score up to 0.942 and AUC of ROC up to 0.976.

2 DATA COLLECTION

To perform this study, we collect data from different sources and in this section we describe in detail our data. First, we obtain lists with manually curated news sites, categorized as "fake" and "real". We then use the "fake" news sites list as input for crawling historical data from Wayback machine to annotate the state of each website. Finally, to explore the web traffic characteristics of the two categories of news sites and how their audience behaves, we collect data from SimilarWeb [26] and CheckPageRank [3].

2.1 Fake & Real News Sites Dataset

For this study, we compose two manually curated lists of news sites. One with sites that are marked as "fake" and one with sites marked as "real". For the fake news sites, we utilize the domains repository provided by the opensources.co website [25]. The repository contains 834 biased news sites, of which 283 domains are manually checked and flagged as "fake". This is a well-accepted list, and has been used in studies related to the fake news ecosystem of 2016 US elections [2, 12], as well as in fake news detection tools such as the *BS-Detector* [41]. Additionally, we compose a second list of same size, for "real" news sites by taking the top Alexa news sites (and ensuring that there are no sites there that are marked as fake in the *opensources* repository).

2.2 Web Traffic & Audience Behavior Analytics

To assess the user engagement in the websites of our dataset, we collect web traffic data from popular data services like SimilarWeb and CheckPageRank (date of crawl: June 2020). SimilarWeb provides Web and other traffic related data per website, while CheckPageRank provides search engine-related information. In summary, we analyze volume of user visits, where the user visits come from, their duration and what subdomains they browse, the number of users who bounce off a domain, as well as Web connectivity of websites with respect to number and type of incoming or outgoing links from and to other sites.

2.3 Historical Data & Annotation of State

Next, we focus on the 834 news sites flagged for spreading misinformation (i.e., marked as fake or biased) and to identify their different states across time, we collect historical data from the Wayback Machine [13]. Specifically, we first query the Wayback CDX server for each such news site in our list, and we get an index of the available timestamps for the particular domain. Then, we proceed with downloading the landing page of each timestamp and storing it locally for further processing. In total, we downloaded the content of these websites from the last 23 years.

The landing pages of each timestamp reflect the state at which the websites were in during that timestamp. The pages could contain material related to the domain crawled, or irrelevant content: if the domain name is not paid and is returned to the market for sale. To understand what is the state of each website for each timestamp, we attempt a manual annotation for each snapshot. By using the Puppeteer [33] framework, we iterate through the timestamped versions of these domains and render them on screen. In each case, a dialogue box is prompted, and we categorize the timestamped website as one of the following:

(i) **alive**: When a website is offering news content
(ii) **zombie**: When a website is offering content other than news (e.g., e-marketing or other news-irrelevant content).

[1] https://github.com/mxalk/fake$_n$ews$_r$esources

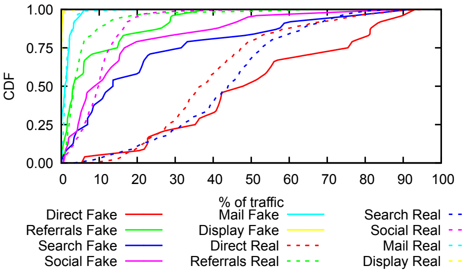

Figure 1: CDF of traffic sources for real and fake news sites. The median fake news site is accessed mostly directly. The median real news site is accessed mostly via search engines.

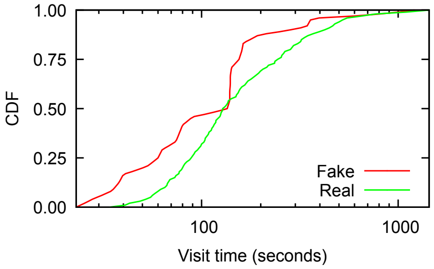

Figure 2: Distribution of visit duration in seconds, for real and fake news sites. In real news sites, the visit duration is longer than in fake news sites and follows a distribution close to power law.

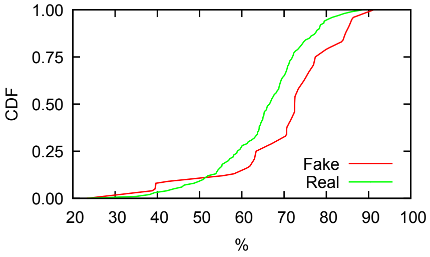

Figure 3: Distribution of bounce rate for real and fake news sites. The median real news site has a lower bounce rate (i.e., 66.55%), compared to the median fake news one (i.e., 72.49%).

(iii) **dead**: If the timestamped content of a website is none of the above (e.g., no HTML content was returned, or HTTP errors were returned), the website is declared "dead".

Given that a website may have been archived multiple times per month on Wayback, we aggregate the state of the website per month, by making the assumption that if it had at least one "alive" timestamp in said month, then it was "alive" for the entire month. Similarly, it was in a "zombie" state, if it had at least one such state in that month, or "dead", if none of the above applied. Finally, if there is a month that Wayback does not have a state for a website, that timestamp is marked as "missing" for said website.

3 USER ENGAGEMENT

As a first step, we set out to analyze and compare various user traffic-related metrics for the fake and real news sites in our lists, in an attempt to understand what is the different behavior of the audience in these two categories of websites. In particular, we focus on (1) where users come from to land on such websites, (2) how many pages they visit within a website, (3) their visit's duration and what sub-domains they browse, (4) number of users who bounce off a domain, and (5) Web connectivity of websites with respect to number and type of incoming or outgoing links to other sites.

Where do users come from? In Figure 1, we study the different sources that drive traffic to fake and real news sites. As we can see, the median fake site is being accessed mostly directly (user navigates directly to the website) by the users (*Direct*), or via links in *Social* media and search engines (*Search*). On the other hand, the median real news site is being accessed mostly via search engines (*Search*), and *Direct* follows. Sources such as *Mail* and *Display* drive similar traffic to fake and real news sites.

How many pages do users visit? In Table 1, we present the mean, standard deviation, median and 90th percentile of our user engagement metrics across all fake and real news sites. If we focus on the average number of pages per visit (in a time window of 6 months), we see that the median real news site tends to have a larger number of pages (i.e., 2.18 pages, on average) visited per user than the median fake news site (i.e., 1.72 pages, on average). Considering the 90th percentile, however, we see that there are fake news sites that have more (up to 3.54) pages visited on average than the corresponding real news sites (i.e., 3.49 pages visited, on average).

How long do users stay per visit? In Figure 2, we present the distribution of the average duration of the user visits per news site. This duration defines the time elapsed between the beginning of the first and the end of last page visit (sessions are considered closed after 30 minutes of user inactivity [42]). As we can observe, in real news sites, the visit duration is a heavy-tailed distribution, with many users visiting a website for up to a few hundred seconds, and very few users visiting for up to thousands of seconds. Additionally, as presented also in Table 1, visits last longer (i.e., 198.8 seconds) in real news sites, than in fake news sites (i.e., 163.4 seconds), with the visits of the 90th percentile lasting around 423.40 seconds in real news sites and 284.20 seconds in fake news sites, on average.

Website Bounce Rate. In Figure 3, we present the percentage of visitors who enter a site and then leave after visiting only the first page (also known as bounce rate). This metric is calculated by dividing the single-page sessions by all sessions [43], and reflects how well a site is doing at retaining its visitors. A very high bounce rate is generally a warning that people are not willing to stick around to explore the website, and instead they choose to leave. As we can see in the figure, the median real news site has a significantly lower bounce rate (i.e., 66.55%), compared to the median fake news one (i.e., 72.49%) with the corresponding rates for the 90th percentile being at 78.33% and 85.08%, respectively. We deduct that fake news sites probably provide content of lower quality, that is less engaging or interesting compared to the real news sites.

Website Backlinks & Referrals. In Figure 4, we plot the distribution of the number of backlinks for fake and real news sites. A backlink (also called citation or inbound/incoming link) of a website *A* is a link from some other website *B* (i.e., the referrer) to website *A* (i.e., the referent). As we can see in the figure, backlinks of fake news sites follow a heavy-tailed distribution, and they are significantly lower compared to the backlinks of real news sites. In particular, the median fake news site in our dataset scores 4.7K backlinks, but the median real news site scores 23.2M backlinks. The 90th percentile of fake news sites has 1.12M backlinks when the corresponding real news site scores as high as 53M backlinks. It

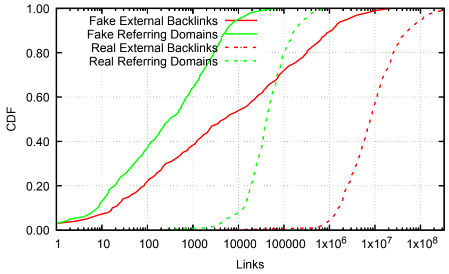

Figure 4: Distribution of backlinks and referring domains, for real and fake news sites. The median fake news site has significantly lower (4.7K) backlinks compared to real news (23.2M), and has ~2 orders lower (i.e., 307) referring domains than real news (i.e., 41.3K).

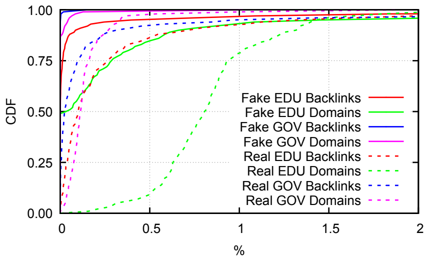

Figure 5: Distribution of EDU/GOV backlinks & referring domains over total, for real and fake news sites. Fake news sites have lower portions of EDU backlinks and referrals, as well as GOV backlinks than the real news sites.

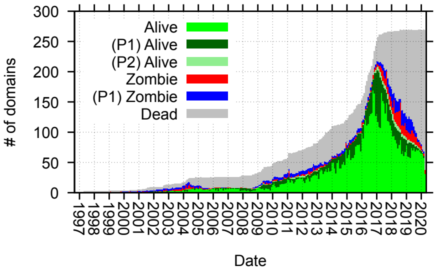

Figure 6: Histogram of the state of fake news sites for the last 20 years. There is a rise in fake news activity during the USA presidential election years 2016 and 2017. After 2017, we observe a sudden fall of "alive" or even "zombie" fake news websites.

Table 1: Summary of all user engagement metrics for the two categories of websites in our dataset, followed by two sample Kolmogorov-Smirnov statistics.

Metric	Type	Mean	St. Dev	Median	90th Perc	KS Statistic	p-value
Rank	Fake	197K	182K	130K	451K	0.62935	<0.00001
	Real	13.8K	16.6K	7.5K	33.7K		
Total Visits	Fake	3.3M	13M	349K	2.4M	0.10386	0.00412
	Real	44.5M	115M	9M	92.2M		
Pages per Visit	Fake	2.33	2.14	1.72	3.54	0.19857	<0.00001
	Real	2.49	1.65	2.18	3.49		
Visit Duration (sec)	Fake	163.4	234.3	135	284.2	0.04664	0.55113
	Real	198.8	190.7	128	423.4		
Bounce Rate	Fake	69.42%	15.81%	72.49%	85.08%	0.11044	0.00184
	Real	65.13%	11.04%	66.55%	78.33%		
Backlinks	Fake	519K	1.94M	4.7K	1.12M	0.94623	<0.00001
	Real	23.2M	4.73M	23.2M	53M		
Referring Domains	Fake	2.2K	5.6K	307	5.4K	0.95693	<0.00001
	Real	78K	106K	41.3K	168K		
Traffic Sources Direct	Fake	50	32	49	100	0.27726	<0.00001
	Real	41	16	39	64		
Traffic Sources Referrals	Fake	5	12	0	13	0.40130	<0.00001
	Real	4	6	2	9		
Traffic Sources Search	Fake	26	29	15	72	0.48964	<0.00001
	Real	42	15	41	63		
Traffic Sources Social	Fake	16	23	6	51	0.27289	<0.00001
	Real	10	6	9	17		
Traffic Sources Mail	Fake	0	2	0	2	0.55406	<0.00001
	Real	1	2	0	3		
Traffic Sources Display	Fake	0	4	0	0	0.35119	<0.00001
	Real	0	1	0	0		

is of no doubt that this difference is caused by the lack of trust that a large portion of websites show to fake news distributing websites.

Similarly, in Figure 4, we plot the distribution of the number of referring domains per website. When backlinks are the links on the websites that link back to a given site, a referring domain is where backlinks are coming from (e.g., think of the referring domain as a phone number and backlinks as the number of times you've gotten a call from that particular number). In median values, fake news sites have about 2 orders of magnitude lower (i.e., 307) number of referring domains than real news (i.e., 41.3K).

Finally, we study a particular class of domains or links from EDU or GOV domains, which could provide more authority and trust to a website when being referenced or linked to. In Figure 5, we plot the portion of backlinks and referring domains related to EDU/GOV domains for fake and real news sites. We see that fake news sites have clearly lower portions of EDU backlinks and referrals, as well as GOV backlinks than the real news sites.

4 LONGITUDINAL STUDY OF FAKE NEWS SITES

In this section, we focus on the fake news ecosystem and perform a historical analysis by studying the following questions: (1) What is the lifetime of a fake news site? (2) Are there any such websites that synchronize their uptime and reproduce the same content through time? (3) Which third-party trackers were persistently embedded in such websites through time?

4.1 What is the lifetime of a fake news site?

We use three terms to study the lifetime of a fake news site. First, we define with "lifespan" the upper limit for which a website may have existed on the Web. This is computed as the time difference from the first and last timestamp with "alive" state. Furthermore, the terms "alive time" and "zombie time" define the number of timestamps (e.g., months) that the website under study has been tagged as "alive" or "zombie", respectively. Consequently, the timestamps for which the Wayback does not provide any data are considered "dead".

During the lifetime of a website, various problems could arise, such as the owner not paying for the domain for some months, or the website being offline due to technical issues, etc.During such periods, and due to the crawling nature of the Wayback Machine, not all websites are archived at the same rate, and therefore, we may not have snapshots of websites for all timestamps studied. In an attempt to infer what the state of a website was in such un-archived or "missing" timestamps, we use a 2-phase interpolation process.

Phase 1. In the first phase ($P1$), we identify for each website any gaps between two timestamps with the same label A (A={$alive$, $zombie$}). Thus, when the two timestamps are non-consecutive, and there is no other labelled timestamp between them, we proceed with propagating label A to all "missing" timestamps of that gap. For example, if website W was found *alive* in timestamps i and j, with m timestamps in-between them (i.e., $j = i + m$), and no

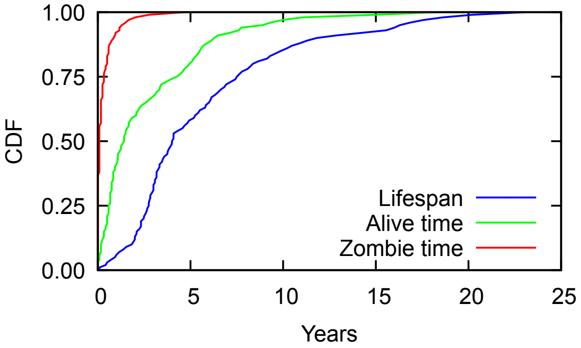

Figure 7: CDF of lifespan, alive and zombie time for the fake news sites. The median alive and zombie times of fake news sites are as low as 2 and 0.08 years, respectively.

other state was captured between i and j (i.e., during the m timestamps), then, we assume that W was alive for all the timestamps between i and j. Similar process was applied if A was *zombie*. This interpolation process can be applied for increasingly larger gaps, i.e., , for $m = 1, 2, \ldots$. Therefore, we applied it for increasingly larger m, and we stopped at 3-year gaps (i.e., m=36), since after that time there was no more correction done to the dataset.

Phase 2. In the second phase ($P2$), we identify gaps on the output of P1 between two "alive" timestamps up to three years apart, but allowing for up to 12 "non-alive" (i.e., "zombie" or "dead") timestamps between them. The "missing" timestamps between these two alive ones were also labelled as "alive".

In Figure 6, we present a histogram of the state of the fake news sites in our dataset through our examined period of more than 20 years. In the y-axis, we show the number of domains for each state. In the histogram, we also show the results from the interpolation phases $P1$ and $P2$, and how the plot smooths out, as expected. In this plot, at a given timestamp, we show the number of alive websites (including our interpolation phases), the number of zombie websites, when the remainder from our list is considered dead. Interestingly, we can see the rise in fake news activity during the USA presidential election years 2016 and 2017, and their sudden fall afterwards. During that fall, a portion of these websites turned into zombie state, while the great majority of them was shut down, especially in the last two years. This can happened either because they fulfilled their purpose (i.e., cause polarization or political bias [1]) or after being included in fake news lists and tools for blocking sources of misinformation.

In Figure 7, we plot the CDF of the lifespan, alive and zombie time for the fake news sites studied. The lifespan is the absolute maximum that such websites were found to exist on the web. As we can see, their lifespan is found to be about 4 years in median values, when on the other hand, alive and zombie times are lower, with median values of only 2 and 0.08 years, respectively.

4.2 Do fake news sites synchronize on their uptime and content?

As a next step, we set out to explore whether fake news sites appear to synchronize (i) on the times they are available on the Web, and (ii) the content they serve.

Uptime synchronization. To investigate the possible synchronization of their uptime, we assume that each website's sequence of alive or zombie states represents a binary time series and we focus on the last 5 years of the fake news activity (i.e., 2015-2020). To retrieve a cleaner signal and differentiate time series that synchronize across websites, we perform an aggregation at the quarter level (i.e., 3-month granularity) instead of monthly level. Thus, the final time series reflects quarters, and each one has 3 possible values (1, 2, or 3) for the number of months from the quarter that alive state was registered. Then, we compute measures of correlation between pairs of websites, using their quarterly-aggregated time series. For the comparison of time series, we use the pairwise euclidean distance for each pair of fake news sites. Since we perform quarter-time aggregation, this forces the euclidean distance analysis to perform similarly to more advanced methods such as Dynamic Time Warping.

Our method was able to identify several couples and a trio of websites with identical time series for the 2015-2020 time frame (i.e., euclidean distance of zero) categorized into the following types:

(1) [the-insider.co, ladylibertynews.com, amposts.com]: The websites were alive at the same time only for 1/12 quarters of their time series.
(2) [dailyinfobox.com, times.com.mx]: The websites were alive at the same time only for 1/12 quarters of their time series.
(3) [coed.com, rickwells.us]: The sites were alive for each of the 12 quarters of the time series.
(4) [dailyinfobox.com, times.com.mx]: These sites were alive for 2-5 quarters in total.
(5) [usapolitics24hrs.com, 24wpn.com]: These sites were alive for 2-5 quarters in total.
(6) [politicalo.com, religionlo.com]: These sites were alive for 2-5 quarters in total.
(7) [aurora-news.us, DonaldTrumpPotus45.com]: These sites were alive for 2-5 quarters in total.
(8) [washingtonpost.com.co, drudgereport.com.co]: These sites were alive for 2-5 quarters in total.
(9) [coed.com, rickwells.us]: These sites were alive for 2-5 quarters in total.
(10) [usaonlinepolitics.com, dailynewsposts.info]: These sites were alive for 2-5 quarters in total.

Content synchronization. To investigate how fake news sites may synchronize their content (in the same time window: 2015-2020), we developed a pipeline to compare pairs of fake news sites with respect to the content they publish. First, using the Beautiful Soup [35] library, we extract the text from each website[2]. After performing text pre-processing tasks on the extracted content (i.e., tokenization, removal of stop-words and lemmatization), we vectorize the documents using a typical TFIDF process [34]. Such vectors were created for each website and timestamp that it had content available. To compare these vectors, we use the cosine similarity metric [5] and we set a threshold of 0.5 to select pairs that appear to have high similarity. With this threshold, we ended up with 22

[2] There are more modern techniques for article content extraction available such as *newspaper3k* [27] or *readability.js*[23], but they are not applicable in our case because they are optimized with heuristics that extract content from full articles rather than landing pages.

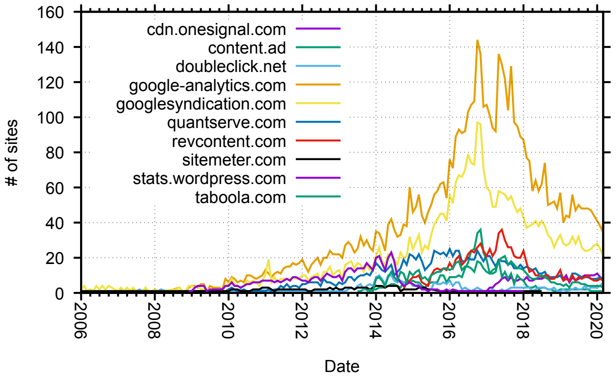

Figure 8: Top 10 third-party domains in fake news sites through time. During the peak of 2016-2017, we see an increase in use of analytics, but not a similar increase in use of ad-related third-parties.

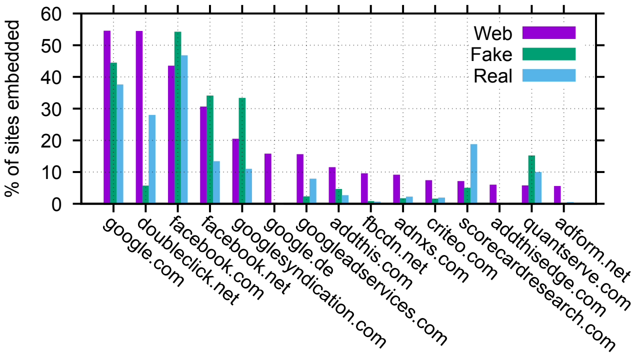

Figure 9: Top 15 third-parties on Web, ranked by *Whotracks.me*, and their respective coverage on fake and real news sites of our data.

distinct pairs of websites. Upon manually inspecting the pairs at their matched timestamps, we make the following observations regarding fake news site content synchronizations:

1. [newslo.com, religionlo.com, politicalo.com, politicops.com, politicot.com]. This group of websites consist of 5 domains, with all of them offering the exact same content. The longest period of synchronization was between [newslo.com, religionlo.com, politicalo.com], which lasted from 09/2015 to 04/2016. These 7 consecutive months of content synchronization were joined by the pair [politicops.com, politicot.com] in the last month.
2. [16wmpo.com, newsdaily12.com, local31news.com]. This was an active group between 07-11/2017.
3. [usatoday.com.co, washingtonpost.com.co, drudgereport.com.co]. This group was synchronized on 07/2015.

We observe that several of the pairs and even portions of the groups above overlap with the uptime synchronization study presented earlier. As a consequence, we believe our proposed methodology of studying synchronization of content and uptime of websites can enable a fake news detection process to select websites that have suspiciously high similarity in their uptime and content, for further examination and even blocking, if needed.

4.3 Which third-party trackers were embedded in fake news sites through time?

Contrary to most popular news sites that progressively move towards paywalling their high quality content [31], fake news sites rely on programmatic ads to make profit [28, 30]. Indeed, some of these websites were even created with the sole purpose of luring ad clicks by publishing clickbait content [4, 8].

To understand which third-party advertising entities provide tracking and other ad-related functionality to fake news sites, we study the third-party domains embedded in these sites through time. Specifically, we parse all collected HTML content per fake news site for each timestamp in our dataset and by using the AdblockPlus blacklist [6], we identify 55 such third-party domains in the HTML body of at least one website for one timestamp.

In Figure 8, we show the number of websites that the top 10 third-party domains were embedded in, per timestamp (i.e., month) in the list of fake news sites. These top 10 third-party domains were selected based on their cumulative appearance across fake news sites and across all timestamps. Evidently, analytics and ad entities dominate the top 10 list residing in 91.8% of all the fake news sites in our dataset. Interestingly, during the aforementioned peak of 2016-2017, we do see a significant increase in the use of analytics (i.e., *google-analytics*, and *googlesyndication*) but not an increase in the use of ad-related third-parties. This phenomenon shows that, the majority of the marked as fake news sites that were created within this time window (US pre-election period), had purposes beyond monetizing their published content (e.g., polarize, deliver misinformation, etc.).

Next, in Figure 9, we use the data provided by *Whotracks.me* [19] to compare the most embedded third-parties on the web, with the ones found in the fake and real news sites of our dataset for the period of 2016-2017. Interestingly, we see Google's *doubleclick* and *googleadservices* residing in less than 6% and 2% of the fake news sites, respectively, but they have presence in more than 27% and 8% of the real news sites of our dataset, respectively. Similarly, *scorecardresearch* (third biggest web beacons-based tracking service, owned by ComScore [44]) is present in less than 6% of the fake news sites, but in more than 19% of the real news sites of our dataset. On the other hand, *facebook* and *quantserve* (second biggest web beacons-based tracking service, owned by Quantcast [14]) are present in more marked-as-fake news sites than real ones.

5 AUTOMATIC FAKE NEWS DETECTION

Our earlier network traffic analysis of such websites revealed that it is possible for some of these features to be good at distinguishing the nature of the news website, such as number of visits, bounce rate, backlinks, etc. Thus, we were inspired to build an automated tool that performs the following tasks:

1. Retrieve network data for each website from common sources such as `similarweb` or `checkpagerank`
2. Preprocess the data and extract related features on network traffic activity

Table 2: Traffic-related features, used for classification.

Feature	Range of values for feature
Global rank	$[48, \ldots, 1.2M]$
Country rank	$[3, \ldots, 816K]$
Category rank	$[1, \ldots, 19K]$
Country (majority of traffic)	32 countries
Category	42 distinct web categories
Total visits	$[0, \ldots, 987M]$
Pages per visit	$[0, \ldots, 13.32]$
Bounce rate	$[0\%, \ldots, 95.37\%]$
Traffic source direct	$[2\%, \ldots, 100\%]$
Traffic source referrals	$[0\%, \ldots, 69\%]$
Traffic source search	$[0\%, \ldots, 87\%]$
Traffic source social	$[0\%, \ldots, 96\%]$
Traffic source mail	$[0\%, \ldots, 25\%]$
Traffic source display	$[0\%, \ldots, 95\%]$

Table 3: Performance metrics from ML binary classification of websites as showing fake or real news. Prec: Precision; Rec: Recall.

#	Method	TP Rate	FP Rate	Prec.	Rec.	F1	AUC
1	Random Forest (RF)	0.942	0.059	0.942	0.942	0.942	0.976
2	Logistic Regression (LR)	0.915	0.091	0.916	0.915	0.915	0.968
3	Naive Bayes (NB)	0.876	0.136	0.882	0.876	0.875	0.948
4	Neural Net (15x45x2)	0.911	0.092	0.911	0.911	0.911	0.955
5	RF on sites 0<Rank≤10K	0.929	0.378	0.924	0.929	0.923	0.925
6	RF on sites 10K<Rank<1.3M	0.917	0.096	0.917	0.917	0.917	0.970
7	RF #6 tested on websites of #5	0.923	0.174	0.932	0.923	0.926	0.942

(3) Apply a machine learning (ML) model that classifies the given website as serving fake or real news

In order to built this classifier, we performed a fresh crawl (February 2021) on our previously mentioned lists of fake and real news websites, and we trained and evaluated our envisioned classifier.

Based on previously mentioned 14 network traffic metrics or features (as summarized in Table 2) we train different ML classifiers for automatic classification of news websites as "real" or "fake". As a basic preprocessing step, we removed features with very little to zero variability. Our dataset for training and testing is fairly balanced, with "real" news websites being 278 and "fake" being 239. The difference from the previous numbers lies in the fact that we did extra steps for removing websites that did not have scores across all metrics. We applied 10-fold cross-validation on the available data, and trained and tested various ML standard techniques, including Random Forest, Logistic Regression, Naive Bayes, as well as more complex techniques such as Neural Network (NN), with three fully-connected layers: an input layer of 15 input variables, a dense layer of 45 neurons and an output layer for classification. We measured standard ML performance metrics such as True Positive and False Positive Rates, Precision and Recall, F1 score and Area Under the Receiver Operating Curve (AUC). The scores were weighted to take into account individual performance metrics per class weight.

Table 3 shows the results achieved with the aforementioned classifiers when all the dataset is used (upper part, classifiers #1-#4). Interestingly, we find that the typical Random Forest classifier performs very well across the board, with high True Positive and low False Positive rates, and higher Precision and Recall than the other ML methods, including the more complex NN approach.

Table 4: Evaluation of feature importance based on information gain with respect to class, for all websites, and when controlling for Rank (lower: 10K<Rank; higher: Rank≤10K).

Feature	Information Gain		
	All ranks	Rank≤10K	10K<Rank
Country rank	0.5199	0.0728	0.4723
Total visits	0.5027	0	0.4616
Global rank	0.4992	0	0.4616
Category rank	0.3754	0.0791	0.3066
Traffic source search	0.2918	0.2472	0.2881
Category	0.1757	0.0683	0.1873
Traffic source display	0.1642	0	0.1061
Country	0.1518	0.0842	0.1119
Traffic source mail	0.1267	0	0.1161
Traffic source social	0.0839	0	0.1219
Traffic source direct	0.0822	0.1273	0.1025
Traffic source refferals	0.0738	0	0.0693
Pages per visit	0.0298	0	0.0372
Bounce rate	0	0.0672	0

Given that the amount of traffic and other features used here are naturally correlated with each other (e.g., a highly ranked website should attract more visits, etc.), we also test the scenario where we split our dataset into two major groups of ranked websites (highly popular with Rank≤10K, and lower popularity websites, i.e., 10K<Rank), to check if the ML classification is still possible under similarly ranked websites. The results, shown in Table 3 (lower part, classifiers #5 and #6) demonstrate that it is possible to achieve very good performance, even when controlling for the rank of websites. This means that even if we focus the classification task on websites of similar ranking (low or high), the performance of the classifier is still high.

Furthermore, classifier row #7 checks the scenario where data from the lower ranked websites (i.e., 10K<Rank, where many fake news websites rank), are used to train a classifier that is then tested on data from higher ranked websites (i.e., Rank≤10K, where fewer fake news sides rank). Interestingly, the performance still remains high, showing that examples of fake news sites from lower ranking can be useful to distinguish such websites at higher ranks.

Finally, in Table 4, we investigate the importance of features as used in the classification effort. We evaluate the worth of a feature by measuring the information gain with respect to the binary class. We perform this evaluation for the three versions of the dataset: (1) all websites, irrespective of rank, (2) highly ranked websites, i.e., Rank≤10K, (3) lower ranked websites, i.e., 10K<Rank. We find that features expressing *ranking* are very important when all websites are considered. Other features such as *traffic source*-related metrics are less important. Interestingly, when top ranked websites are considered only, the features that are most important are *traffic source search* and *direct* and then 2/4 *ranking* features. The rest of features do not contribute to the model. On the other hand, when lower ranked websites are considered only, the order of importance of features is almost the same as when all websites are considered, with *ranking*-related features being most important, and some *traffic source*-related features being less important.

6 RELATED WORK

Numerous studies are attempting to explore the characteristics of fake news and its spread. In [7], authors conducted a traffic analysis to websites known for publishing fake news in the months preceding the 2016 US presidential election. Although the study also includes features as traffic sources and temporal trends, our work significantly diverges from that. We analyze a much greater set of websites, we compare fake with real news websites and we do not focus on social networks or the elections. In [45] authors perform a 3-year long study on fake news websites prior (2014-2016). Their analysis includes time-series modelling for causality testing. This is one of the few studies that are including time-series analysis. However, the method differs from ours in substantial manners.

As important as understanding fake news dynamics is, we cannot avoid mentioning the significant efforts to identify and flag fake news stories. Based on a survey [38], fake news can be identified with content-based, feedback-based and intervention-based methods. In [39] authors characterize detection as knowledge-based, stance-based, style-based and propagation-based. While there is a large number of publications in the Fake News detection area, we will only give specific examples. Check-it [32] is an ensemble method that combines different signals to generate a flag about a news article or social media post. The aforementioned signals are the domain name, linguistic features, reputation score and others. NELA [12] creates and combines credibility scores of the news article and the news source. Although it is possible to combine different methods to solve this problem, most papers focus on a single approach. For example, in a different publication, Shu [40] also argues about the role of social context for fake news detection.

In [48], authors provide a comprehensive overview of existing research on the false information ecosystem. In [9] authors show that fake news aim to affect the emotions of the readers, and ultimately deceive them. The authors create a neural network capable of detecting false news from such an effect. In [7], it is clear that aggregate voting patterns were strongly correlated with the average daily fraction of users visiting websites serving fake news. In [2], authors dive into the dynamics and influence of fake news on Twitter during the 2016 US presidential election. With a dataset of 30 million tweets and the opensources.co list, it finds that 25% of these tweets spread either fake or extremely biased news. Based on [46], the online social network ecosystems seem to interplay, and information shared in one network can affect the information flow in another network.

A lot of work has been done on Twitter disinformation. In [15], authors present a thorough analysis of rumour tweets from the followers of two presidential candidates. It is also shown by [47] that many trolls have been sponsored externally for ultimate goals. [37] proves that bots play a key role in the Twitter misinformation ecosystem, targeting influential users for misinformation spreading. As shown by [11], the vast majority of fake news is spread by an extremely small number of sources, forming clusters. This study sheds light on the target groups as well, being conservative-leaning, older and highly engaged with political news individuals. Following a different trajectory, [16] points to images as being very crucial content for the news verification process.

7 DISCUSSION

Summary. Our findings from the various measurements performed, can be summarized as follows:

- The median site serving real news tends to have a larger number of pages (i.e., 2.18 pages, on average) visited per user than the median fake news site (i.e., 1.72 pages, on average).
- On average, visits last longer (i.e., 198.8 seconds) in real news sites, than in fake news websites (i.e., 163.4 seconds).
- The median fake news site is being accessed mostly directly, when the corresponding median real news site is accessed via search engines.
- The median real news site has a significantly lower bounce rate, compared to the median fake news site.
- The median fake news site scores 4.7K backlinks, but the median real news site has more than 23.2M backlinks.
- Fake news sites have lower portions of EDU backlinks and referrals, as well as GOV backlinks and referrals, than the real news sites.
- Fake news sites have about 2 orders of magnitude lower number of referring domains than real news.
- The median alive and zombie times of fake news sites are as low as 2 and 0.08 years, respectively.
- There was a significant rise in fake news website birth and activity during the USA presidential election years 2016-2017, and there was a rapid fall afterwards.
- We detect numerous clusters of websites synchronizing their uptime and content for long periods of time.
- During this period, we see a significant increase in the use of analytics, but not an increase in the use of ad-related third-parties from the fake news sites.
- Domains like *doubleblick*, *googleadservices* and *scorecardresearch* tend to have higher presence in real news sites than in marked-as-fake ones. On the contrary, *facebook* and *quantserve* have higher presence in fake news sites.
- We show that it is possible to train supervised ML classifiers to detect fake from real news websites, achieving a very good performance: F1 score up to 0.942 and AUC of ROC up to 0.976.

Implications. Our traffic analysis revealed that websites serving real news are, on average, more engaging and the visitors stay longer in the website, visit more pages in each domain and are less likely to bounce, i.e., abandon quickly the site. On the other hand, it also showed that fake news sites are accessed primarily due to direct visits or social sources, and have fewer backlinks, regardless of type (EDU, GOV, etc.) in comparison to real news sites. In fact, they tend to have a lifetime of a couple of years, and groups of such websites synchronize their uptime and what type of content they serve within the group. These findings point to an opportunistic (eco)system of websites, whose goal is not to keep users engaged, informed and recurring, but rather to track users to sell ads and turn quick profit. These websites aim to get visitors by spreading their existence via social media or other means, which can directly bring users to the fake news content for consumption and ad-targeting and ad-delivery.

In fact, this study found the top players in user Web tracking such as Google's *googlesyndication* and *google.com*, as well as Facebook's *facebook.net* and *facebook.com*, cumulatively, were in the great majority of such fake news websites. Perhaps the easy integration of such ad-trackers and services has allowed them to exist in many of these fake news websites. What is of particular importance here is the potential data aggregation that these "monopoly" players can perform through their intense tracking, in conjunction with their sketchy practices when collecting user-consent for this tracking [29]. Upcoming privacy laws should pay attention to these aggregation practices, since such companies can be the target of legal or illegal querying for information on specific individual's preferences or large user audiences, by governmental or other advertising agencies, and for political purposes such as during elections, or for other marketing purposes.

Moreover, our effort to automatically detect with ML models the websites that serve fake news based on their traffic profile showed that it is possible to do this with high accuracy. In fact, something of future interest would be to investigate how such ML models can be updated at near-real time, with data collection that happens at regular intervals or at the discovery of a news website. This effort can be done in a crowd-sourced fashion across multiple online users, by deploying the ML pipeline we outlined earlier into a browser plugin, e.g., such as the Check-It plugin [32]. Then, the plugin can (1) perform the crawling of network metadata for the website visited from its user, and (2) apply the ML model we provide. The plugin can also report these metadata collected per website to a centralized location for updating the ML model. In case privacy of users is at stake, privacy-preserving methodologies can be used, that employ Federated Learning techniques for training the ML model, coupled with local differential privacy applied at the user devices, or even using Trusted Execution Environments of PPFL [22]. Also, different news mobile apps can collaborate to build on-device better ML models, by sharing their collected data or pre-models using FLaaS [20].

Limitations. Our crawling was performed on websites included in well-accepted lists of fake news websites. However, such lists do not have perfect classification, and are also not always up-to-date. In fact, at any given moment, they capture only portion of the fake news websites ecosystem that is or has been active. Also, this capture is with a time delay, depending on how quickly the list maintainer can be alerted of new websites potentially serving fake news. Furthermore, network traffic data crawled per website from online services such as SimilarWeb, Alexa and CheckPageRank are not always up-to-date, since such services need time (usually a few months) to detect new websites and measure their traffic performance, ranking across the Web, etc. Therefore, future efforts to detect fake news websites should consider lists, network metrics, traffic data and labels that are available at real-time.

8 CONCLUSION

In this paper, we performed a first of its kind investigation on the fake news ecosystem. We studied and compared the user engagement on such websites by analyzing traffic-related metrics for fake and real news websites. Additionally, we explored the lifetime of fake news sites and their typical uptime periods over a time range of more than 20 years. We proposed a methodology to study how websites may be synchronizing their uptime periods, and the content they serve during these uptime periods. Our findings enabled us to characterize the traffic and behavior of fake news sites, study differences between them and real news sites and also build a novel, content-agnostic machine learning (ML) classifier for automatic detection of fake news websites that are not yet included in manually curated blacklists. We tested various supervised ML methods for our classifier which achieved a very good performance: F1 score up to 0.942 and AUC of ROC up to 0.976. We discussed practical implications of our findings and ML classifier, and offered future directions of research.

ACKNOWLEDGMENTS

The research leading to these results received partial funding from the EU H2020 Research and Innovation programme under grant agreements No 830927 (Concordia), No 871793 (Accordion), No 871370 (Pimcity) and Marie Sklodowska-Curie No 690972 (Protasis). These results reflect only the authors' view and the Commission is not responsible for any use that may be made of the information it contains.

REFERENCES

[1] Pushkal Agarwal, Sagar Joglekar, Panagiotis Papadopoulos, Nishanth Sastry, and Nicolas Kourtellis. 2020. Stop Tracking Me Bro! Differential Tracking of User Demographics on Hyper-Partisan Websites. In *Proceedings of The Web Conference 2020 (WWW '20)*. ACM, New York, NY, USA, 1479–1490.

[2] Alexandre Bovet and Hernán A Makse. 2019. Influence of fake news in Twitter during the 2016 US presidential election. *Nature communications* 10, 1 (2019), 1–14.

[3] Caye Group. 2004. Check Page Rank - Check Your PageRank Free! www.checkpagerank.net. (2004).

[4] Saska Cvetkovska, Aubrey Belford, Craig Silverman, and J. Lester Feder. 2018. The Secret Players Behind Macedonia's Fake News Sites. https://www.occrp.org/en/spooksandspin/the-secret-players-behind-macedonias-fake-news-sites. (2018).

[5] DeepAI. 2019. Cosine Similarity. https://deepai.org/machine-learning-glossary-and-terms/cosine-similarity. (2019).

[6] fanboy, MonztA, Famlam, Khrin. 2020. EasyList filterlist project. https://easylist.to/pages/about.html. (2020).

[7] Adam Fourney, Miklos Z Racz, Gireeja Ranade, Markus Mobius, and Eric Horvitz. 2017. Geographic and Temporal Trends in Fake News Consumption During the 2016 US Presidential Election.. In *CIKM*, Vol. 17. 6–10.

[8] BBC Future. 2019. I was a Macedonian fake news writer. https://www.bbc.com/future/article/20190528-i-was-a-macedonian-fake-news-writer. (May 2019).

[9] Bilal Ghanem, Paolo Rosso, and Francisco M. Rangel Pardo. 2019. An Emotional Analysis of False Information in Social Media and News Articles. *ACM Transactions on Internet Technology (TOIT)* 20 (2019), 1 – 18.

[10] David A. Graham. 2019. Some Real News About Fake News. https://www.theatlantic.com/ideas/archive/2019/06/fake-news-republicans-democrats/591211/. (2019).

[11] Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. 2019. Fake news on Twitter during the 2016 US presidential election. *Science* 363, 6425 (2019), 374–378.

[12] Benjamin D Horne, William Dron, Sara Khedr, and Sibel Adali. 2018. Assessing the news landscape: A multi-module toolkit for evaluating the credibility of news. In *Companion Proceedings of the The Web Conference 2018*. 235–238.

[13] Internet Archive. 2001. Wayback Machine. https://archive.org/web/. (2001).

[14] Joanna Geary James Ball. 2012. Quantserve (Quantcast): What is it and what does it do? https://www.theguardian.com/technology/2012/apr/23/quantcast-tracking-trackers-cookies-web-monitoring. (2012).

[15] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, Yu Wang, and Jiebo Luo. 2017. Rumor Detection on Twitter Pertaining to the 2016 U.S. Presidential Election. *ArXiv* abs/1701.06250 (2017).

[16] Z. Jin, J. Cao, Y. Zhang, J. Zhou, and Q. Tian. 2017. Novel Visual and Statistical Image Features for Microblogs News Verification. *IEEE Transactions on Multimedia* 19, 3 (2017), 598–608.

[17] Arit John. 2020. Satanism and sex rings: How the QAnon conspiracy theory has taken political root. https://www.latimes.com/politics/story/2020-07-15/qanon-conspiracy-theory-congressional-candidates. (2020).

[18] Claire Pedersen Juju Chang, Jake Lefferman and Geoff Martz. 2016. When Fake News Stories Make Real News Headlines. https://www.theatlantic.com/ideas/archive/2019/06/fake-news-republicans-democrats/591211/. (2016).

[19] Arjaldo Karaj, Sam Macbeth, Rémi Berson, and Josep M. Pujol. 2018. WhoTracks.Me: Shedding light on the opaque world of online tracking. (2018). arXiv:cs.CY/1804.08959

[20] Nicolas Kourtellis, Kleomenis Katevas, and Diego Perino. 2020. FLaaS: Federated Learning as a Service. In *Workshop on Distributed ML*. ACM CoNEXT.

[21] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096.

[22] Fan Mo, Hamed Haddadi, Kleomenis Katevas, Eduard Marin, Diego Perino, and Nicolas Kourtellis. 2021. PPFL: Privacy-preserving Federated Learning with TrustedExecution Environments. In *19th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys)* (June 24 – July 2). ACM, New York, NY, USA.

[23] Mozilla Foundation. 2020. Readability.js: A standalone version of the readability library used for Firefox Reader View. https://github.com/mozilla/readability. (2020).

[24] NPR/Ipsos. 2020. Topline NPR Misinformation Poll. https://www.ipsos.com/sites/default/files/ct/news/documents/2020-12/topline$_n$pr$_m$isinformation$_p$oll$_1$23020.pdf. (2020).

[25] OpenSources. 2017. https://github.com/BigMcLargeHuge/opensources/blob/master/sources/sources.csv. (2017).

[26] Nir Cohen Or Offer. 2007. SimilarWeb: Website Traffic Statistics & Analytics. www.larweb.com. (2007).

[27] Lucas Ou-Yang. 2013. Newspaper3k: Article scraping & curation. https://newspaper.readthedocs.io/en/latest/. (2013).

[28] Michalis Pachilakis, Panagiotis Papadopoulos, Evangelos P. Markatos, and Nicolas Kourtellis. 2019. No More Chasing Waterfalls: A Measurement Study of the Header Bidding Ad-Ecosystem. In *Proceedings of the Internet Measurement Conference (IMC '19)*. ACM, New York, NY, USA, 280–293.

[29] Emmanouil Papadogiannakis, Panagiotis Papadopoulos, Nicolas Kourtellis, and Evangelos P. Markatos. 2021. User Tracking in the Post-cookie Era: How Websites Bypass GDPR Consent to Track Users. In *Proceedings of the WWW*.

[30] Panagiotis Papadopoulos, Nicolas Kourtellis, Pablo Rodriguez Rodriguez, and Nikolaos Laoutaris. 2017. If you are not paying for it, you are the product: How much do advertisers pay to reach you?. In *Proceedings of the 2017 Internet Measurement Conference*. 142–156.

[31] Panagiotis Papadopoulos, Peter Snyder, Dimitrios Athanasakis, and Benjamin Livshits. 2020. Keeping out the Masses: Understanding the Popularity and Implications of Internet Paywalls. In *Proceedings of The Web Conference 2020 (WWW '20)*. ACM, New York, NY, USA, 1433–1444.

[32] Demetris Paschalides, Chrysovalantis Christodoulou, Rafael Andreou, George Pallis, Marios D Dikaiakos, Alexandros Kornilakis, and Evangelos Markatos. 2019. Check-It: A plugin for detecting and reducing the spread of fake news and misinformation on the web. In *2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. IEEE, 298–302.

[33] Puppeteer Developers. 2020. Puppeteer: Headless Chrome Node.js API. https://pptr.dev/. (2020).

[34] Juan Ramos et al. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, Vol. 242. Piscataway, NJ, 133–142.

[35] Leonard Richardson. 2020. Beautiful Soup Documentation. https://www.crummy.com/software/BeautifulSoup/bs4/doc/. (2020).

[36] Joel Rose. 2020. Even If It's 'Bonkers,' Poll Finds Many Believe QAnon And Other Conspiracy Theories. https://www.npr.org/templates/story/story.php?storyId=951095644&live=1&t=1620650686416. (2020).

[37] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Alessandro Flammini, and Filippo Menczer. 2017. The spread of fake news by social bots. *arXiv preprint arXiv:1707.07592* 96 (2017), 104.

[38] Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, and Yan Liu. 2019. Combating Fake News: A Survey on Identification and Mitigation Techniques. *ACM Trans. Intell. Syst. Technol.* 10, 3, Article 21 (April 2019), 42 pages. https://doi.org/10.1145/3305260

[39] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *SIGKDD Explor. Newsl.* 19, 1 (Sept. 2017), 22–36. https://doi.org/10.1145/3137597.3137600

[40] Kai Shu, Suhang Wang, and Huan Liu. 2019. Beyond News Contents: The Role of Social Context for Fake News Detection. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (WSDM '19)*. ACM, New York, NY, USA, 312–320.

[41] Daniel Sieradski. 2017. B.S. Detector. https://gitlab.com/bs-detector/bs-detector. (2017).

[42] SimilarWeb. [n. d.]. Similar Web Average Visit. https://support.similarweb.com/hc/en-us/articles/115000501485-Average-Visit-Duration. ([n. d.]).

[43] SimilarWeb. [n. d.]. Similar Web Bound Rate. https://support.similarweb.com/hc/en-us/articles/115000501625-Bounce-Rate. ([n. d.]).

[44] Joanna Geary Teodora Beleaga. 2012. ScorecardResearch (ComScore): What is it and what does it do? https://www.theguardian.com/technology/2012/apr/23/scorecardresearch-tracking-trackers-cookies-web-monitoring. (2012).

[45] Chris J Vargo, Lei Guo, and Michelle A Amazeen. 2018. The agenda-setting power of fake news: A big data analysis of the online media landscape from 2014 to 2016. *New media & society* 20, 5 (2018), 2028–2049.

[46] Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. 2017. The Web Centipede: Understanding How Web Communities Influence Each Other through the Lens of Mainstream and Alternative News Sources. In *Proceedings of the 2017 Internet Measurement Conference (IMC '17)*. ACM, New York, NY, USA, 405–417.

[47] Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. 2019. Disinformation warfare: Understanding state-sponsored trolls on Twitter and their influence on the web. In *Companion Proceedings of The 2019 World Wide Web Conference*. 218–226.

[48] Savvas Zannettou, Michael Sirivianos, Jeremy Blackburn, and Nicolas Kourtellis. 2019. The web of false information: Rumors, fake news, hoaxes, clickbait, and various other shenanigans. *Journal of Data and Information Quality (JDIQ)* 11, 3 (2019), 1–37.