

Who Funds Misinformation? A Systematic Analysis of the Ad-related Profit Routines of Fake News Sites

This work is published in WWW '23. Please cite DOI: 10.1145/3543507.3583443

Emmanouil Papadogiannakis
FORTH & University of Crete
Greece

Panagiotis Papadopoulos
FORTH
Greece

Evangelos P. Markatos
FORTH & University of Crete
Greece

Nicolas Kourtellis
Telefonica Research
Spain

ABSTRACT

Fake news is an age-old phenomenon, widely assumed to be associated with political propaganda published to sway public opinion. Yet, with the growth of social media, it has become a lucrative business for Web publishers. Despite many studies performed and countermeasures proposed, unreliable news sites have increased in the last years their share of engagement among the top performing news sources. Stifling fake news impact depends on our efforts in limiting the (economic) incentives of fake news producers.

In this paper, we aim at enhancing the transparency around these exact incentives, and explore: Who supports the existence of fake news websites via paid ads, either as an advertiser or an ad seller? Who owns these websites and what other Web business are they into? We are the first to systematize the auditing process of fake news revenue flows. We identify the companies that advertise in fake news websites and the intermediary companies responsible for facilitating those ad revenues. We study more than 2,400 popular news websites and show that well-known ad networks, such as Google and IndexExchange, have a *direct advertising* relation with more than 40% of fake news websites. Using a graph clustering approach on 114.5K sites, we show that entities who own fake news sites, also operate other types of websites pointing to the fact that owning a fake news website is part of a broader business operation.

CCS CONCEPTS

• **Information systems** → **Online advertising**; **Traffic analysis**; **Data extraction and integration**.

KEYWORDS

Fake News, Online Advertising, Web Monetization

1 INTRODUCTION

Misinformation, or formally, spreading “incorrect or misleading information” [71], is not a new phenomenon, but the tools people use to spread misinformation have dramatically improved with the Internet and social media. Fake news and misinformation, not only pose serious threats to the integrity of journalism, but have also created societal turmoils in the economy [95], the political world [4, 70] and even in human life [105]. Unlike the yellow newspapers of the past that have been capitalizing on fake news for decades, social media and search engines pose an additional threat to truth: the more luring the content of a website is, the more it is promoted by

the algorithms underpinning these platforms. BBC interviewed 50 experts about the “grand challenges of the 21st century” and many of them named propaganda and fake news [50] as a key challenge.

Considering its significant impact, tech firms, researchers, governments and stakeholders have explored various methods to identify and curtail the spread of fake news. Google and other tech companies, signed up to a voluntary EU code of conduct which required them to “improve the scrutiny of ad placements to reduce revenues of the purveyors of disinformation” [49]. Also, there is an abundance of academic works aiming at analyzing [1, 18, 71, 92, 106] or detecting [90, 96, 108, 110] fake news sources on the Web.

Despite these important actions, unreliable news sites significantly increased (2.1×) their share of engagement among top performing news sources in the past year alone [42]. The success of curbing fake news primarily depends on the efforts to reduce or even eliminate the incentives of fake news producers. But, admittedly, there is little we know about the incentives and funding of fake news on the Web. Aside from various political gains that may motivate the spread of doctored narratives [1, 98], disseminating fake news has been a lucrative Web business [66]. The ad industry provides wide avenues for high revenues: for every \$2.16 spent on news websites in USA, \$1 is spent on misinformation [53, 62]. In fact, ad-tech agencies intensely track [40, 85] and programmatically bid [80, 82, 88] for ad spaces (of lower cost) that reside in websites of questionable content. Thus, ad budgets move from high quality news websites to low-cost, controversial ones [44], with various examples of ads from prestigious companies (e.g., Microsoft, Citigroup, IBM) and small business owners being placed on (and thus unwittingly funding) websites that promote fake or even illegal (e.g., Jihadi [36] and neo-nazi [39] related) content.

In this study, we shed light on the revenue flows of fake news websites by investigating who supports and maintains their existence. We do not examine what misinformation is, rather, we investigate who provides revenue to fake news websites. We systematize the auditing process of digital advertising in those websites by developing a methodology, which enables us to identify (i) the intermediary companies that sell the ad space of fake news websites to the ad ecosystem, (ii) the advertisers who buy the ad space on such sites, and (iii) the type of ads they place.

The contributions of this paper are summarized as follows:

- (1) We develop a novel ad detection methodology which enables us to identify the advertisers that collaborate with fake news

websites. We find that about 70% of the fake news websites advertise “Business” products and services, and close to 40% display “Entertainment” advertisements.

- (2) We study who provides the ad revenues of fake news websites and show that the most well-known legitimate advertising networks (such as google.com, indexexchange.com, and appnexus.com) have a *direct* advertising relation with more than 40% of the fake news websites in our dataset, and have a *reseller* relation with more than 60% of those sites.
- (3) We show that owners of fake news websites own other types of websites as well, including “Entertainment”, “Business”, and “Politics”. This implies that the operation of an average fake news website is not an isolated or outlying event, but instead is probably part of a wider business function.
- (4) We make our lists of fake and real news websites, ad creatives collected on top 100 websites, fake news clusters, and code of crawler and ad detection method publicly available [83, 84].

2 RELATED WORK

Fake News: There has been a lot of effort to create datasets that enable future research on misinformation [76]. Most recently, in [65], authors produced a dataset regarding fake news information related to the COVID-19 pandemic, while in [89], authors manually annotated news articles and social media posts of real or fake COVID-19 stories. In [46], authors collected and evaluated news articles regarding American Politics, resulting in a dataset of fake news and satirical articles, along with a factual article that disproves them. In [109], authors analyzed over 2K news articles and 140K tweets on the COVID-19 pandemic, formed lists of reliable or unreliable news publishers, and explored spread of COVID-19 articles on Twitter.

Similar to our work, in [7] authors explored the advertising market of traditional, fake news and low-quality news websites. Using a manually curated list of popular ad servers, they found that fake publishers rely on credible ad servers to display ads and monetize their traffic. In [55], authors utilized non-perceptual features (e.g., domain name, DNS config) to train a multi-class model that detects disinformation websites in the wild. In [52], Han et al. studied how Web infrastructure supports misinformation and hate speech websites. They found that fake news websites disproportionately rely on hosting providers (e.g., Cloudflare), and that they mainly rely on Revcontent and Google to generate revenue.

Bakir et al. [2] discussed how the lack of understanding and control advertisers have regarding where their ads appear, enables fake news websites to generate revenue. They explained how fake news websites can proliferate by moving to a new ad network once blocked in another. Zeng et al. [107] studied problematic ads (e.g., clickbait, scams) and their prevalence across news websites, as well as the ad platforms that serve them. Similar to our work, they discover that fake news websites work with the same ad platforms as real news websites and that similar ads are served in both categories. However, contrary to our work, authors did not study the revenue flow associated with ads and only focus on the ad content.

Finally, the Global Disinformation Index often conducts studies to assess the ad companies that inadvertently facilitate misinformation websites [58, 59]. The Check My Ads Institute reviews the

adtech industry and attempts to disrupt the revenue flows of disinformation and hate speech outlets [60]. Similarly, the Sleeping Giants activism movement creates awareness regarding how ads are distributed across the ecosystem and has managed to reduce the ad revenue of fake news websites [8, 73].

Website Administration: Academic research has focused on identifying the legal entities that control and operate websites. The methodology followed in this work is closely related to the one presented in [86]. Specifically, authors proposed a graph-based model of website administration using ad network and tracking services relationships. Through a large-scale analysis on the monetization models of ad networks and Web publishers, they detected patterns of preferential administration of websites. In our work, we make use of the proposed Metagraph to detect websites operated or even owned by the same entity. We refrain from analyzing the behavior of intermediary publishing partners since they do not provide any additional information to this work. In [99], authors studied security threats and the involved entities by making use of HTTPS certificates to extract organization names. In [10] the authors utilized email addresses found in WHOIS records to extract groups of domains owned by the same entity.

Ad Detection: The study, detection and exclusion of advertisements in websites has been the focus of research work for a long period of time. In [74], the authors proposed *MadTracer*, a system that detects malicious advertisements in websites based on the redirection chains among publishers and ad networks. In [7], similar to our methodology, authors crawl websites and extract URLs embedded in the webpage and in iFrames. Using EasyList, they form a list of popular ad servers, against which URLs are matched. Contrary to our work, they do not examine network traffic and delivered content for ad detection. In [61], authors presented *AdGraph*, a graph-based system that detects ads and other tracking resources in websites. AdGraph provides a graph representation of the website rendering, network traffic and Javascript execution. In [100], the authors presented *PageGraph*, a similar but more robust graph representation system. In [97] authors proposed *WebGraph* which builds a graph representation of the webpage but focuses on the actions of ads instead of their content. Contrary to these techniques, we do not focus on the rendering of websites or the execution of code, nor do we use a trained model. Our methodology combines external block lists with network traffic monitoring, making it more agile to adapt, as it does not require (re)training models.

3 OVERVIEW OF METHODOLOGY

In this section, we outline the methodological steps we follow to investigate fake news websites and the entities that support them. As illustrated in Figure 1, we first construct a list of fake and real news websites and crawl them to collect ad-related data on each.

3.1 Fake and real news website lists

We utilize publicly available datasets to create a corpus of fake and real news websites, and ensure the reproducibility of our work.

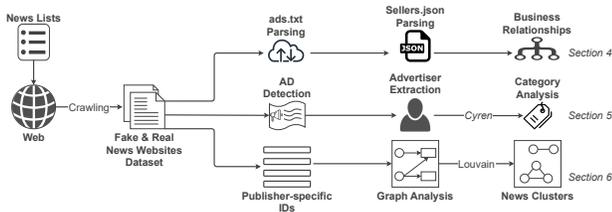


Fig. 1: Overall methodology of the present study.

- (1) MediaBias/FactCheck (MBFC) [20]: MBFC is an independent organization that aims to detect bias of media and other information sources by following a very strict manual methodology [19], that makes use of a combination of objective measures¹. We download the list on 15 Nov 2021 and extract the websites that have been labeled as “Questionable Source” or “Conspiracy-Pseudoscience” and have “Low” or “Very Low” factual reporting and their credibility has been described as “Low”. These websites manifest extreme bias, obvious propaganda, lack of proper sourcing to credible information, complete lack of transparency, and focus on spreading fake news. From the original list of 3,915 websites, we conclude with 816 fake news websites that meet the above criteria.
- (2) Columbia Journalism Review (CJR) [93]: CJR is a journal for professionals of various disciplines. Their list was created by merging some of the most common fake news lists (e.g., OpenSources [81], Politifact [91] and Snopes [101]). CJR curated the list to remove high-partisan websites that do not serve fake news content. From the resulting list, we consider websites that have been labeled as “fake news”, “conspiracy” or “extremely biased websites”, and end up with a list of 350 websites².
- (3) Golbeck et al. [46]: In this dataset, authors focus on fake news and satirical articles related to USA politics, posted after January 2016. They follow a manual investigation process, where each article is evaluated by two researchers. Additionally, for fake news articles, they provide a link to a well-researched, factual article that rebutted the fake news story. From this list, we select 55 websites that have been found to have published at least three such articles by both evaluators. This threshold has been determined based on empirical analysis.
- (4) Zhou et al. [109]: Authors created a dataset of 2,029 news articles and 140,820 tweets related to COVID-19. Regarding the news articles, they extracted knowledge using *NewsGuard* [57] and MBFC. Similar to the above, we extract 31 websites that have published at least three fake stories.

Real News: Additionally, we form a list of credible news websites that serve factual content, cite credible sources and usually cover both sides of reported stories. We focus on websites that have been evaluated by MBFC and have been found to have minimal or no bias. Specifically, we extract websites that have been labeled as “Pro-Science”, “Least-Biased”, “Left-Center” or “Right-Center” and have “High” or “Very High” factual reporting and “High” credibility. We do not make any assumptions about the spread of misinformation across the political spectrum, however, MBFC uses the labels “Left-Center” and “Right-Center” for websites which are less biased and

¹Already used in numerous past studies [33, 34, 45, 51, 109]

²At the moment of this writing, it is accessible through an archive site [93].

Source	Description	# Websites
1. Media Bias/Fact Check [20]	Questionable Sources	816
2. CJR [93]	Fake News & Biased	350
3. Golbeck et al. [46]	Fake & Satire Articles	55
4. Zhou et al. [109]	News articles	31
5. Media Bias/Fact Check [20]	High Credibility	1,368
Total unique fake news websites		1,044
Total unique real news websites		1,368

Table 1: Sources of fake and real news sites and unique total used.

generally trustworthy. This, in conjunction with the fact that we also require that they have been labeled with high factual reporting, ensures that such websites are credible. This approach results in a list of 1,368 credible websites, which we refer to as *real news*.

Fake & Real News Lists: By combining these sources, we construct a list of 1,044 unique fake news and 1,368 unique real news websites. Please note that there is an overlap across fake news lists and there are websites which can be found in multiple lists. For instance, the website *infowars.com* has been labeled as a misinformation source in all 4 sources. Table 1 summarizes the aforementioned sources of fake and real news websites. Our lists are publicly available [84]. To understand the popularity of the sites in our dataset, in Figure 2 we plot their ranking based on the Tranco list [72], from 18.10.2021 to 16.11.2021³. We see that 45 fake news websites are among the top 10K most popular sites and such rankings usually translate in a wide audience with millions of visitors per month. We observe that websites in the two lists have very similar rankings, suggesting that they attract a similar number of visitors and therefore are directly comparable.

3.2 Website crawling

We develop a puppeteer-based crawler that stores (i) the HTML content of the visited website, (ii) a cookie-jar for both first-party and third-party cookies, (iii) the `ads.txt` file (if present), (iv) a screenshot of the landing page, and (v) the HTTP(S) network traffic. We also implement the ad-detection mechanism, described later in Section 5.1. The implementations of both the crawler and the ad detection methodology are publicly available [83]. Using this crawler, we visit the landing page of real and fake news websites on 13 Dec 2021. The crawler was located in an EU-based institution and collected about 31GB of data. The timeout for loading each website was set up to 60 seconds. Ethical aspects of our study are discussed in Appendix A.

4 NEWS WEBSITE FINANCING

4.1 Who sells ad space on fake news sites?

First, we study the entities selling ad space to understand who facilitates the monetization of news websites. To achieve this, we utilize `ads.txt` files served by websites. An `ads.txt` file [67] is a simple text file located at the root of a website that explicitly states which auctioneers are authorized to sell the impression inventory of this website. In order for the entire ad ecosystem to work as expected, Supply-Side Platforms (SSPs) should ignore inventory which they are not authorized to sell, while Demand-Side Platforms (DSPs) should not buy inventory from unauthorized sellers. As

³<https://tranco-list.eu/list/YKQG/full>

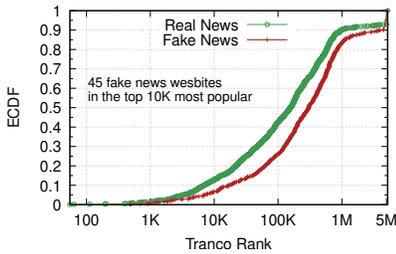


Fig. 2: Distribution of rank of news websites based on Tranco. There are both fake and real news websites with very high popularity. In addition, we see that both lists contain websites of comparable popularity.

shown in Figure 3, each record in `ads.txt` is an entry with comma-separated fields and it authorizes a specific SSP to sell impressions for this website. These fields are: (i) the domain of the SSP, (ii) an identifier which uniquely identifies the account of the publisher within the service (iii) the relationship for this account (can be either DIRECT i.e., the publisher is the owner of the specified account or RESELLER i.e., a third party has been assigned by the website owner to manage the specified account), and (iv) optionally, an identifier that maps to the company listed in (i) and uniquely identifies it within a certification authority. Every such entry defines a business relationship between the owner of the website and the seller [3].

We parse and analyze the content of these files, fetched by the crawler described in Section 3.2. In total, we find 198 fake news websites and 627 real news websites serving a valid `ads.txt` file that follows the specification [67]. According to specification [67], relationships of DIRECT type indicate that the publisher (i.e., the owner of the content) directly controls the specific account in the respective service. Consequently, these relationships are of special interest, since they disclose a direct business contract between the publisher and the ad network. An analysis of the RESELLER relationships is presented in Appendix B.

For each ad network, we measure the portion of websites that provide an `ads.txt` file and have a business relationship with it. We find that, on average, fake news websites in our dataset form direct business relationships with 27 ad systems, while surprisingly, real news websites do so with 41 systems. In Figure 5, we illustrate the top 10 most popular digital sellers of ads for the DIRECT relationships that appear in both real and fake news websites (i.e., intersection). As the figure suggests, a large portion of real news websites tends to form DIRECT business relationships with well-known ad networks (e.g., 96% of real news with `google.com`, and 82.1% with `indexexchange.com`). Even though ad platforms are found in these files, they might not end up serving any ads due to the nature of programmatic advertising. However, there is still a business relationship between the website and the ad network.

What is more interesting, however, is that a lot of fake news websites also have direct business relationships with these ad networks. Indeed, 80.8% of fake news websites have a direct business relationship with `google.com`, 49% of fake news websites with `indexexchange.com`, and 52.5% with `appnexus.com`. By independently examining the top ad systems for fake and real news websites, we find that `revcontent.com` is the only ad system that is popular (i.e.,

```
google.com, pub-0007030388228657, DIRECT
google.com, pub-5519830896693885, DIRECT
google.com, pub-4764333688337558, DIRECT
google.com, pub-4573231550355221, RESELLER
pubmatic.com, 160907, DIRECT
indexexchange.com, 194273, DIRECT
media.net, 8CU5DERG1, DIRECT
rhythmone.com, 895733750, DIRECT
video.unrulymedia.com, 895733750
sovrn.com, 357833, DIRECT
```

Fig. 3: Snippet of the `ads.txt` file served by `ainarsgames.com` on October 2022. Each entry is a unique business relationship. Identifiers can be matched against `sellers.json` files (see Fig 4).

```
{
  "seller_id": "pub-0007030388228657",
  "seller_type": "PUBLISHER",
  "name": "Iveta Veinberga",
  "domain": "ainarsgames.com"
},
...
{
  "seller_id": "pub-4573231550355221",
  "seller_type": "BOTH",
  "name": "AdPlus Media Inc.",
  "domain": "ad.plus"
}
```

Fig. 4: Snippet of the `sellers.json` served by Google on October 2022. Each entry represents an entity with which Amazon has a business relationship.

ranked 5th) among the ad networks integrated with fake news websites, but ranked very low (i.e., 51st) among the ad networks of real news websites, which suggests that this network is preferred by fake news websites. Contrary, we find `yahoo.com` being preferred by real news websites: 68% of them form a business relationship with `yahoo.com`, while only 30% of fake news websites do so.

We observe that only a portion of fake news websites in our list provide an `ads.txt` file. We recrawl our list of fake news websites on January 31, 2023 and find that 262 websites now serve `ads.txt` files (up from 198). We find similar results, with the top ad-networks being almost identical. 83.9% of fake news websites have a DIRECT relationship with Google, 47.32% with IndexExchange, etc. Studying the third parties that fake news websites interact with, shows that for the 198 websites serving `ads.txt` files, 94.95% of them interact with Google-owned tracking or ad-serving domains. We classify domains as trackers based on the list provided by Disconnect [56]. Looking into all the crawled fake news websites, regardless if they serve `ads.txt` files, we find 84.08% of them interacting with Google. The above support our findings that (i) the fake news websites with `ads.txt` files we studied are representative of the ad-ecosystem; (ii) popular ad systems provide ad revenue to fake news websites.

Finding: Although the percentages vary from one ad network to the next, Figure 5 suggests that, on average, popular ad networks have DIRECT business relationship with about half of the fake news websites we analysed. Consequently, fake news websites rely on popular and credible ad networks to generate revenue. It is interesting to note that before starting such a business relationship between an ad network and a website, there is a vetting process to be followed. For example, Google’s AdSense ensures that the website complies with its policy [14]. One might expect that during the review process previously described, the popular ad networks would not approve requests of fake news websites, or of websites proven to publish misinformation.

4.2 Business Relationships

Although `ads.txt` files provide a clear view of the ad ecosystem in the analysed fake news websites, this view is based on data provided by the fake news websites themselves. To provide the point of view of the sellers, we utilize `sellers.json` files as provided by the advertising services. `sellers.json` is a complementary mechanism to `ads.txt` introduced by the IAB Tech Lab to oppose ad fraud and profit from counterfeit inventory. Specifically, `sellers.json` files (format shown in Figure 4) can be used by buyers to discover the

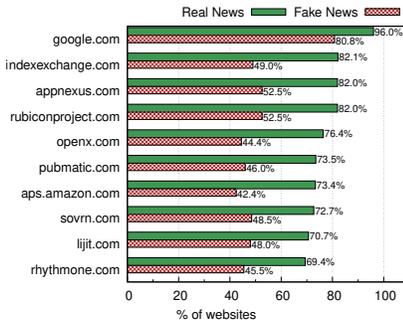


Fig. 5: Most popular authorized digital sellers with DIRECT relationship in ads .txt files. Identifiers of such entries indicate that the publisher is the direct owner of the account. We observe that the majority of news websites have business relationship with Google.

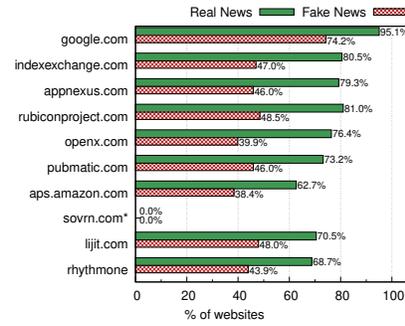


Fig. 6: Direct business relationships between news websites and ad networks verified by entries in ads.txt and sellers.json files. sovrn.com is excluded since its sellers .json file could not be retrieved.

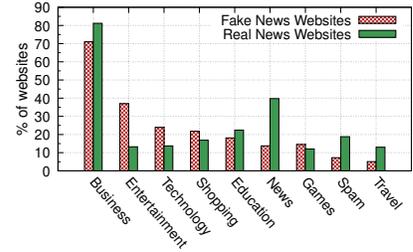


Fig. 7: Distribution of categories of advertisers appearing in fake and real news websites. For both types the majority of advertisers provide business-related information.

final sellers of a bid request (either direct sellers or intermediaries). In an attempt for a more transparent marketplace, each seller (i.e., SSP) publishes in its own `sellers.json` file all entities, with which it has business relationships. According to the specification [68], the list of entities which are represented by the ad network must be included in this file, even if their identity is confidential. For that reason, a seller ID is required. This ID is the same as the one that appears in the website’s `ads.txt` file. Finally, for each seller ID, the type of the account must be specified (i) as PUBLISHER, if ad inventory is sold on a website directly owned by the company and the ad network pays the company directly, (ii) as INTERMEDIARY, if ad inventory is sold by an entity which does not directly own it. Using this information, we are able to extract reliable information and match it against the one provided by websites to ensure that there are no falsely listed business relationships.

To verify the business relationships we found in `ads.txt` files, on 12 Jan 2022, we download and parse the `sellers.json` files of all popular ad services found in our previous analysis. We exclude `sovrn.com` from this experiment as we were unable to retrieve its `sellers.json` file. For each identifier with DIRECT relationship found in `ads.txt` files of news websites, we verify whether the respective business relationship is also registered by the advertising system in its `sellers.json` file. Note that we do not investigate whether there is a relationship mismatch since they are considered as out of scope for this work and left for future research. Instead, we focus on whether there is a business relationship of any kind between a news website and the respective ad network.

In Figure 6, we present our findings for the top 10 most popular sellers. We find that for all ad networks, the results reported in Figure 5 and Figure 6 are very similar (or even the same). We attribute the small disparities between the two figures to (i) the fact that `ads.txt` files might not be all-inclusive, up-to-date or syntactically correct [3], and (ii) the common discrepancies and mislabeled relationships between `ads.txt` and `sellers.json` files [35].

Despite the differences that may exist, the important thing to focus is that both points of view agree. A substantial percentage of fake news websites receive ads through well-known services including `google.com`, `indexexchange.com`, `appnexus.com`, etc. Even though according to Google’s Terms of Service, content that makes false claims or contradicts scientific consensus is not eligible for monetization [48], this is not the case. 74.3 - 80.8% of the fake news websites in our analysis have a DIRECT relationship with

`google.com` (i.e., receive ads through Google), 47.0 - 49.0% with `indexexchange.com`, and 46.0 - 52.5% with `appnexus.com`. Please note that compared to previous work [107], these findings have not been inferred or detected using a custom methodology. The importance of these results along with the ones presented in Section 4.1, is that they are reported by the involved entities themselves.

Finding: It is evident that news websites tend to form business relationships with ad networks in order to monetize their published content and generate revenue. Based on our analysis, we find that not all such networks evaluate their clients or refuse deals with fake news websites. Such ad companies prefer to increase their profits at the expense of a more transparent, reliable and safe Web. Therefore, even if these business relationships have been formed due to lack of thorough examination of news websites, it is evident that some ad networks facilitate fake news content on the Web.

5 ADVERTISING ON FAKE NEWS WEBSITES

5.1 Ad Detection

Detecting ads embedded in websites is not trivial [100]. The main difficulty is that the final advertiser may be selected after an auction and is accessed after several re-directions. To detect ads embedded in websites and identify the actual advertisers, we propose and implement a novel methodology as outlined in Figure 8. The novelty of this methodology lies in the fact that it consists of two distinct components: *external blocking lists* and *network traffic monitoring*.

First (step 1), we extract all URLs from the landing page of the website. Using the Chrome DevTools protocol, we extract all hyperlinks that can be found even in iFrames, where ads are most commonly found, or the Shadow DOM. From the extracted URLs, we consider only URLs to other domains. Next (step 2), we search for URLs belonging to ad networks and represent ads. When users click on such URLs, either directly, or because they clicked on an image, they are redirected to the advertiser’s landing page. We make use of Brave’s adblock engine [9] and the popular open-source filter lists *EasyList* [41] and *uBlock Origin* [54] to evaluate URLs, and detect (step 3) those which are ads.

Additionally, our methodology is able to detect ad URLs that belong to the actual advertiser, and not to an ad network that redirects to the advertiser. We perform an application-level network traffic analysis and trace HTTP(S) requests. For each request, we extract the body of the response (step 4) and the request URL (step

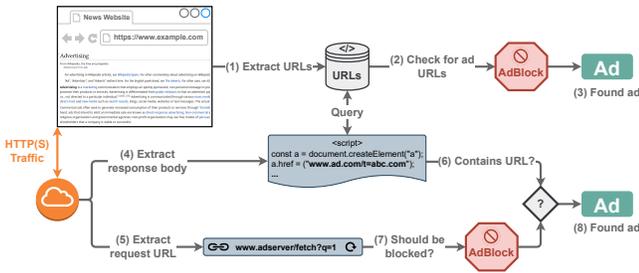


Fig. 8: Ad detection methodology that combines both external block lists and network traffic monitoring.

5). For a more robust and thorough approach, we follow all redirect chains. If we find a URL in the response, and we know that this URL has been placed in the website (step 6), we determine whether the original request was towards an advertising domain (step 7), using EasyList and uBlock Origin filter lists. If so, we deduce that this URL has been placed into the website through an ad network, and consequently, it is an ad URL (step 8). By combining the two approaches (steps 1, 4, and 5), our methodology is able to detect ad URLs that are either direct URLs to the actual advertiser, or URLs of ad networks that eventually redirect to the advertiser. To establish and attribute the actual advertiser, we navigate to the detected ad URLs and extract the landing page.

Manual Verification: To validate our methodology, we use a list of popular websites. Using *SimilarWeb* [75], we extract the 50 most popular websites from the “News and Media” and “Sports” categories, for a total of 100 websites. We select these categories based on empirical analysis, since they are more likely to contain ads. We apply our methodology for ad detection and advertiser attribution on these websites, while at the same time storing a screenshot of the website. Next, we manually evaluate how accurately our method can detect ads on these 100 websites. We find that our approach has both high Precision (92% of “ads” marked in the websites are actual ads), and Recall (87% of actual ads in the websites were correctly detected). These results indicate that our method detects very accurately most ads in websites, with very few false positives. To ensure the reproducibility of our study, we release a collection of annotated screenshots with ads detected by our methodology [84].

5.2 Who buys ad space on fake news sites?

Using our ad detection methodology, we extract the actual entities that advertise in news websites. Using a clean browser state (i.e., no synthetic personas), we visit each ad that our methodology detects and extract the domain of the advertiser. Our methodology was able to detect ~900 distinct advertisers in real news websites and ~200 advertisers in fake news websites.

We find that a considerable number of fake news websites does not monetize their content via ads. This is inline with the finding of Section 4.1 and is further discussed in Appendix E. However, on those who do, we discover that entertainment advertisers with captivating and luring ads are the most popular ads. In particular, *newscityhub.com* and *inspiredot.net* are the most popular advertisers, appearing in 15% and 14% of fake news websites, respectively. These advertisers are known for using click-bait ads with “catchy” titles that entice the visitor’s curiosity. Even unintentionally, these advertisers are common among the misinformation websites we

study and provide great revenue to their operators. Please note that often, advertisers have control over where their ads will appear and, therefore, share a portion of the ethical responsibility for the proliferation of fake news content. For example, in the Google Ads platform, advertisers can choose where their ads are displayed [15] and even exclude specific websites [13]. Similarly, Rubiconproject (now called Magnite) respects advertisers’ blocklists regarding where their ads will appear [103].

Next, we extract the categories of advertisers by utilizing *Cyren* [37], whose classification engine has already been used in previous academic works (e.g., [12, 38, 87]), and has been proven that it can classify a greater set of websites than other similar systems [11]. Using their classification service, we are able to extract the categories of over 95% of advertisers in our dataset. For websites assigned to multiple categories, we single out the most frequent label in our dataset. Figure 7 illustrates the types of distinct advertisers collaborating with real news and fake news websites.

The majority of advertisers in both fake and real news websites come from the “Business” category. This behavior is expected, since these advertisers promote websites that contain business-related information in an attempt to popularize their services or products. Also, we observe that a large number of fake news websites (almost 40%) display ads from the “Entertainment” websites. These ads contain captivating, and, sometimes even click-bait, content from celebrity websites, television and movie programs, as well as entertainment news that tempt users. The rest of the advertisers fall into “Technology”, “Shopping”, “Education”, “News”, etc. On the other hand, real news websites place ads coming primarily from advertisers of other businesses, news, and education-related services. “Spam” category seems less prominent in fake than real news sites.

Finding: We observe that click-bait and captivating ads are more likely to appear on fake news websites. Such advertisers fuel fake news content and, through their ad impressions, financially support part of the ecosystem. We also find that advertisers on fake news sites seem to be normal and legitimate business. Our results suggest that fake news websites host ads from legitimate advertisers, thus doing serious ad business and avoiding ads from malicious or dodgy sites such as SPAM, which could risk their monetization avenues or jeopardize their existence in the ad ecosystem.

6 NEWS WEBSITES OWNERSHIP

In this section, our goal is to answer: *Who owns fake news websites?* and *What other websites do the owners of fake news websites operate?* Towards this goal, we expand our dataset, since so far we focused on websites which were clearly categorized as either fake or real news. Thus, in this analysis, we include a corpus of 1,548 extra news websites from the sources of Section 3.1, which were not clearly categorized as either fake or real, for a total of 3,960 news websites.

6.1 Community Detection

To be able to answer what kind of other websites the owners of fake news websites own, we first need to determine who the owner of a fake news website is. Although this question is rather tricky to answer, we capitalize on the methodology described in [86]. The methodology makes use of four different types of Publisher-specific IDs used in three separate Google Services. Contrary to

Description	Volume	% of total	FN	RN
Initial set of websites	3,960	100.00%	-	-
Websites successfully crawled	3,311	83.61%	-	-
Websites that errored	649	16.39%	-	-
Websites with no ad-related identifiers	737	22.26%	325	172
Websites with at least one identifier	2,574	77.74%	385	1,025
Websites with all types of identifiers	184	5.56%	2	62

Table 2: Summary of crawled News websites. “FN” stands for fake news websites while “RN” stands for real news websites.

Sections 4 and 5, the analysis of this section is bound to websites that make use of such Google services. Then, websites can be linked together if they contain common such identifiers.

Publisher-specific ID detection: Such identifiers are alphanumeric values that follow strict formats and uniquely identify user accounts in popular services, such as AdSense and Google Analytics. Administrators embed these identifiers in their websites in order to use the respective service. For example, admins need to embed an identifier in the form of UA-123456-7 in order to use Google Analytics. Since some of these identifiers are associated with the receipt of the funds generated via ads, it is generally safe to assume that websites that share the same identifier (i.e., give their ad revenue to the same entity) are closely related, or even owned by the same entity [86]. Using regular expressions and common data cleaning techniques, identifiers are extracted from the HTML code of websites, network traffic and first- or third-party cookies. Then, values that are words of the English dictionary, or match a custom list of common keywords are removed. Table 2 summarizes the websites containing Publisher-specific IDs. We find that there are 385 fake news websites and 1,025 real news websites with at least one type of identifier. A rundown of the detected identifiers can be found in Appendix C. We find that for most types of identifiers, there are more domains than actual identifiers, indicating that there are identifiers being re-used in more than one domain.

Graph Analysis & Cluster Construction: Using the aforementioned detected identifiers, we construct a *Metagraph*, a graph that represents the relationships among websites. This graph contains only website nodes and those that share an identifier (that uniquely identifies an account) are connected through an edge. The weight of the edge is proportional to the number of identifiers websites share. A large edge weight represents greater confidence that these two websites are indeed operated and managed by the same entity. The notion of the Metagraph has been assessed in [86], where the authors show that it can accurately detect websites operated by the same entity and validated its performance against other techniques. Figure 9 illustrates the construction of such a toy Metagraph.

To detect clusters of websites operated or even owned by the same entity, a graph community detection algorithm is applied on the Metagraph. Contrary to [86] which uses the Girvan-Newman method, in this work we apply the Louvain method [5]. Our decision is only based on the performance benefits of the Louvain method: it is faster, scalable, and able to accommodate the entire Metagraph without performing any edge-pruning. Additionally, we integrate information from the 1MT crawl dataset [86], which contains Publisher-specific IDs found in the top 1M most popular websites of April, 2021. The resulting Metagraph contains over 114.5K website nodes and 443K edges. More information about the community detection process can be found in Appendix C.

We define a *Fake news cluster* as a community of websites that contains at least one fake news website. This implies that a community is operated or owned by an entity which, among other business, also spreads fake news. Similarly, we define a *Real news cluster* as a community with at least one real news website. It is worth noting that, in these definitions, we do not label other websites inside the clusters, we simply characterize the clusters they belong to.

6.2 Categories of website clusters

By definition, each Fake news cluster contains at least one fake news website. However, it also contains other websites as well. Figure 10 shows the types of other websites contained in each such cluster. We see that for the Fake news clusters (red bars) about 29.5% of the websites are news. The rest (almost 70%) are “not news” websites and encompass “Entertainment”, “Business”, “Politics”, “Technology”, etc. This verifies that most fake news website owners are also engaged in other types of businesses. Contrary, we observe that entities that own or operate real news websites, also tend to manage other news websites in order to reach a wider audience or even convey other types of news. It seems that both types of cluster have some diversity, but it is not clear which is more diverse.

To clarify this, we study Shannon’s diversity index [79], a statistical measure that can indicate how many different categories there are in a community, while at the same time reflecting the relative abundance of website categories. Shannon’s diversity index is defined as $H' = -\sum_{i=1}^S p_i \ln p_i$, where S is the number of different categories in the dataset (i.e., richness) and p_i is the proportion of websites belonging to category i . When all categories in a community are equally common, the Shannon index takes the maximum value $\ln(R)$. The more unequal the categories are, the smaller the index is. Shannon’s diversity index equals zero when there is only one category of websites in a community.

We apply this statistical measure to communities that contain fake news or real news websites. Figure 11 illustrates the distribution of the diversity index for fake news and real news clusters. This index is normalized by $\ln(R)$, which is the case where all categories are equally common. Consequently, the case of 0% in Figure 11 indicates that there is only one category of websites in the community, while the case of 100% suggests that the categories are equally distributed, thus revealing a diverse community. We see that Real news clusters tend to cluster higher and to the left (for the same value of y) of the Fake news clusters.

Finding: We find that Real news clusters are more homogeneous: owners of these clusters tend to focus on a smaller number of different Web businesses. At the same time, owners of Fake news clusters seem to engage in higher diversity in their business. Combined with the fact that fake news website owners have a preference towards “Entertainment” and “Business” websites, we speculate that their goal is to monetize their websites and generate revenue, and that fake news websites might be a way to make “quick buck”.

6.3 Who owns fake news websites?

In order to study fake news websites owners, we manually investigate communities⁴ that contain at least one fake news website, and discover the legal entity that operates the websites of each

⁴For the communities, we rely on the accuracy of the methodology as presented in [86].

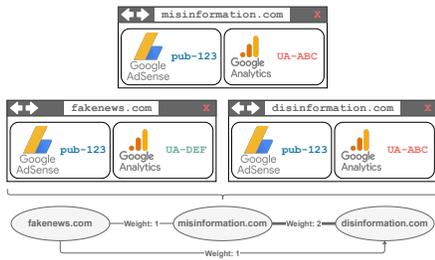


Fig. 9: Example of a Metagraph construction. Websites that share identifiers are linked together in the resulting graph. The more identifiers they share, the greater the edge weight.

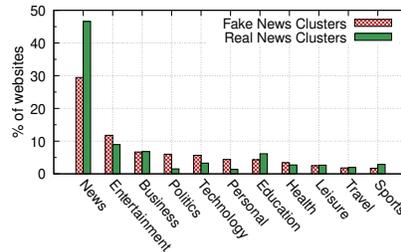


Fig. 10: Distribution of website categories in each cluster. We see that most sites in a cluster are “News”, followed by “Entertainment”, and “Business”.

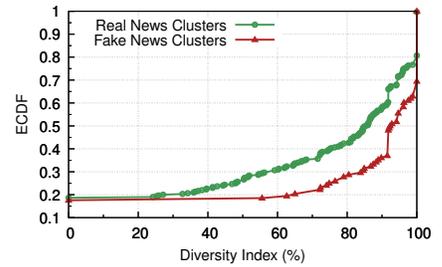


Fig. 11: Diversity score based on Shannon's Diversity Index. Real News (green circles) seem to cluster higher from Fake News (red triangles) indicating more homogeneity.

community. To provide a better understanding of the fake news ecosystem and highlight its social effect, we selectively report some of these communities. More information about the methodology we follow along with other striking examples of fake news websites ownership can be found in Appendix D. We make the clusters of fake news websites publicly available [84].

We detect a community of websites related to the *Family Research Council* (FRC), an activist group with an affiliated lobbying organization. However, one of these websites has been labeled as a “Questionable Source” by MBFC since it promotes far-right propaganda, it lacks transparency regarding funding and it has numerous failed fact checks [25]. The Southern Poverty Law Center (SPLC) designated FRC as a hate group [17]. Additionally, we discover the websites *thetruthaboutcancer.com* and *thetruthaboutvaccines.com*, owned by Ty and Charlene Bollinger. Their websites promote both unproven and dangerous remedies (i.e., pseudoscience), as well as information regarding COVID-19 and vaccines which has been proven to be false [32]. In fact, Ty and Charlene Bollinger have been identified as part of the “Disinformation Dozen”, a set of 12 individuals that produce 65% of the misinformation and misleading claims regarding COVID-19 on social media [6, 43].

Finding: These examples, along with others excluded for brevity, demonstrate the correctness and efficiency of our methodology. That is, we are able to accurately detect communities of fake news websites, owned or operated by the same entity that pushes a specific political or ideological agenda, and tries to shift the public opinion. In fact, people may be led to accept false beliefs or even make life-altering decisions based on this false information [63]. We believe our methodology can play a vital role in this problem: if a person can be informed that the website they are visiting is owned by an entity that also operates or owns fake news site(s), the visitor will most likely view the content with more caution.

7 DISCUSSION & CONCLUSION

7.1 Limitations

Even though we study what the Web ad-ecosystem uses in its majority [3], we understand that there are limitations to ads.txt files [35, 102]. Moreover, the analysis of advertisers relies on the methodology presented in Section 5.1, and, though our methodology has a high precision score, we acknowledge that it might fail to detect some ads. Also, our network monitoring approach will miss fragmented ad URLs. These limitations do not reduce the credibility of our findings, since we still study a big portion of the ad

ecosystem. Furthermore, we made efforts to exclude intermediary publishing partners from communities of websites operated by the same entity, as discussed in [86]. Finally, domain classification services might suffer from classification and disagreement flaws and no service is error-free [104]. We choose Cyren because it (i) accepts miss-classification reports, (ii) is language and content agnostic, and (iii) has a vast database of 140 million classified domains.

7.2 Summary

The success of curbing fake news primarily depends on the ability of stakeholders to remove the incentives of fake news producers. One may think that fake news sources are supported only by shady organizations enlisting people in remote countries [70] and legitimate ad-networks have pulled out from such misinformation sources. In this work, we show that this, unfortunately, is not the case.

We identify and study the companies that advertise in fake news websites and the middlemen responsible for keeping the avenues of ad revenue open. We show that popular, legitimate advertising systems (such as Google, Indexexchange and AppNexus) have a *direct* advertising relation with more than 40% of the fake news websites in our list. Through clustering based on advertiser IDs present in such websites, we report that operators of fake news sites usually operate a set of websites that include entertainment, business, politics, etc. This indicates that the operation of a fake news website is part of a larger business and not an isolated event.

We believe that the Metagraph described in [86] and used in this work provides clear understanding of relationships among websites. We plan on exploiting the sensitive information related to advertising and analytics services to develop a content-agnostic classifier that can automatically detect fake news websites. Contrary to common content-aware fake news detection schemes and manual fact-checking campaigns, we believe that such a classifier can effectively detect fake news websites that have just spawned through the entities that own or operate them.

ACKNOWLEDGMENTS

This project received funding from the EU H2020 Research and Innovation programme under grant agreements No 830927 (Concordia), No 830929 (CyberSec4Europe), No 871370 (Pimcity), No 871793 (Accordion), No 101021808 (Spatial), and No 883543 (CC-DRIVER). 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] Vian Bakir and Andrew McStay. 2018. Fake news and the economy of emotions: Problems, causes, solutions. *Digital Journalism* 6, 2 (2018), 154–175.
- [3] Muhammad Ahmad Bashir, Sajjad Arshad, Engin Kirda, William Robertson, and Christo Wilson. 2019. A Longitudinal Analysis of the Ads.Txt Standard. In *Proceedings of the Internet Measurement Conference (IMC '19)*. ACM, New York, NY, USA, 294–307. <https://doi.org/10.1145/3355369.3355603>
- [4] Marco T. Bastos and Dan Mercea. 2019. The Brexit Botnet and User-Generated Hyperpartisan News. *Social Science Computer Review* 37, 1 (2019), 38–54. <https://doi.org/10.1177/0894439317734157> arXiv:<https://doi.org/10.1177/0894439317734157>
- [5] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008, 10 (2008).
- [6] Shannon Bond. 2021. Just 12 People Are Behind Most Vaccine Hoaxes On Social Media, Research Shows. <https://www.npr.org/2021/05/13/996570855/disinformation-dozen-test-facebooks-twitthers-ability-to-curb-vaccine-hoaxes?t=1628769021116>.
- [7] Lia Bozarth and Ceren Budak. 2021. Market forces: Quantifying the role of top credible ad servers in the fake news ecosystem. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM '21, Vol. 15)*. 83–94.
- [8] Joshua A Braun, John D Coakley, and Emily West. 2019. Activism, advertising, and far-right media: The case of sleeping giants. *Media and Communication* 7, 4 (2019).
- [9] Brave Software. 2021. [adblock-rust](https://github.com/brave/adblock-rust). <https://github.com/brave/adblock-rust>.
- [10] Frank Cangialosi, Taejoong Chung, David Choffnes, Dave Levin, Bruce M. Maggs, Alan Mislove, and Christo Wilson. 2016. Measurement and Analysis of Private Key Sharing in the HTTPS Ecosystem. In *Proceedings of SIGSAC Conference on Computer and Communications Security (CCS '16)*. ACM, 628–640.
- [11] Juan Miguel Carrascosa, Jakub Mikians, Ruben Cuevas, Vijay Erramilli, and Nikolaos Laoutaris. 2014. Understanding interest-based behavioural targeted advertising. *arXiv preprint arXiv:1411.5281* (2014).
- [12] Juan Miguel Carrascosa, Jakub Mikians, Ruben Cuevas, Vijay Erramilli, and Nikolaos Laoutaris. 2015. I always feel like somebody's watching me: measuring online behavioural advertising. In *Proceedings of the 11th ACM Conference on Emerging Networking Experiments and Technologies (CoNEXT '15)*. 1–13.
- [13] Google Help Center. 2022. Exclude specific webpages and videos. <https://support.google.com/google-ads/answer/2454012>.
- [14] Google Help Center. 2022. Sign up for AdSense. <https://support.google.com/adsense/answer/10162>.
- [15] Google Help Center. 2022. Where your ads can appear. <https://support.google.com/google-ads/answer/1704373>.
- [16] Southern Poverty Law Center. [n.d.]. Alliance Defending Freedom. <https://www.splcenter.org/fighting-hate/extremist-files/group/alliance-defending-freedom>.
- [17] Southern Poverty Law Center. [n.d.]. Family Research Council. <https://www.splcenter.org/fighting-hate/extremist-files/group/family-research-council>.
- [18] Manolis Chalkiadakis, Alexandros Kornilakis, Panagiotis Papadopoulos, Evangelos Markatos, and Nicolas Kourtellis. 2021. The Rise and Fall of Fake News Sites: A Traffic Analysis. In *13th ACM Web Science Conference 2021 (Virtual Event, United Kingdom) (WebSci '21)*. Association for Computing Machinery, New York, NY, USA, 168–177. <https://doi.org/10.1145/3447535.3462510>
- [19] Media Bias/Fact Check. 2022. Methodology. <https://mediabiasfactcheck.com/methodology/>.
- [20] Media Bias/Fact Check. 2022. Search and Learn the Bias of News Media. <https://mediabiasfactcheck.com/>.
- [21] Media Bias Fact Check. 2020. Need to Know. <https://mediabiasfactcheck.com/need-to-know/>.
- [22] Media Bias Fact Check. 2020. Salem Radio Network News (SRN News). <https://mediabiasfactcheck.com/salem-radio-network-news-srn-news/>.
- [23] Media Bias Fact Check. 2021. Alliance Defending Freedom. <https://mediabiasfactcheck.com/alliance-defending-freedom/>.
- [24] Media Bias Fact Check. 2022. CheckYourFact. <https://mediabiasfactcheck.com/check-your-fact/>.
- [25] Media Bias Fact Check. 2022. Family Research Council. <https://mediabiasfactcheck.com/family-research-council/>.
- [26] Media Bias Fact Check. 2022. Health Impact News. <https://mediabiasfactcheck.com/health-impact-news/>.
- [27] Media Bias Fact Check. 2022. Medical Kidnap. <https://mediabiasfactcheck.com/medical-kidnap/>.
- [28] Media Bias Fact Check. 2022. PJ Media. <https://mediabiasfactcheck.com/pj-media/>.
- [29] Media Bias Fact Check. 2022. Vaccine Impact. <https://mediabiasfactcheck.com/vaccine-impact/>.
- [30] Media Bias Fact Check. 2022. The Vaccine Reaction. <https://mediabiasfactcheck.com/the-vaccine-reaction/>.
- [31] Media Bias Fact Check. 2023. Daily Caller. <https://mediabiasfactcheck.com/daily-caller/>.
- [32] Media Bias Fact Check. 2023. The Truth About Cancer. <https://mediabiasfactcheck.com/the-truth-about-cancer/>.
- [33] Zhouhan Chen and Juliana Freire. 2020. Proactive discovery of fake news domains from real-time social media feeds. In *Companion Proceedings of the Web Conference*. 584–592.
- [34] Rajdipa Chowdhury, Sriram Srinivasan, and Lise Getoor. 2020. Joint Estimation of User And Publisher Credibility for Fake News Detection. In *Proceedings of the 29th ACM International Conference on Information Knowledge Management (Virtual Event, Ireland) (CIKM '20)*. 1993–1996.
- [35] Credibility Coalition. 2021. Examining Opaque Programmatic Markets with the Credibility Coalition AdSellers Dataset. <https://misinfocon.com/examining-opaque-programmatic-markets-with-the-credibility-coalition-adsellers-dataset-b9ff5d6781c4>.
- [36] Robert Cookson. 2016. Jihadi website with beheadings profited from Google ad platform. <https://www.ft.com/content/b06d18c0-1bf6-11e6-8fa5-44094fd9c46/>.
- [37] Cyren. 2022. Website URL Category Check. <https://www.cyren.com/security-center/url-category-check-gate>.
- [38] Michalis Diamantaris, Francesco Marcantoni, Sotiris Ioannidis, and Jason Polakis. 2020. The Seven Deadly Sins of the HTML5 WebAPI: A Large-Scale Study on the Risks of Mobile Sensor-Based Attacks. *ACM Trans. Priv. Secur.* 23, 4, Article 19 (jul 2020), 31 pages. <https://doi.org/10.1145/3403947>
- [39] John Ellis. 2018. Dear Google: Please stop using my advertising dollars to monetize hate speech. <https://qz.com/1177168/dear-google-please-stop-using-my-advertising-dollars-to-monetize-hate-speech/>.
- [40] Steven Englehardt and Arvind Narayanan. 2016. Online tracking: A 1-million-site measurement and analysis. In *Proceedings of the SIGSAC Conference on Computer and Communications Security*. 1388–1401.
- [41] fanboy, MonztA, Famlam, and Khirin. 2021. EasyList. <https://easylist.to/>.
- [42] Sara Fischer. 2020. “Unreliable” news sources got more traction in 2020. <https://www.axios.com/unreliable-news-sources-social-media-engagement-297bf046-c1b0-4e69-9875-05443b1dca73.html>.
- [43] Center for Countering Digital Hate. 2021. The Disinformation Dozen. Why platforms must act on twelve leading online anti-vaxxers. <https://www.counterhate.com/disinformationdozen>.
- [44] Augustine Fou. 2020. Marketers Helped Fake News Kill Real News With Bullets Supplied By Ad Tech. <https://www.forbes.com/sites/augustinefou/2020/10/18/marketers-helped-fake-news-kill-real-news-with-bullets-supplied-by-ad-tech/?sh=19ac37b15d73>.
- [45] Bilal Ghanem, Simone Paolo Ponzetto, Paolo Rosso, and Francisco Rangel. 2021. FakeFlow: Fake News Detection by Modeling the Flow of Affective Information. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume (EACL '21)*. 679–689.
- [46] Jennifer Golbeck, Matthew Mauriello, Brooke Auxier, Keval H Bhanushali, Christopher Bonk, Mohamed Amine Bouzaghrane, Cody Buntain, Riya Chanduka, Paul Cheakalos, Jennine B Everett, et al. 2018. Fake news vs satire: A dataset and analysis. In *Proceedings of the Web Science Conference*. 17–21.
- [47] Google. 2021. Certified Publishing Partner. <https://www.google.com/ads/publisher/partners/>.
- [48] Google. 2022. Publisher Policies. <https://support.google.com/publisherpolicies/answer/10502938>.
- [49] Google News Initiative. 2022. Battling Misinformation. <https://newsinitiative.withgoogle.com/dnifund/report/battling-misinformation/>.
- [50] Richard Gray. 2017. Lies, propaganda and fake news: A challenge for our age. <https://www.bbc.com/future/article/20170301-lies-propaganda-and-fake-news-a-grand-challenge-of-our-age>.
- [51] Mauricio Gruppi, Benjamin D Horne, and Sibel Adali. 2021. NELA-GT-2020: A Large Multi-Labelled News Dataset for The Study of Misinformation in News Articles. *arXiv preprint arXiv:2102.04567* (2021).
- [52] Catherine Han, Deepak Kumar, and Zakir Durumeric. 2022. On the Infrastructure Providers That Support Misinformation Websites. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM '22, Vol. 16)*.
- [53] Arvind Hickman. 2021. Advertisers spend \$2.6bn on misinformation websites, study finds. <https://www.campaignlive.co.uk/article/advertisers-spend-26bn-misinformation-websites-study-finds/1725293>.
- [54] Raymond Hill. 2021. uBlock Origin. <https://github.com/gorhill/uBlock>.
- [55] Austin Hounsel, Jordan Holland, Ben Kaiser, Kevin Borgolte, Nick Feamster, and Jonathan Mayer. 2020. Identifying disinformation websites using infrastructure features. In *10th USENIX Workshop on Free and Open Communications in the Internet (FOCI)*.
- [56] Disconnect Inc. 2022. Tracker Protection Lists. <https://github.com/disconnectme/disconnect-tracking-protection>.

- [57] NewsGuard Technologies Inc. 2022. NewsGuard - Combating Misinformation with Trust Ratings for News. <https://www.newsguardtech.com/>.
- [58] Global Disinformation Index. 2019. The Quarter Billion Dollar Question: How is Disinformation Gaming Ad Tech? <https://www.disinformationindex.org/research/2019-9-1-the-quarter-billion-dollar-question-how-is-disinformation-gaming-ad-tech/>.
- [59] Global Disinformation Index. 2020. Ad Tech Fuels Disinformation Sites in Europe – The Numbers and Players. <https://www.disinformationindex.org/research/2020-3-1-research-brief-ad-tech-fuels-disinformation-sites-in-europe-the-numbers-and-players/>.
- [60] Check My Ads Institute. 2022. Check My Ads. <https://checkmyads.org/about/>.
- [61] Umar Iqbal, Peter Snyder, Shitong Zhu, Benjamin Livshits, Zhiyun Qian, and Zubair Shafiq. 2020. Adgraph: A graph-based approach to ad and tracker blocking. In *IEEE Symposium on Security and Privacy (SP '20)*. IEEE, 763–776.
- [62] Mark Di Stefano Javier Espinoza. 2020. Fake news websites still profit from Google advertising. <https://www.ft.com/content/5f8a405c-c132-4d9b-a86f-c52884535f3e>.
- [63] Kate Kelland. 2020. Fake news makes disease outbreaks worse, study finds. <https://www.reuters.com/article/us-health-fake-idUSKBN208028>.
- [64] Erin Kenneally and David Dittrich. 2012. The menlo report: Ethical principles guiding information and communication technology research. Available at SSRN 2445102 (2012).
- [65] Jisu Kim, Jihwan Aum, SangEun Lee, Yeonju Jang, Eunil Park, and Daejin Choi. 2021. FibVID: Comprehensive fake news diffusion dataset during the COVID-19 period. *Telematics and Informatics* 64 (2021), 101688.
- [66] Nir Kshetri and Jeffrey Voas. 2017. The Economics of “Fake News”. *IT Professional* 19, 6 (2017), 8–12. <https://doi.org/10.1109/MITP.2017.4241459>
- [67] IAB Technology Laboratory. 2019. ads.txt Specification Version 1.0.2. <https://iabtechlab.com/wp-content/uploads/2019/03/IAB-OpenRTB-Ads.txt-Public-Spec-1.0.2.pdf>.
- [68] IAB Technology Laboratory. 2019. sellers.json Specification. https://iabtechlab.com/wp-content/uploads/2019/07/Sellers.json_Final.pdf.
- [69] Emanuel Landau. [n.d.]. World without Cancer: the Story of Vitamin B17. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1653400/>.
- [70] Alexander Lawrence and Craig Silverman. 2016. How Teens In The Balkans Are Duping Trump Supporters With Fake News. <https://www.buzzfeednews.com/article/craigsilverman/how-macedonia-became-a-global-hub-for-pro-trump-misinfo>.
- [71] 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* (2018).
- [72] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczyński, and Wouter Joosen. 2019. Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation. In *Network and Distributed System Security Symposium (NDSS '21)*.
- [73] Yevgeniya Li, Jean-Grégoire Bernard, and Markus Luczak-Roesch. 2021. Beyond clicktivism: What makes digitally native activism effective? An exploration of the sleeping giants movement. *Social Media+ Society* 7, 3 (2021).
- [74] Zhou Li, Kehuan Zhang, Yinglian Xie, Fang Yu, and XiaoFeng Wang. 2012. Knowing your enemy: understanding and detecting malicious web advertising. In *Proceedings of the Conference on Computer and Communications Security*.
- [75] Similarweb LTD. 2023. Website Traffic. <https://www.similarweb.com/>.
- [76] Taichi Murayama. 2021. Dataset of Fake News Detection and Fact Verification: A Survey. *arXiv preprint arXiv:2111.03299* (2021).
- [77] International Fact-Checking Network. [n.d.]. IFCN Code of Principles - Check Your Fact. <https://ifcncodeofprinciples.poynter.org/application/public/check-your-fact/16EBE6DB-6072-CE51-EDC0-0D6347FF9605>.
- [78] NewsGuard. 2020. Tracking Facebook’s COVID-19 Misinformation “Super-spreaders”. <https://www.newsguardtech.com/special-reports/superspreaders/>.
- [79] Kathleen A Nolan and Jill E Callahan. 2006. Beachcomber biology: The Shannon-Weiner species diversity index. In *Proc. Workshop ABLE*, Vol. 27. 334–338.
- [80] Lukasz Olejnik, Tran Minh-Dung, and Claude Castelluccia. 2014. Selling off privacy at auction. In *21st Annual Network and Distributed System Security Symposium* (San Diego, California, USA) (NDSS '14).
- [81] OpenSources. 2017. Sources. <https://github.com/BigMcLargeHuge/opensource/blob/master/sources/sources.csv>.
- [82] 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)*.
- [83] Emmanouil Papadogiannakis. 2022. Scrape Titan. <https://gitlab.com/papamano/scrape-titan>.
- [84] Emmanouil Papadogiannakis. 2023. Open-source Datasets. <https://gitlab.com/papamano/who-funds-misinformation>.
- [85] 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 Web Conference 2021 (Ljubljana, Slovenia) (WWW '21)*. Association for Computing Machinery, 12.
- [86] Emmanouil Papadogiannakis, Panagiotis Papadopoulos, Evangelos P. Markatos, and Nicolas Kourtellis. 2022. Leveraging Google’s Publisher-Specific IDs to Detect Website Administration. In *Proceedings of the ACM Web Conference 2022 (Virtual Event, Lyon, France) (WWW '22)*. Association for Computing Machinery, New York, NY, USA, 2522–2531. <https://doi.org/10.1145/3485447.3512124>
- [87] Elias P. Papadopoulos, Michalis Diamantaris, Panagiotis Papadopoulos, Thanasis Petsas, Sotiris Ioannidis, and Evangelos P. Markatos. 2017. The Long-Standing Privacy Debate: Mobile Websites vs Mobile Apps. In *Proceedings of the 26th International Conference on World Wide Web (Perth, Australia) (WWW '17)*.
- [88] 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 (IMC '17)*. 142–156.
- [89] Parth Patwa, Shivam Sharma, Srinivas Pykl, Vineeth Gupta, Gitanjali Kumari, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. 2021. Fighting an infodemic: COVID-19 fake news dataset. In *International Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation*. Springer, 21–29.
- [90] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. (2018), 3391–3401.
- [91] Politifact. 2017. Fake News Almanac. <https://infogram.com/politifact-fake-news-almanac-1gew2vjdxl912nj>.
- [92] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A Stylometric Inquiry into Hyperpartisan and Fake News. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 231–240.
- [93] Columbia Journalism Review. 2021. Index of fake-news, clickbait, and hate sites. <https://web.archive.org/web/20210720140548/https://www.cjr.org/fake-beta>.
- [94] Caitlin M. Rivers and Bryan L. Lewis. 2014. Ethical research standards in a world of big. *F1000Research* 3 (2014). Issue 38. <https://doi.org/10.12688/f1000research.3-38.v2>
- [95] Jeff John Roberts. 2017. Hoax Over ‘Dead’ Ethereum Founder Spurs \$4 Billion Wipe Out. <https://fortune.com/2017/06/26/vitalik-death/>.
- [96] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *SIGKDD Explor. News* 19, 1 (sep 2017), 22–36. <https://doi.org/10.1145/3137597.3137600>
- [97] Sandra Siby, Umar Iqbal, Steven Englehardt, Zubair Shafiq, and Carmela Troncoso. 2022. {WebGraph}: Capturing advertising and tracking information flows for robust blocking. In *31st USENIX Security Symposium*. 2875–2892.
- [98] Craig Silverman, Jane Lytvynenko, Lam Thuy Vo, and Jeremy Singer-Vine. 2017. Inside The Partisan Fight For Your News Feed. <https://www.buzzfeednews.com/article/craigsilverman/inside-the-partisan-fight-for-your-news-feed>.
- [99] Milivoj Simeonovski, Giancarlo Pellegrino, Christian Rossow, and Michael Backes. 2017. Who controls the internet? analyzing global threats using property graph traversals. In *Proceedings of the Web Conference*. 647–656.
- [100] Alexander Sjösten, Peter Snyder, Antonio Pastor, Panagiotis Papadopoulos, and Benjamin Livshits. 2020. Filter list generation for underserved regions. In *Proceedings of The Web Conference (WWW '20)*. 1682–1692.
- [101] Snopes. 2016. Field Guide to Fake News Sites and Hoax Purveyors. <https://www.snopes.com/news/2016/01/14/fake-news-sites/>.
- [102] CHECK MY ADS Team. 2021. Kargo’s “no fake news” guarantee is fake news. <https://checkmyads.org/branded/kargos-no-fake-news-guarantee-is/>.
- [103] Magnite Team. 2020. On Content Standards. <https://www.magnite.com/blog/on-content-standards/>.
- [104] Pelayo Vallina, Victor Le Pochat, Álvaro Feal, Marius Paraschiv, Julien Gamba, Tim Burke, Oliver Hohfeld, Juan Tapiador, and Narseo Vallina-Rodriguez. 2020. Mis-shapes, mistakes, misfits: An analysis of domain classification services. In *Proceedings of the Internet Measurement Conference*. 598–618.
- [105] World Health Organization. 2021. Fighting misinformation in the time of COVID-19, one click at a time. <https://www.who.int/news-room/feature-stories/detail/fighting-misinformation-in-the-time-of-covid-19-one-click-at-a-time>.
- [106] Savvas Zannettou, Michael Sirivianos, Jeremy Blackburn, and Nicolas Kourtellis. 2019. The Web of False Information: Rumors, Fake News, Hoaxes, Clickbait, and Various Other Shenanigans. *J. Data and Information Quality* 11, 3, Article 10 (May 2019), 37 pages. <https://doi.org/10.1145/3309699>
- [107] Eric Zeng, Tadayoshi Kohno, and Franziska Roesner. 2020. Bad news: Clickbait and deceptive ads on news and misinformation websites. In *Workshop on Technology and Consumer Protection (ConPro '20)*.
- [108] Xichen Zhang and Ali A Ghorbani. 2020. An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management* 57, 2 (2020), 102025.
- [109] Xinyi Zhou, Apurva Mulay, Emilio Ferrara, and Reza Zafarani. 2020. reCOVary: A Multimodal Repository for COVID-19 News Credibility Research. In *Proceedings of the International Conference on Information and Knowledge Management*.
- [110] Xinyi Zhou, Reza Zafarani, Kai Shu, and Huan Liu. 2019. Fake News: Fundamental Theories, Detection Strategies and Challenges. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (Melbourne VIC, Australia) (WSDM '19)*. Association for Computing Machinery, 836–837.

A ETHICAL CONSIDERATIONS

The execution of this work has followed the principles and guidelines of how to perform ethical information research and use of shared measurement data [64, 94]. We keep our crawling to a minimum to ensure that we do not slow down or deteriorate the performance of any web service in any way, and make concerted effort not to perform any type of DoS attack to the visited website. Therefore, we crawl only the landing page of each website and visit it only once. We do not interact with any component inside a website, and only passively observe network traffic. Consequently, we emulate the behavior of a normal user that stumbled upon a website.

In accordance to the GDPR and ePrivacy regulations, we did not engage in collection of data from real users. Also, we do not share with any other entity any data collected by our crawler. We intentionally do not make our crawled dataset public (but only the fake and real news lists), to ensure that there is no infringement of copyrighted material from any website.

Finally, regarding the ad detection methodology, we were cautious not to affect the advertising ecosystem or deplete advertiser budgets. The development and testing of our methodology was performed on offline captures of websites. Additionally, for each website we process, we visit only the landing page and “click” on advertisements only once.

B AD NETWORKS

Complementary to the analysis of Section 4.1, we examine how many fake news websites have a RESELLER relationship with the ad networks studied so far. A RESELLER business relationship expresses cases where a third party has been authorized to control the ad space [67]. Table 3 presents the results. We find that 67.71 - 73.73% of fake news websites in our dataset have a RESELLER relationship with *appnexus.com*, *openx.com*, *rubiconproject.com*, *indexexchange.com*, and *pubmatic.com*. We note that these percentages reported in Table 3) are even higher than those reported in Figure 5. For example, although as many as 52.5% of fake news websites engage in a DIRECT relationship with *appnexus.com*, an even higher percentage of them (73.73%) engage in a RESELLER relationship with it. The same trend is true for the rest of the ad networks, which means that roughly six out of ten fake news websites have RESELLER relationships with the major ad networks.

Real News		Fake News	
Service	Portion	Service	Portion
<i>appnexus.com</i>	86.92%	<i>appnexus.com</i>	73.73%
<i>openx.com</i>	85.32%	<i>rubiconproject.com</i>	69.19%
<i>rubiconproject.com</i>	85.00%	<i>pubmatic.com</i>	68.68%
<i>indexexchange.com</i>	85.00%	<i>spotxchange.com</i>	67.67%
<i>pubmatic.com</i>	84.37%	<i>spotx.tv</i>	67.17%

Table 3: Most popular ad networks with RESELLER relationships. Publishers authorize intermediary entities to operate their accounts.

C CLUSTER COMPOSITION

For the construction of the Metagraph we make use of the detected Publisher-specific IDs (Table 4). However, very large communities of websites may arise due to the presence of intermediary publishing partners. These are third-party services that help publishers manage their websites and increase website popularity, and consequently

generate more revenue. In this work, we focus only on identifiers which can be found in more than 1 but at most 50 websites. We use this threshold based on the analysis of [86], declaring these as *Small* and *Medium* classes of website administrators. Larger clusters are considered intermediary publishing partners and not an actual administrator or an owner [47]. The use of these two classes in the Metagraph eliminates the issue of intermediary partners.

Description	Unique Identifiers	Unique Domains of landing URLs	% successful websites
Publisher IDs	642	872	26.34
Tracking IDs	2,638	2,365	71.43
Measurement IDs	393	584	17.64
Container IDs	1,113	1,221	36.88

Table 4: Detected Publisher-specific IDs in news websites.

In order to detect communities of websites operated by the same entity, we employ the Louvain community detection algorithm [5]. We perform hierarchical clustering by successive instances of the algorithm, and extract a dendrogram, where each level is a partition of the metagraph nodes. Level 0 contains the smallest communities while moving to higher levels results to bigger communities.

Table 5 summarizes the detected communities of websites operated by the same entity. We observe that for higher levels of the dendrogram (i.e., levels 1 and 2), fake news clusters contain thousands websites, and each such cluster is very big in size (i.e., 50.43 for level 2). Communities in higher levels are formed due to the presence of intermediary publishing partners that control hundreds or even thousands of websites and according to [86], do not indicate a clear co-administration relationship. Thus, we focus only on the first level of the dendrogram, containing small and more accurate communities. We find 73 fake news clusters that remain identical across different dendrogram levels.

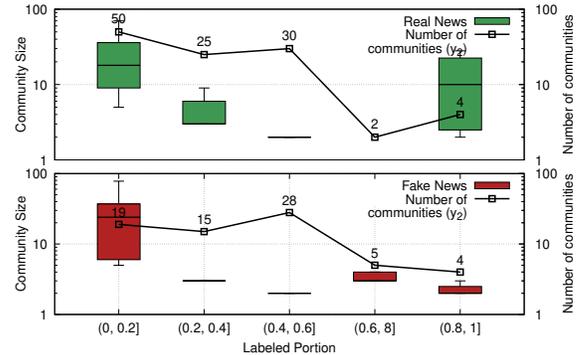


Fig. 12: Distribution of sizes of detected clusters and portion of websites labeled as fake or real news. Real news sites cluster more together (green rectangle, range (0.8, 1]).

For each identified cluster (Fake news and Real news) we compute the portion of its websites labeled (fake or real, respectively) based on our lists. For example, a portion of 0.5 for a Fake news cluster indicates that half of the websites in the cluster were labeled as fake news. In Figure 12, we illustrate the size of the detected fake news and real news clusters, as well as their labeled portions. We see that both Fake and Real news clusters behave similarly apart from the (0.8, 1] range. Indeed, in that range, large communities of real news (the big green rectangle at right) show that we have Real

Level	Number of communities	Fake News clusters	Websites in clusters	Average Cluster size
0	32,961	108	883	8.17
1	30,832	95	2,843	29.92
2	30,687	91	4,590	50.43

Table 5: Detected website communities using Louvain method.

news clusters of decent size where more than 80% of them are categorized as real news. This implies that such clusters contain several websites that disseminate credible information. On the contrary, for the Fake news clusters, in the same range (i.e., (0.8, 1)) we see that the red rectangle is very thin with a value close to two. This implies that fake news websites do not tend to cluster together - at least not as much as the real news websites do. These results are inline with the ones presented in Section 6.2.

D FAKE NEWS WEBSITES OWNERS

In this section, we provide some examples of fake news sites, their owners, and their ecosystem. We establish the entity by manually reviewing the copyrights claim, the privacy notice or the terms of services provided voluntarily by the websites themselves. We do not utilize external resources to ensure that any information about the ownership of a website is willingly provided by their administrators. We focus on smaller clusters, for which we have greater confidence about their miss-informative nature, as discussed in Appendix C.

For example, we find a cluster of 6 websites published by *Sophia Media*. 4 of these websites are part of the *Health Impact News Network*. These websites have been marked as “Pseudoscience websites” by MBFC, since they promote anti-vaccination propaganda and have multiple failed fact checks [26, 27, 29]. In fact, in 2020 *NewsGuard* [57], a journalism company that tracks online misinformation, identified Health Impact News as one of the greatest spreaders of COVID-19 misinformation on Facebook [78].

We also find a set of two websites owned and published by the National Vaccine Information Center (NVIC), an American organization for information about diseases and vaccines. One of these two websites has been marked as “Pseudoscience website” by MBFC since it promotes anti-vaccination propaganda and has multiple failed fact checks [30]. Health Impact News was also identified as one of the greatest spreaders of COVID-19 misinformation by NewsGuard [78]. Moreover, we find the websites *adfmedia.org* and *adflegal.org* being controlled by the same entity with the latter having been labeled as an extreme biased website due to propaganda [23]. The Alliance Defending Freedom (ADF) is a multi-million organization, which has also been classified as a hate group by the Southern Poverty Law Center (SPLC) [16].

Furthermore, we find that not only coordinated organizations, but also individuals are behind communities of fake news websites. Specifically, we find a pair of websites, *freedomforceinternational.org* and *needtoknow.news*, founded and powered by G. Edward Griffin. In his websites, he generally promotes right-wing beliefs, but also conspiracy theories and pseudoscience treatments [21]. Some of his beliefs about cancer treatment have been debunked by the American Journal of Public Health, since he promoted a banned chemical compound without any scientific evidence [69].

Ambiguous Website Ownership: Finally, we observe some contradicting communities that contain both a real (i.e., credible) and

a fake news website. These communities are not formed because of a clustering mistake, contrariwise, we show that specific entities operates both types of websites. First, we observe a community of three websites consisting of *checkyourfact.com*, *smokeroom.com* and *dailycaller.com*. CheckYourFact, an accepted signatory of the International Fact Checking Network [77], is considered by MBFC a credible fact checker with high factual reporting that utilizes proper sources and adheres to credible fact checking principles [24]. However, according to their *About* page, CheckYourFact is a news product of TheDailyCaller, a conservative news website that deliberately publishes misleading information and false stories [31].

In addition to this, we also find another contradicting community of 51 different websites. We manually visited and explored all of these websites and deducted that they belong to *Salem Media Group* and its subsidiaries. One of the websites in this community, *snnnews.com*, is part of our real news dataset since it has been rated HIGH for its factual reporting and has a clean fact check record [22]. In the same community, we also find *pjmedia.com*. This website is labeled as a questionable source since it displays extreme right-wing bias, it regularly promotes propaganda, as well as conspiracy theories, and it has published multiple false stories that failed fact checks [28].

E ADS IN FAKE NEWS WEBSITES

In Sections 4 and 5 we surprisingly discovered that only a portion of fake news websites display digital ads. This behavior was verified by both the ads.txt files served by the websites themselves, as well as by our ad detection methodology. To further understand this issue we manually investigated a random subset of 100 fake news websites in our list. We found that a big number of fake news websites do not display ads because they received funding from various other sources. For example, both *infowars.com* and *brighteon.com* have online stores, *21stcenturywire.com* and *navarreport.com* receive funding from publishing magazines and books respectively, while *cosmicintelligenceagency.com* provides paid webinars. We also discovered a great number of websites that sustain themselves by receiving money from their visitors either through donations (e.g., *canadafreepress.com* and *infiniteunknown.net*) or through paid memberships (e.g., *aapsonline.org*). Undoubtedly, there are also these fake news publishers that do not really care about monetizing their content but focus only on pushing their political or ideological agendas (e.g., *911truth.org* and *channel18news.com*). The sources of external funding, which fake news websites receive, are considered a different topic and left for future research.

From a technical point of view, we discovered that some websites do not display ads in their landing pages and only do so if you click on specific articles (e.g., *12minutos.com* and *24aktuelles.com*), while others display ads that come from static campaigns that stem from direct business contracts (e.g., *abovetopsecret.com*). Our methodology was not able to handle such cases. Finally, we found that a great portion of the evaluated websites is no longer active (e.g., *24wpn.com* and *embols.com*) and that some websites are no longer maintained and even though they contain ad scripts, these scripts no longer work and cannot fetch ads (e.g., *dcgazette.com*).