

An Analysis of the Partnership between Retailers and Low-credibility News Publishers

Lia Bozarth

University of Michigan, School of Information, U.S.A.

Ceren Budak

University of Michigan, School of Information, U.S.A.

Abstract

In this paper, we provide a large-scale analysis of the display ad ecosystem that supports low-credibility and traditional news sites, with a particular focus on the relationship between retailers and news producers. We study this relationship from both the retailer and news producer perspectives. First, focusing on the retailers, our work reveals high-profile retailers that are frequently advertised on low-credibility news sites, including those that are more likely to be advertised on low-credibility news sites than traditional news sites. Additionally, despite high-profile retailers having more resources and incentive to dissociate with low-credibility news publishers, we surprisingly do not observe a strong relationship between retailer popularity and advertising intensity on low-credibility news sites. We also do not observe a significant difference across different market sectors. Second, turning to the publishers, we characterize how different retailers are contributing to the ad revenue stream

of low-credibility news sites. We observe that retailers who are among the top-10K websites on the Internet account for a quarter of all ad traffic on low-credibility news sites. Nevertheless, we show that low-credibility news sites are already becoming less reliant on popular retailers over time, highlighting the dynamic nature of the low-credibility news ad ecosystem.

Keywords: low-credibility news, fake news, misinformation, display ads, online advertisement, online retailers

1. Introduction

Recent scholarship on combating disinformation focuses on five high-level approaches: i) legislative measures, ii) platform affordances, iii) community norms, iv) market forces, and finally v) individual learning. First, legal scholars (Butler, 2018; El-Khoury, 2020; Feingold, 2017; Kraski, 2017) argue that policy-makers can and should pass laws that criminalize hoaxes and misinformation that have a direct negative effect on individuals or the general public. Alternatively, online platforms can choose to implement various models to detect low-credibility news (Rubin et al., 2016; Shao et al., 2016; Wang, 2017; Boididou et al., 2018; Monti et al., 2019; Nguyen et al., 2020) and then provide flags or indicators to assist users in discerning reputable news from misleading content (Iosifidis and Nicoli, 2020; Geeng et al., 2020; Yaqub et al., 2020). Third, social and community norms are another powerful force that regulates behavior in online communities (Mele et al., 2017; Chandrasekharan et al., 2018). Fourth, many low-credibility news providers are profit-driven (Subramanian, 2017; Lazer et al., 2018; Munger, 2020). As such, market forces can disincentivize these providers by curtailing their revenue stream (Kshetri and Voas, 2017; Bakir and McStay, 2018; Braun and Eklund, 2019; Bozarth and Budak, 2020). Finally, raising an individual's aptitude including tech literacy and critical thinking abilities through education and interventions can reduce their susceptibility to low-credibility news (Jang and Kim, 2018; Jones-Jang et al., 2019; Roozenbeek et al., 2020).

In this study, we focus on market forces and examine the relationship between low-

credibility news sites and their advertisers. We first aggregate a list of 913 low-credibility news sites and 3.98K traditional news sites. We then scrape ads on their homepages daily using a browser in incognito mode for a span of approximately a year. Using this dataset, we study this relationship from both the retailer and publisher perspectives. We note that the term *retailer* (also referred to as *advertiser*) incorporates all individuals, firms, parties, and entities that advertise themselves, their products, services, and brand on news sites.

The first set of our analyses focus on the retailer perspective. Retailers benefit from advertising on low-credibility news sites by navigating potential customers to their websites through placed advertisements. A careful analysis of these benefits necessarily requires reliable data on how click-through rates vary across low-credibility and traditional news publishers. Given the lack of such reliable data, we do not attempt to perform this analysis. Instead, we focus on the potential *cost* of this relationship.

Specifically, past literature on brand and marketing suggests that consumers avoid retailers that they perceive to be socially irresponsible and unethical (Hoewe and Hatemi, 2017; Matos et al., 2017; Abu Zayyad et al., 2021). Indeed, several well-known advertisers had already faced brand concerns due to their partnership with low-credibility news sites (Braun et al., 2019; Berthon and Pitt, 2018). In our paper, we identify the retailers and attributes of retailers associated with a close connection with low-credibility news sites. We identify individual retailers, market segments, and retailer popularity ranges that are most closely related to low-credibility news sites. Most surprisingly, we find that an average high-profile retailer is very similar to an average low-profile retailer in the probability of appearing on low-credibility news sites, despite having more resources and incentive to invest efforts in brand management (Braun et al., 2019; Berthon and Pitt, 2018). In fact, our analysis of individual advertisers identifies numerous high-profile retailers that have a significantly higher odds of appearing on low-credibility news sites over traditional news sites. Specifically, we identify popular conservative-leaning websites including `donaldjtrump.com` and `usconcealedcarry.com`. More interestingly, we also see popular retailers including `amazon.com` and `ebay.com` are also disproportionately advertised on low-credibility new sites. While the latter pair of retail giants could potentially face consumer backlash, `donaldjtrump.com` and `usconcealedcarry.com` may be unlikely to experience brand concerns from their core consumer base. This is because conservatives

tend to share views congruent with low-credibility news publishers and also place more trust in these publishers (Verma et al., 2018; Budak, 2019; Calvillo et al., 2020). And, consumers also significantly prefer publishers and retailers with which they perceive to have shared ideals and beliefs (Gentzkow et al., 2015; Hoewe and Hatemi, 2017; Matos et al., 2017; Abu Zayyad et al., 2021). This differential cost for retailers may pressure certain types of retailers to cease collaborating with low-credibility news publishers while incentivizing others to retain this partnership. Results here suggest that brand-related cost for a given advertiser and its incentive to partnering with low-credibility news publishers is significantly dependent on its consumer base.

Next, we turn to the low-credibility news publishers' perspective. These news publishers, much like other content providers, commonly rely on ad revenue to stay profitable (Tamibini, 2017; Graham, 2017). Such ad revenue depends on three types of actors: *retailers* that pay content providers for placing their ads on their webpages, *ad firms* (e.g. Google DoubleClick) that act as intermediaries between retailers and content providers, and *news consumers* whose browsing behaviors dictate monetary flow in this marketplace. Past work (Bozarth and Budak, 2020) demonstrates that the vast majority of low-credibility news sites depend on only a handful of credible ad servers to generate ad revenue. Thus, *ad firms*, i.e. the intermediaries, that own these ad servers blacklisting low-credibility news sites could lead to a significant reduction in ad revenue for these news sites. Retailers can also take similar actions against low-credibility news sites. For instance, in the first two months of 2017, the number of retailers on Breitbart dropped by 90% (Berthon and Pitt, 2018) due to brand concerns. As such, in this paper, we focus on the concentration of ad traffic support across retailers. This helps us identify which retailers are most beneficial to low-credibility news publishers. We find that low-credibility news sites are moderately reliant on high-profile advertisers. Though, the level of reliance is significantly less than their reliance on top credible ad servers (Bozarth and Budak, 2020). Further, with retailers and platforms increasingly taking action against such harmful publishers (Wingfield et al., 2016; Iosifidis and Nicoli, 2020; Berthon and Pitt, 2018), we also examine how this reliance has changed over time. We observe a sizable reduction in low-credibility news publishers' reliance on top retailers.

2. Data and Methods

Our primary analysis is based on daily scrapes of low-credibility and traditional news publishers collected using a Web emulator software. Sites were scraped in incognito mode to minimize personalized advertising. These scrapes were further parsed to identify the display ads served on news publisher domains. These collected data were then further processed to identify retailer popularity ranks and market sectors.

2.1 Identifying Publisher Types

To quantify the relationship between low-credibility news producers and retailers, we first identify the set of low-credibility news publishers. We use the aggregated list of low-credibility news sites provided by 5 distinct sources: i) *the Daily Dot* (2018), ii) *Media Bias/Fact Check* (2018), iii) *PolitiFact* (2017), vi) Allcott et al (2018), and v) Zimdars et al. (2016). Several of these sources differentiate between types of low-credibility news sites. For instance, Zimdars (2016) uses *fake* to label news sites that have published completely fabricated news articles, and *unreliable* or *clickbait* to label sites that are deemed unreliable and publish articles with unverified claims. Here, we group all domains into *low-credibility*.

To provide a richer context, we also identify the set of traditional news sites. This allows us to identify the relationship between traditional news producers and retailers and determine which retailers, or types of retailers, publish *more* on low-credibility news sites compared to traditional news. Here, we use the combined list of traditional news sites from two sources: i) *Media Bias/Fact Check* (2018), and ii) Vargo et al (2018). We refer readers to related work (Grinberg et al., 2019; Bozarth et al., 2020) for details of these lists. In total, the aggregation results in 1.64K low-credibility news sites and 4.01K traditional news sites. Next, previous work shows that majority of the labeled low-credibility news sites are no longer active. For instance, 68.9% of all low-credibility news sites provided by *PolitiFact* were defunct by September 2019. At the time of writing, December 2020, 44.1% of all low-credibility news sites in the aggregated list are defunct. Additionally, 0.8% of all real news sites are also no longer active. Removing defunct domains results in 913 low-credibility news sites and 3.98K traditional news sites. Sites that had become defunct in the middle of our data collection are excluded from further analysis.

Name (Short Name)	Per- cent	Example subcategories	Top Retailers	Random Retailers
community &recreation (community)	22.9%	government agencies, non-profit organization, senior and child care, retirement homes, religion, church	usconcealedcarry.com, donaldjtrump.com, etrade.com	pgatoursuperstore.com, braddockfilms.com, whitney.org
other (other)	15.5%		amazon.com, ebay.com, iresults.com	peuterey.com, svlg.org, bookit.com
health & beauty (health)	10.1%	health, patient, skin care, dental, glasses, eye care	zennioptical.com, internalbeauty solution.com, arthrozene.com	innatewellnessaz.com, mainehealth.org, accel- eratedurgentcare.com
home (home)	9.5%	home renovation, home maintenance, home accessories, furniture, appliances, kitchen	foreverredwood.com, naturefreshpuri- fier.com, casper.com	autooneinc.com, bobs- discountmattresses.com, lgdconstruct.com
computer, science & technology (computer)	8.5%	computer, electronics, software, on- line services, large machinery, industrial systems	vimeo.com, adremover.org, comparisons.org	idg.com, fightwithlights.com, argusml.com
food & drink (food)	5.4%	food, grocery, wine, coffee, tea, organic	plummarket.com, vusevapor.com, angeli- nos.com	riunitecans.com, smithflathouse.com, broadwaybean.com
banking & finance (finance)	5.2%	bank, financial services, financial products, life insurance, health insurance, investment	lowermybills.com, quickenloans.com, americanexpress.com	elconfidencial.com, fmsbank.com, boneandbailey.com
vehicle (auto)	4.6%	car, truck, motorcycle, small auto- motive vehicle, dealership, tire	stuffanswered.com, carfax.com, carsge- nius.com	tacomasubaru.com, whitemanchevrolet.com, proformparts.com
school (school)	3.8%	college, campus, university, learn- ing, high school, academic	uc.edu, purdueglobal.edu, on- linevirginia.net	portobelloinstitute.com, laurel.edu, pnwboces.org
news media (news)	3.3%	news, journalism, blog, political commentaries, entertainment news, science news and journal	cbs.com, popcornnews.com, aarp.org	1430wcmy.com, traveller24.com, wglt.org
real estate (real estate)	3.0%	real estate, property management, commercial properties, home auc- tion, condos, apartment complex	internationalliv- ing.com, ups.com, kingston.com	michaelsaunders.com, myhome.ie, eppendorf.com
apparel (apparel)	2.7%	clothing, apparel, shoes, shirts, bag, dress	nike.com, menswearhouse.com, atmtee.com	shoplovestitch.com, voisins.com, unified- manufacturing.com
legal services (law)	2.4%	lawyer, attorney, law firm, legal practice, personal injury, car acci- dent	jeffdavislawfirm.com, legalshield.com, samhenrylaw.com	sellersandmitchell.com, hearingsolution- softx.com, pattersonlegalgroup.com
jewelry & special occasions (jewelry)	1.6%	jewelry&watch, holidays, special oc- casions, wedding, christmas, invita- tion	birchgold.com, jewelryexchange.com, brilliantearth.com	lightninglabels.com, juliesgraphics2006.com, arlingtonwatches.com
farm, garden & animals (farm)	1.3%	garden,farm, plant, seed, tree, dog	provenwinners.com, viralsharks.net, dogfoodexpose.com	aerogarden.com, nwf.org, greenlifewaco.com

Table 1: Topic Modeling Results. Sectors are sorted by the column “percent” which indicates the fraction of retailers that belong to a given market sector (e.g., the sector *community* constitutes to 23% of all retailers.). Example subcategories of retailers in a sector are determined through top keywords. Finally, we provide the top 3 retailers for each sector and three randomly sampled example retailers.

2.2 Identifying Advertisements on News Publisher Sites

We first use the Selenium WebDriver API (Avasarala, 2014) to scrape ad-related URLs from news sites. An URL is *ad-related* if its domain is a known ad server (see details in Section 1 in Supplementary Materials). We then obtain the corresponding retailer from the URLs using regular expression matching and URL redirects. The former approach extracts the retailer directly from an ad URL ¹. The latter entails first making a *URL get request* using the ad URL and then obtaining the retailer from the redirected landing URL ². Please see the Supplementary Materials (Section 1) for a detailed description of this process. Note that direct advertising where the retailers work directly with publishers is not captured using our approach. An example of direct advertising is the retailer **adobe.com** employs **nytimes.com** to promote its brand through a sponsored article. For reference, according to **emarketer.com**, a leading market research company, direct advertising accounts for approximately 31% of all display ad spending³.

We collect data for ads and retailers through 2 different time periods. The first dataset is collected between 09/17/2019 and 12/02/2019, and the second between 03/13/2020 to 12/18/2020. Due to technical issues (e.g., the processes running out of memory) and human error (e.g., not responding to the issues on time), there are several time gaps in our data collection, notably between 10/15/2019 to 10/21/2019, 04/24/2020 to 06/01/2020, 06/14/2020 to 06/30/2020, and 07/26/2020 to 08/12/2020. As a whole, our data collection results in 8.3M ad-related URLs scraped from 640 low-credibility news sites and 2.8K traditional news sites. Of the 8.3M URLs, we are able to match 3.6M (or 43.1%) to 63.3K unique retailers (85.9% and 14.1% of the ads are obtained through regex matching and URL redirect respectively). Matched ads are more likely to be from ad servers that are credible and not listed in malware lists (Bozarth and Budak, 2020). Specifically, 46% of ads from credible ad servers are matched, compared to 29% of ads served by risky ad servers.

¹As an example: the corresponding retailer is **nike.com** for the ad URL <https://adclick.g.doubleclick.net/pcs/click?xai=AKA&urlfix=1&adurl=https://www.nike.com/>.

²As an example: typing the link <http://api.content.ad/Lib/TrackOutboundClick.aspx?wid=690055> into the browser redirects you to the landing URL **funnelwide.com**, which is the retailer.

³<https://www.emarketer.com/content/driven-by-social-native-accounts-for-nearly-two-thirds-of-display-ad-spend>

In addition, 47% of ads by the top-10 ad firms are matched. Finally, we observe that We observe that 30% of low-credibility news sites and 28% of traditional news sites are free of ads.

2.2.1 *Determining Retailer Popularity Ranks*

Past work demonstrates differences in retailer size and reliance on display ads (Budak et al., 2016). This suggests that there could be significant differences in reliance on display ads served on low-credibility and traditional news sites. To determine this effect, we approximate a retailer’s online popularity using the list of 10 million top-ranked websites provided by Open PageRank Online Tool ⁴. We see that the most popular retailer in our dataset is `amazon.com`, which has a rank of 25. Additionally, 41.5% of retailers are not listed within the top 10 million sites and are assigned *rank* $\geq 10M$.

2.2.2 *Constructing Retailer Market Sectors*

Similar to past work (Budak et al., 2016), we use topic modeling to assign each retailer into its corresponding market sector (e.g., the website `toyota.com` belongs to the sector *auto*) as follows: We first scrape main text content from each retailer’s homepage and use the text content as input to train a semi-supervised LDA model (Jagarlamudi et al., 2012). This model assigns each retailer to its most probable market sector (i.e., topic). Crowdsourced evaluation of our approach on Amazon Mechanical Turk shows an F1 score of 0.73. This score is comparable to an average crowd worker’s ability to detect market sectors. The process is described in Section 1 in Supplementary Materials and resulting market sectors are summarized in Table 1. As shown, there are 14 distinct sectors. The market sector *community & recreation* (or, *community* for short) is the largest sector and constitutes $\approx 23\%$ of all retailers. Additionally, *farm, garden, & animals* (or, *farm* for short)

⁴See the list generation process at <https://www.domcop.com/top-10-million-websites>. Approximately 3 billion web pages are crawled in a time span of 7 years and then used to generate pagerank scores for online websites. The latest list of top 10M websites is generated on 10/18/2020.

constitutes $\approx 1\%$ of all retailers, and is the smallest market sector. Finally, approximately 16% of retailers are in the *other* category, which is the sector for uncategorized retailers and also retailers that may in fact belong to one of the 14 known sectors.

3. Results

We begin our empirical analysis by characterizing the relationship between news publishers and retailers from the retailers' perspective. We then discuss the publisher side of the market, with a particular focus on low-credibility news publishers. This latter analysis informs strategies for curbing disinformation online.

3.1 *Retailer-centric Analysis*

Our retailer-centric analysis centers on the branding costs for retailers to advertise on low-credibility news publishers. Indeed, extensive literature has addressed consumer activism within the online advertising space (Newman, 2004; Minocher, 2019). Retailers risk brand contamination and reduced profit when partnering with entities or support causes that consumers view as controversial or unethical (Swimberghe et al., 2011; Minocher, 2019). Within the field of low-credibility news, activist organizations have encouraged concerned consumers to boycott and publicly shame retailers that are promoting on sites like `breitbart.com` (Berthon and Pitt, 2018). As such, exposure on low-credibility news sites will likely cause a retailer to face consumer backlash and experience brand damage. Here, we operationalize exposure for each retailer i through two measures:

1. **low-credibility news advertising frequency** ($freq_i$) which represents the overall frequency that i had been promoted on low-credibility news sites.
2. **Disproportionate low-credibility news advertising** ($zscore_i$) low-credibility news advertising frequency can be driven by the overall advertising frequency of a given retailer. As an example, the retailer `samsung.com` has 615 ads on low-credibility news sites and 7.7K ads on traditional news sites, and the retailer `6dollarshirts.com` has ≈ 100 ads on low-credibility news sites and ≈ 200 ads on traditional news sites. While `samsung.com` has a larger absolute number of ads on low-credibility news sites, `6dollar-shirts.com` has a higher odds ratio. To account for this, we also compute a

measure of *disproportionate low-credibility news advertising* by using log-odds-ratio with informative Dirichlet priors. We then compute the z-scores of the log-odds-ratios to account for variance. A $zscore_i \geq 2$ indicates that i is significantly more likely to advertise on low-credibility news sites over traditional news sites. Likewise, $zscore_i \leq -2$ suggests i is significantly less likely to promote on low-credibility news sites.

We examine i) the *individual retailers* that are disproportionately advertising on low-credibility news sites and thereby endangering their brand. We then determine whether an advertiser’s ii) *market sector* and iii) *popularity* is associated with higher odds of being promoted on low-credibility news sites. We choose to examine *market sector* and *popularity* because prior work (Budak et al., 2016) demonstrates that retailer reliance on display ads is inversely correlated with the size of the market sector and the popularity of the retailer. Additionally, related work (Braun et al., 2019) also suggests that high-profile retailers have more resources readily available as well as higher incentives to protect their brand. As such, we hypothesize that larger market sectors and more popular retailers have lower exposure on low-credibility news sites.

3.1.1 *Individual retailer*

Focusing on advertising frequency, we observe that high-profile retailers including `amazon.com`, `donaldjtrump.com`, and `menswearhouse.com` are among those that have the highest advertising frequency on low-credibility news sites. For instance, Amazon has 12.6K ads on low-credibility news sites. This is a surprising finding. Further analysis shows that `prepperwebsite.com`, a low-credibility news site that advertises many survival-related supplies, alone accounts for 45.2% (5.7K) of Amazon’s ads on low-credibility news sites. It’s followed by the low-credibility news sites `patriotrising.com` and `breakingburgh.com` which account for 10.7% and 10.5% of Amazon’s ads on low-credibility news sites. In comparison, ads of `donaldjtrump.com` and `usconcealed-carry.com` are more equally distributed across low-credibility news sites. Indeed, the normalized Gini coefficient (Gini, 1921) for ads distribution is 0.94 for Amazon, but 0.61 and 0.63 for `donaldjtrump.com`

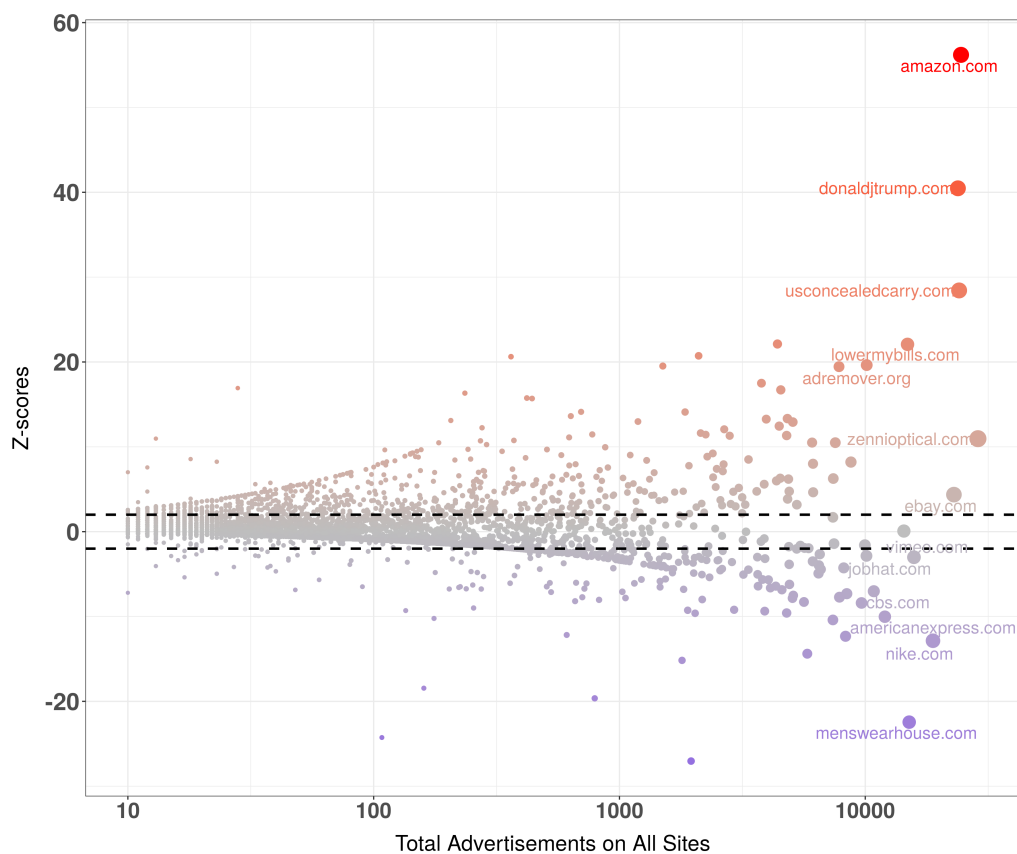


Figure 1. Top Retailers advertising on low-credibility news sites measured using log-odds-ratio with informative Dirichlet prior. The x-axis denotes a retailer’s total number of ads on low-credibility and traditional news sites. Only retailers with 10 ads or more are included in the plot. The y-axis is the z-scores of the log-odds-ratios with informative Dirichlet prior. Finally, dash lines mark $y = \pm 2$. Retailers with z-scores ≥ 2 are significantly more likely of being promoted on low-credibility news sites over traditional news sites.

and `usconcealedcarry.com` (we only include sites that have at least 1 ad for a given retailer when calculating Gini).

Next, we focus on disproportionate low-credibility news advertising measured using z-scores. Results are summarized in Figure 1. The x-axis denotes overall ad frequency (i.e., $freq_i$) and y-axis denotes the zscores of the log-odds-ratios (i.e., $zscore_i$). We also label the top retailers that are significantly more and less likely to promote on low-credibility news sites. As shown, retailers including `amazon.com`, `donaldjtrump.com`, and `usconcealedcarry.com` are retailers that have the highest z-scores, indicating that they

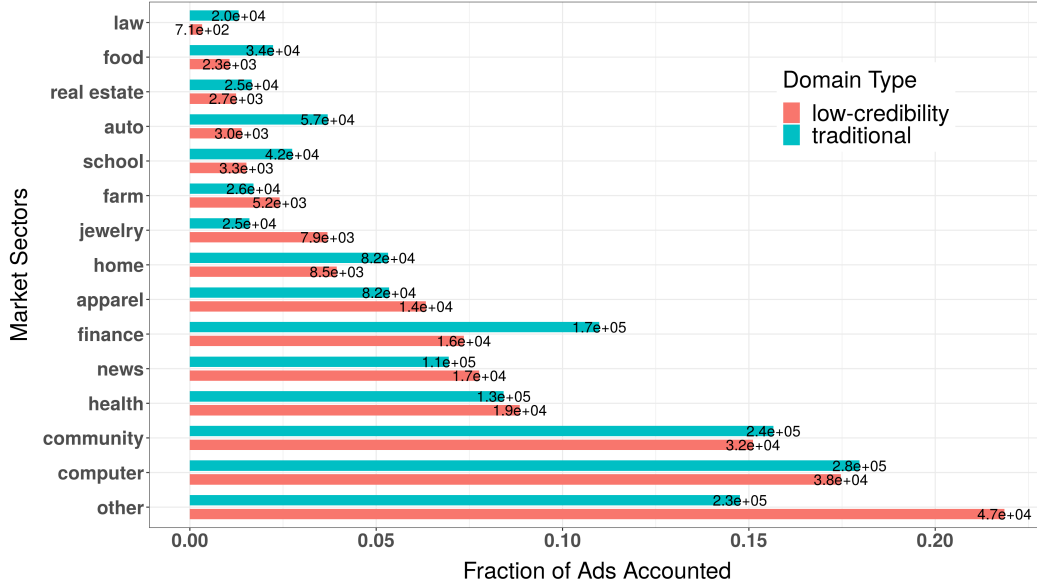


Figure 2. Advertising frequency accounted for by each market sector. Results show that *other*, *apparel* and *jewelry* retailers notably contribute to a higher fraction of ads on low-credibility news sites than on traditional news sites. In contrast, retailers in *finance*, *school*, *auto*, et cetera contribute to a lower fraction of ads on low-credibility news sites than on traditional news sites.

are significantly more likely to be promoted on low-credibility news sites. In Amazon’s case, excluding the low-credibility news publisher `prepperwebsite.com` stills results in its $z\text{-score} \geq 2$. In comparison, while retailers including `menswearhouse.com` and `nike.com` also appear frequently on low-credibility news sites (e.g., `menswearhouse.com` has over 2K ads on low-credibility news sites), their $z\text{-scores}$ are ≤ -2 , suggesting that they have a significantly lower probability than random chance to appear on low-credibility news sites.

3.1.2 Market Sectors

We first examine advertising frequency by deriving the total low-credibility news advertising frequency (*freq*) accounted by retailers from each of the 15 distinct market sectors. We also include the frequency distribution across sectors for traditional news sites for contextualization. Results are summarized in Figure 2. We see that the market sectors *other*, *computer* and *community* have the highest number of ads on low-credibility news sites.

The 3 sectors account for 22.0% (47.0K), 17.5%(37.5K), and 15.1% (32.5K) of total ads on low-credibility news sites respectively. Additionally, we also see that *other*, *jewelry*, and *apparel* account for a noticeably higher fraction on low-credibility new sites than they do on traditional news sites. Particularly, we see that *other* accounts for less than 15% of ads on traditional news sites, but approximately 22% on low-credibility news sites. Further analysis shows that the difference in *other* is mainly due to **amazon.com**. If we discount **amazon.com**, the remaining retailers in *other* account for 17.0% of ads on low-credibility news sites.

Next, we examine whether certain market sectors are disproportionately advertising on low-credibility news sites. As shown in Figure 3(a), we observe that an average retailer from *jewelry* or *apparel* has a higher z-score than an average retailer from *auto* or *finance*. Additionally, 6.9% of retailers in *apparel* have $zscore \geq 2$, the highest among all sectors. It's followed by *news* and *other* at 5.7% and 4.3%. However, the vast majority of retailers' z-scores are between $\{-2, 2\}$ regardless of market sector. For instance, 92.3% retailers in *apparel* have $-2 < zscore < 2$. We observe a similar pattern across the sectors. Results here suggest that there is no substantial difference between the market sectors in terms of preferentially promoting on low-credibility news sites.

For robustness check, we also reassign the retailers in the *other* sector to the remaining 14 well-defined sectors when possible. Specifically, if a *other* retailer's 2nd most probable sector is one of the 14 known sectors, we assign it to that sector instead. We rerun the previous analysis and observe that results are consistent (see details in Section 2.1 in the Supplementary Materials).

3.1.3 Retailer Popularity Tiers

Here, we first bin each retailer into the following 4 tiers according to its popularity rank (see Section 2 for how the popularity of a retailer is obtained and ranked): $\leq 1M$, $1M - 5M$, $5M - 10M$, and $10M+$. A retailer (e.g. **ebay.com**) belongs to the bucket $\leq 1M$ if it's ranked as one of the top 1 million most popular websites. We see that 22.1%, 20.8%, 15.6%, and 41.5% of the retailers belong to the tiers $\leq 1M$, $1M - 5M$, $5M - 10M$, and $10M+$ respectively.

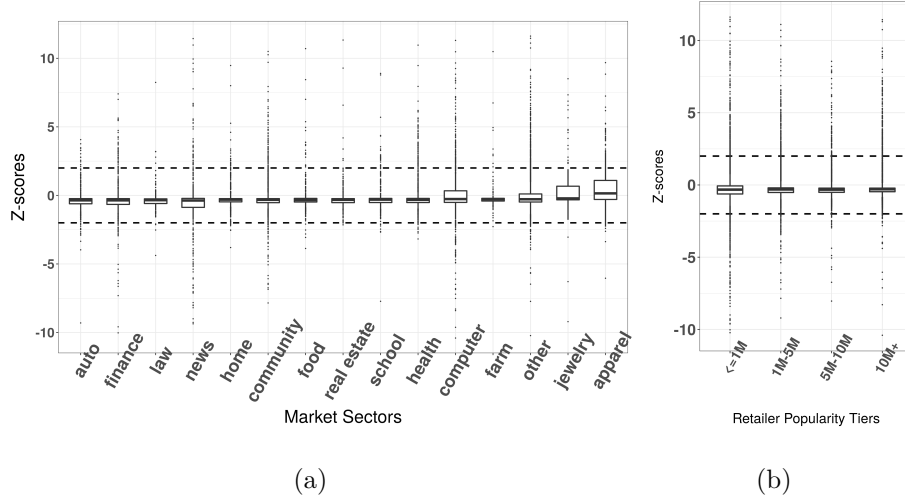


Figure 3. Figure 3(a) shows results for disproportionate low-credibility news advertising across different market sectors. The x-axis denotes the sectors, and the y-axis indicates the z-scores, and the dashed lines mark $y = \pm 2$. We see that an average *jewelry* or *apparel* retailer has a higher z-score than an average *auto* or *finance* retailer. More importantly, however, the vast majority of retailers have $-2 < zscore < 2$, suggesting that most retailers are not significantly more likely (or less likely) to promote on low-credibility news sites. Similarly, Figure 3(b) depicts the disproportionality results for retailers from different popularity tiers. Surprisingly, we see little difference between the tiers. This suggests that an average high-traffic retailer site’s likelihood of advertising on low-credibility news sites is similar to an average low-traffic retailer’s.

First, focusing on advertising frequency, we note that the $\leq 1M$ tier has a total of 114.4K (53.3%) ads on low-credibility news sites even though they constitute less than a quarter of all retailers. It’s followed by the 10M+ tier which accounts for 24.4% of all ads on low-credibility news sites. Finally, the tiers 1M – 5M and 5M – 10M account for 15.4% and 7.2% of ads on low-credibility news sites respectively.

Next, we also assess the disproportionality of advertising for retailers of different popularity tiers. We plot the z-scores of each tier of retailers in Figure 3(b). As shown, we again see that the vast majority of retailers from each tier have $-2 < zscore < 2$, thus they are not significantly more likely (or less likely) to advertise on low-credibility news sites than traditional new sites. In fact, an average high-popularity retailer ($\leq 1M$) is very similar to an average low-popularity retailer (10M+) in its odds of appearing on low-credibility news sites. This suggests that high-profile retailers are not taking more brand-management actions than low-profile retailers to avoid advertising on low-credibility news sites.

The analysis presented here examines the retailer rank and market sector separately. Models that account for these characteristics jointly, while controlling for other factors (e.g. number of unique publishers retailers advertised on), lead to similar qualitative findings and are presented in Supplementary Materials (Section 2.1 and Table 2).

3.1.4 *Implications*

What implications do our findings have for brand and marketing? First, we see that at the individual level, retailers such as `amazon.com`, `ebay.com`, `donald-jtrump.com`, and `usconcealedcarry.com` were disproportionately promoted on low-credibility news sites. For Amazon and eBay, this could potentially lead to a tarnished brand image and backlash from consumers. As a solution, these retailers may consider working with their ad servers to actively blacklist known low-credibility news sites. Yet, in the case of `donaldjtrump.com`, its brand is unlikely to face backlash from its core “consumers” given conservatives’ lower trust for mainstream media and tendency to share views align with low-credibility news publishers (Budak, 2019; Calvillo et al., 2020). In fact, prior work in marketing suggests that retailers whose consumer base match the population who frequent low-credibility news may be incentivized to and benefit from advertising on these sites (Matos et al., 2017; Kim et al., 2018). The difference in potential impact can even vary for apolitical retailers given how our apolitical interests can correlate highly with political behaviors and identities (DellaPosta et al., 2015).

Next, focusing on market sectors, we observe that *other*, *computer*, and *community* are the largest market sectors advertising on low-credibility news sites. The same sectors are also the largest on traditional news sites. Additionally, we observe that retailers from *apparel* and *jewelry* are more likely to advertise on low-credibility news sites than the other sectors, but the difference is small. Note that our data collection is performed in incognito mode. Therefore, our results show that there is no clear difference across sectors in contextual advertising—whether the results will look similar when accounting for news consumer types and targeted display advertising is an open question.

Finally, we expected high popularity retailers (e.g., `zennoptical.com`, `vimeo.com`)

to be more brand-conscious and have more capital/resources available to avoid being seen on low-credibility news sites. Yet, our results demonstrate that it's not the case. This could be due to various factors including: i) a lack of transparency on where retailers advertise their content, ii) insufficient consumer demands to disassociate the brand from the low-credibility news ecosystem, or iii) retailer concerns about weighing in on a politicized and polarized issue. We believe that quantitative work can bring more transparency to this ecosystem, addressing the first factor listed.

3.2 *Publisher-centric Analysis*

Here, we center our research on the issue of combating low-credibility news and examine low-credibility news sites' *ad dependence* on retailers. Past work (Bozarth and Budak, 2020) studying the advertising ecosystem⁵ of low-credibility news publishers demonstrates that two-thirds of low-credibility news sites rely on a handful of ad servers to generate ad revenue. The authors suggest that top ad firms who own these servers blacklisting low-credibility news sites can be an effective strategy to combat low-credibility news. Similar to this work, one might ask whether having i) individual retailers that low-credibility news sites rely on the most, or ii) retailers from the highest popularity tiers blacklist low-credibility news sites is a viable strategy to curtail low-credibility news sites' ad revenue. While answering this casually would necessitate additional data (e.g., news publishers' profit margins), here we quantify the degree to which these groups are currently supporting low-credibility sites. Finally, iii) we examine how low-credibility news publishers' dependence on retailers has changed over time⁶.

We measure low-credibility news site dependence on a group of retailers using two measures adopted from related work (Bozarth and Budak, 2020):

⁵We provide a simplified description of a typical advertising ecosystem. An online site may own 1 or more ad servers. When a user lands on one of the site's pages, each of its ad servers builds and sends out bidding requests to virtual ad marketplaces. Retailers on these marketplaces can then bid to have their ads displayed on the site. Note that ad servers are owned by ad firms (e.g., Google DoubleClick). Further, ad firms also provide tools to retailers for ad-filtering and ad-buying

⁶Note that market sector-based analysis is omitted given that we did not observe a clear difference between the sectors.

Panel A: The number and fraction of ads of each retailer that are displayed on different publishers.				Panel B: Monthly traffic for each publisher.	
	P1	P2	P3	P1	90K
R1	100 (50%)	0	0	P2	9K
R2	50 (25%)	75 (75%)	0	P3	1K
R3	50 (25%)	25 (25%)	10 (100%)		

Panel C: Weighted domain share $f(R1) = \frac{0.5 + 0 + 0}{3} = 0.17$ $f(R2) = \frac{0.25 + 0.75 + 0}{3} = 0.33$ $f(R3) = \frac{0.25 + 0.25 + 1}{3} = 0.50$		Panel D: Total weighted domain share on top-2 retailers: $f(R1) = 0.5$, $f(R2) = 0.33$, $R_2 = \{R1, R2\}$ $f(R_2) = 0.5 + 0.33 = 0.83$	
---	--	---	--

Panel E: Weighted traffic share $y(R1) = \frac{0.5 * 90k}{100k} = 0.45$ $y(R2) = \frac{0.25 * 90k + 0.75 * 9k}{100k} = 0.29$ $y(R3) = \frac{0.25 * 90k + 0.25 * 9k + 1 * 1k}{100k} = 0.26$		Panel F: Total weighted traffic share on top-2 retailers: $y(R1) = 0.45$, $y(R2) = 0.29$, $R_2 = \{R1, R2\}$ $y(R_2) = 0.45 + 0.29 = 0.74$	
--	--	--	--

Figure 4. A simple example using 3 low-credibility news publishers and 3 retailers to summarize the two measures of dependence. Panel A and B are data panels; ii) panel C and D show an example of weighted domain share; and iii) panel E and F show an example of weighted traffic share. Adopted from (Bozarth and Budak, 2020).

1. **Weighted Domain Share:** A retailer i 's weighted domain share, denoted as $f(i)$, is calculate as $f(i) = \frac{\sum_{j \in J} p_{j,i}}{|J|}$, where $p_{j,i}$ is the fraction of ads on j that's of i , and J is the set of low-credibility news sites and $|J|$ is the size of this set. See example on Figure 4 (Panel C). This metric informs a retailer's overall presence across all low-credibility news sites.
2. **Weighted Traffic Share:** The previous metric weighs all low-credibility news sites equally. Yet, popular sites such is [breitbart.com](#) and [info-wars.com](#) have a significantly higher viewer traffic (and therefore ad revenue) than fringe low-credibility news sites like [prepperwebsite.com](#). To account for viewer traffic difference, we calculate i 's weighted traffic share, denote as $y(i)$, using the formula $y(i) = \frac{\sum_{j \in J} p_{j,i} * w_j}{\sum_{j \in J} w_j}$ where w_j is low-credibility news publisher j 's viewer traffic (data is obtained from related work (Bozarth et al., 2020)). See an example of the measurement in Figure 4 (Panel E). Conceptually, a retailer that advertises on popular low-credibility news sites has a higher weighted traffic share than one that advertises on fringe/unpopu-

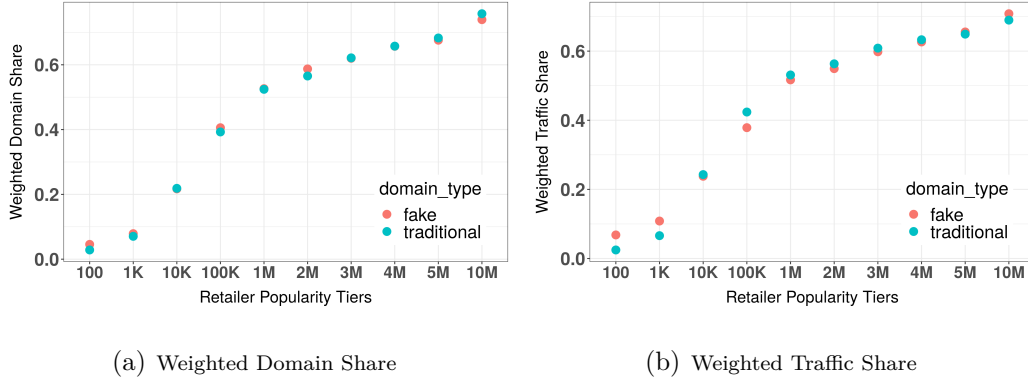


Figure 5. low-credibility New Publishers' Dependence on Retailers of Different Popularity Tiers. Figure 5(a) contains the result for weighted domain share. The y-axis is the cumulative weighted domain share accounted for by retailers that have a popularity rank of at least X. We see that retailers with popularity rank of $\leq 100K$ accounts for approximately 40% of weighted domain share. Likewise, Figure 5(b) contains the result for weighted traffic share.

lar low-credibility news sites, provided that they have comparable weighted domain share.

3.2.1 Dependence on Individual Retailers

Here, we examine how low-credibility news publishers could potentially be affected if the retailers that they collectively rely on the most choose to blacklist low-credibility news sites. We first rank individual retailers by their weighted domain share in descending order. That is, a retailer has *rank* = 1 if it has the highest weighted domain share. We observe that `donaldjtrump.com` has the highest weighted domain share, followed by `zennioptical.com` and `amazon.com`. Additionally, we also observe that the top-10 retailers (*rank* ≤ 10) account for 16.6% of weighted domain share. By comparison, the weighted domain share for `doubleclick.com`, the most popular ad server, is approximately 39%, more than double that of the top-10 retailers. Next, focusing on weighted traffic share, we see that `amazon.com` has *rank* = 1 followed by `ebay.com` and `urbanoutlit.com`. Further, the top-10 retailers contribute to 20.2% of weighted traffic share. We refer readers to the Supplementary Materials (Table 3) for details of the top-10 retailers. Furthermore, the top retailers ranked in this fashion include those that are not high profile (e.g., `mediaplayer10.com`) as well as

political advertisers (e.g. `donaldjtrump.com`). These advertisers might lack the motivation and/or resources to better structure their ad placement. As such, this analysis provides an informative upper bound on the impact of any k retailers can have as opposed to a strategy that is likely.

3.2.2 *Dependence on Retailers Across Different Popularity Tiers*

Prior reporting shows that many high-profile, popular retailers already express reluctance to partner with low-credibility news sites (Berthon and Pitt, 2018). We next determine low-credibility news sites' dependence on retailers of different popularity tiers. Results are summarized in Figure 5. We see that retailers with popularity rank $\leq 10K$ contribute to approximately 20% of both weighted domain share and weighted traffic share. Retailers with ranking $\leq 1M$ account for 52% of both measurements while constituting 22% of all retailers.

3.2.3 *Temporal Change in Dependence*

In this section, we examine whether there are temporal changes in dependence. Specifically, our two datasets (the first collected between Sep 2019 to Dec 2019 and the second collected between March 2020 to Dec 2020) allow us to analyze whether low-credibility news sites' reliance on high-profile retailers has increased or decreased over time. We again bin each retailer into the following 4 tiers: $\leq 1M$, $1M - 5M$, $5M - 10M$, and $10M+$. We then calculate low-credibility news sites' reliance, measured using weighted domain share and weighted traffic share, on retailers from each popularity tier per day and plot the results in Figure 6.

Focusing on weighted domain share in Figure 6(a), we see a steady drop in share for popular retailers in the tier of $\leq 1M$ and a steady rise of share for low popularity retailers in tier $10M+$. Using a simple regression (See Section 2.2 and Table 4 in Supplementary Materials), we observe that weighted domain share for the $\leq 1M$ tier drops by approximately 0.05% per day and increases for the $10M+$ tier by 0.04% per day. In consequence, weighted domain share for the $\leq 1M$ tier retailers reduces from $\approx 65\%$ to $\approx 50\%$, and increases

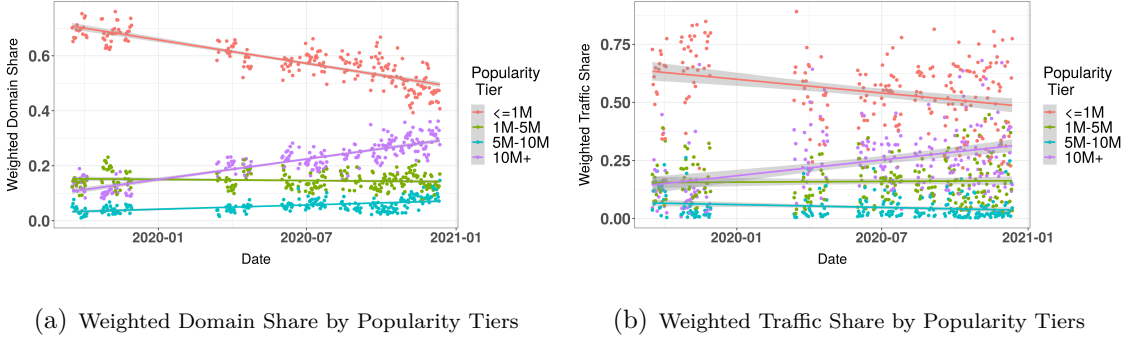


Figure 6. low-credibility News Publishers' Dependence on Different Tiers of Retailers over time. Figure 6(a) depicts the change over time in weighted domain share (y-axis) across different popularity tiers of retailers. We see that the share is decreasing for retailers in the tier $\leq 1M$, increasing for the tier $10M$, and stationary for the middle 2 tiers. Similarly, Figure 6(b) shows the change over time in weighted traffic share.

for the $10M+$ retailers from $\approx 10\%$ to $\approx 30\%$. Results here highlight that low-credibility news sites are becoming less reliant on high-profile retailers and more reliant on low-profile retailers. Next, focusing on weighted traffic share, we again see a drop in share for popular retailers and a rise for low popularity retailers. However, as shown in Figure 6(b), there is a higher variance (as indicated by the data points being more spread out) and the changes are less substantial. For instance, weighted traffic decreases by approximately 0.03% per day for the $\leq 1M$ tier (See Table 4 in Supplementary Materials), compared to a decrease of 0.05% per day for weighted domain share in the same time period. Finally, shares for the mid-level tiers $1M - 5M$ and $5M - 10M$ remain relatively stationary. Results here suggest that high-profile retailers are being promoted less on low-credibility news sites over time. Though, the progression is less substantial for high-traffic low-credibility news sites.

Thus far, we have treated our 2 datasets (the first is from Sep 2019 to Dec 2019 and the second is from March 2020 to Dec 2020) as a single time-series. Yet, there is a 4-month time gap between the 2 datasets. For robustness check, we also separate out these two datasets and use piecewise regression models to analyze temporal changes in low-credibility news publishers' dependence on popular retailers. We show that single time-series models are a better fit and that observations are also consistent between the models (see Table 4 in Supplementary Materials).

3.2.4 *Implications*

Activists and pundits alike have contended that corporations, especially the ones who are resourceful and whose products and services are ubiquitous, have a social responsibility or duty to fight misinformation and protect public interest (Creech, 2020). Results from this section suggest that doing so would have a moderate impact on low-credibility news sites' ad revenue. Specifically, if the top-10 individual retailers cease partnering with low-credibility news publishers, the publishers as a whole could stand to lose one-fifth of their total ad revenue (approximated using viewer traffic). Additionally, assuming that retailers with rank $\leq 10K$ are brand-conscious, incentivized, and easily persuadable ⁷, we again see that having these popular retailers block low-credibility news sites can be moderately effective. Next, we show that low-credibility news sites are becoming less dependent on high-profile retailers over time. This aligns with observations from related work (Berthon and Pitt, 2018; Iosifidis and Nicoli, 2020), which show that certain corporations and popular retailers had already taken actions to disassociate with low-credibility news publishers. If this trend continues, we may see a continued decline of low-credibility news sites' reliance on high-profile retailers. Though, as stated in Section 3.1, retailers vary significantly in design and intention. Specific retailers are likely to continuously partner with low-credibility news publishers without facing significant drawbacks. These problematic sites can also attract new high-profile retailers with particular characteristics. As such, results here only serve as the upper bound for how much retailer withdrawal could impact low-credibility news sites.

Finally, past work focusing on the role of ad servers (Bozarth et al., 2020) demonstrates that the top-10 credible ad servers alone account for half of ad revenue. In other words, low-credibility sites are significantly more dependent on the ad firms (e.g., Google) that own these ad servers than they do on retailers. This suggests that changes in behaviors of top ad firms, and not retailers, hold more promise in limiting low credibility news site revenue streams.

⁷For reference, `youtubekids.com` has a rank of $\approx 1K$, `hillaryclinton.com` and `barackobama.com` have a rank of $\approx 10K$, `walkingdead.wikia.com` has a rank of $\approx 100K$, and both `mustardsgrill.com` and `powerfulpython.com` have a rank of $\approx 1M$.

4. Discussion

In this work, our goal was to provide the first large-scale quantification of the relationship between low-credibility news publishers and retailers. We did so through a retailer and publisher-centric analysis.

Our retailer centric analysis reveals numerous high-profile retailers with an outsized affiliation with low-credibility news publishers. We also observe that retailers across market sectors and with varying scales are generally similar in their advertising intensity on low-credibility news domains. What does this tell us? The data examined in this study are limited to ads served by ad firms that act as intermediaries between publishers and retailers. In these ad networks, retailers generally do not have or employ significant autonomy in which publishers serve their ads, so long as ads are served effectively. The ad firms generally employ targeted advertising strategies. But, our data scraping is done using an incognito browser, thus limiting ad firms' ability for ad targeting. This partially explains why the behavior did not vary much across market sectors or sizes. Nevertheless, large retailers have marketing departments and access to resources to better control where their brand is being advertised. Therefore, it is worthwhile to note that such retailers are not, on average, using these resources to avoid affiliation with low-credibility news sites. Despite the ideological divide that also informs how consumers feel about disinformation and ways to address it (Minocher, 2019; van der Linden et al., 2020), there is enough consumer demand which in turn is leading retailers to signal their desire to disassociate with low-credibility news sites (Berthon and Pitt, 2018). So what explains the continuing presence of high-profile retailers on these problematic sites? One possibility is that these retailers are not aware that they are being promoted on low-credibility news sites (Braun and Eklund, 2019). We believe ongoing quantitative work can bring more transparency to the ad ecosystem that supports low-credibility news by providing such information. Another explanation is that consumer activism only addressed a small set of high-profile retailers that advertise on low-credibility news sites (Verma et al., 2018; Budak, 2019; Calvillo et al., 2020). Indeed, certain retailers (e.g., `donaldjtrump.com`) may face little to no backlash from their core consumers, thus have no incentive to disassociate with these publishers.

Next, through a publisher focused analysis, we observe that low-credibility news

sites are only moderately dependent on high-profile retailers. In comparison, the vast majority of low-credibility news sites are dependent on a handful of ad servers and their corresponding ad firms. As such, having ad firms blacklist low-credibility news sites is likely a better strategy for undercutting low-credibility news sites' ad revenue. Nevertheless, ad firms are, first and foremost, profit-driven entities. Thus, they are more likely to block low-credibility news publishers if such a need is demanded by their clients, the retailers. In other words, high-profile retailers may individually make the choice to dissociate with low-credibility news publishers and potentially have a moderate impact on these publishers' revenue. In contrast, retailers may also collectively pressure their partnering ad firms to block low-credibility news sites, which could have a more significant impact.

Lastly, focusing on temporal changes, we also see that low-credibility news sites' dependence on high-profile retailers has reduced significantly over time. A potential explanation is that advertising firms (e.g., Google and Facebook) and large retailers that may have frequently advertised on low-credibility news publishers in the past (e.g., Kellogg, Audi, Lyft) have been increasing their efforts to disassociate with low-credibility news sites (Berthon and Pitt, 2018). It's worth noting, however, that retailers' willingness to disassociate with low-credibility sites is complicated by their divergent interests and preferences. Certain retailers are incentivized to drop out and others buy into the low-credibility news ad ecosystem. One possible outcome is that the ad ecosystem will consolidate overtime along the alignments and preferences of their consumers. This, in turn, will potentially contribute to what prior work described as pluralistic collapse (DellaPosta et al., 2015; DellaPosta, 2020). Thus, it's crucial to conduct additional longitudinal studies in the future to observe meaningful changes.

There are several limitations to our study. First, our work doesn't differentiate between different subtypes (e.g., *junksci*, *satire*, *clickbait*) of low-credibility news publishers. Given that consumer response to a retailer being present on a satire site is likely going to be different from a *hate* site, future work should provide more granular analysis focusing on the subtypes. In addition, a significant portion of low-credibility news publishers are defunct, it's conceivable that the remaining publishers significantly differ from the ones that are no longer active (e.g., the operational publishers may have had more financial success with their sites). Further, the lists of low-credibility and traditional news sites used in our work

do not encompass all publishers. Moreover, the lists were generated using varied coding schemes by researchers and experts of different background.

Second, as mentioned above, our data only include ads served through ad firms that act as intermediaries between publishers and retailers. The cases where publishers allocate space to retailers directly on their pages (e.g., sponsored content) are not captured here. Moreover, data is collected using an empty user profile and incognito browser, as such we cannot address the characteristics of targeted ads. Further, our datasets were collected through a time period with many external shocks (e.g. covid crisis and the presidential election in United States). Additional longitudinal studies are needed to determine whether the trend we observed persists beyond these external shocks. For instance, high-profile ad firms and retailers may be especially cautious about partnering with low-credibility news publishers during an election year.

Third, we were unable to match a significant fraction of ad-related URLs to advertisers and the matched ads are more likely to be served by credible ad servers. It's possible that retailers employing risky ad servers significantly differ from those that operate through credible ad servers. Hypothetically, these retailers can be, on average, smaller and less well-known. As a consequence, low-credibility news publishers' reliance on high-profile retailers could potentially be lower than what's observed in our work. Future work that implements better ad collection pipelines will reduce potential biases.

Finally, our work simply quantifies the ad ecosystem as is. That is, our work is descriptive and not casual. We do not attempt to provide a counterfactual analysis on the strategies low-credibility news producers are likely to employ if the retailers blacklist them. We also do not claim that using the market, as opposed to other regulators (e.g. law), is the right strategy for limiting misinformation online.

Nevertheless, to the best of our knowledge, this is the first in-depth quantitative work that characterizes the connection between low-credibility news sites and the advertisers that they promote. We hope that results here can inform researchers who are interested in understanding the economics of low-credibility news, and assist parties who seek to leverage market forces to combat the fake news pandemic.

References

- Abu Zayyad, H. M., Obeidat, Z. M., Alshurideh, M. T., Abuhashesh, M., Maqableh, M., and Masa'deh, R. (2021). Corporate social responsibility and patronage intentions: The mediating effect of brand credibility. *Journal of Marketing Communications*, pages 1–24.
- Allcott, H., Gentzkow, M., and Yu, C. (2018). Trends in the diffusion of misinformation on social media. *arXiv preprint arXiv:1809.05901*.
- Avasarala, S. (2014). *Selenium WebDriver practical guide*. Packt Publishing Ltd.
- Bakir, V. and McStay, A. (2018). Fake news and the economy of emotions. *Digital Journalism*, 6(2):154–175.
- Berthon, P. R. and Pitt, L. F. (2018). Brands, truthiness and post-fact: managing brands in a post-rational world. *Journal of Macromarketing*, 38(2):218–227.
- Boididou, C., Papadopoulos, S., Zampoglou, M., Apostolidis, L., Papadopoulou, O., and Kompatsiaris, Y. (2018). Detection and visualization of misleading content on Twitter. *International Journal of Multimedia Information Retrieval*, 7(1):71–86.
- Bozarth, L. and Budak, C. (2020). Market forces: Quantifying the role of top credible ad servers in the fake news ecosystem. *Available at SSRN*.
- Bozarth, L., Saraf, A., and Budak, C. (2020). Higher ground? how groundtruth labeling impacts our understanding of fake news about the 2016 u.s. presidential nominees. *ICWSM*.
- Braun, J. A., Coakley, J. D., and West, E. (2019). Activism, advertising, and far-right media: The case of sleeping giants. *Media and Communication*, 7(4).
- Braun, J. A. and Eklund, J. L. (2019). Fake news, real money: Ad tech platforms, profit-driven hoaxes, and the business of journalism. *Digital Journalism*, 7(1):1–21.
- Budak, C. (2019). What happened? the spread of fake news publisher content during the 2016 us presidential election. In *The World Wide Web Conference*, pages 139–150.

- Budak, C., Goel, S., Rao, J., and Zervas, G. (2016). Understanding emerging threats to online advertising. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, EC '16, pages 561–578, New York, NY, USA. ACM.
- Butler, A. (2018). Protecting the democratic role of the press: A legal solution to fake news. *Wash. UL Rev.*, 96:419.
- Calvillo, D. P., Ross, B. J., Garcia, R. J., Smelter, T. J., and Rutchick, A. M. (2020). Political ideology predicts perceptions of the threat of covid-19 (and susceptibility to fake news about it). *Social Psychological and Personality Science*, 11(8):1119–1128.
- Chandrasekharan, E., Samory, M., Jhaver, S., Charvat, H., Bruckman, A., Lampe, C., Eisenstein, J., and Gilbert, E. (2018). The internet’s hidden rules: An empirical study of reddit norm violations at micro, meso, and macro scales. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–25.
- Couts, Andrew and Wyrich, Andrew (2018). Here are all the fake news sites to watch out for on facebook. [Online; accessed 27-October-2018].
- Creech, B. (2020). Fake news and the discursive construction of technology companies’ social power. *Media, Culture & Society*, page 0163443719899801.
- DellaPosta, D. (2020). Pluralistic collapse: The “oil spill” model of mass opinion polarization. *American Sociological Review*, 85(3):507–536.
- DellaPosta, D., Shi, Y., and Macy, M. (2015). Why do liberals drink lattes? *American Journal of Sociology*, 120(5):1473–1511.
- El-Khoury, M. (2020). Fake news: A legal perspective. *Fake News in an Era of Social Media: Tracking Viral Contagion*, page 149.
- Feingold, R. (2017). Fake news & misinformation policy practicum.
- Geeng, C., Yee, S., and Roesner, F. (2020). Fake news on facebook and twitter: Investigating how people (don’t) investigate. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

- Gentzkow, M., Shapiro, J. M., and Stone, D. F. (2015). Media bias in the marketplace: Theory. In *Handbook of media economics*, volume 1, pages 623–645. Elsevier.
- Gillin, J. (2017). Politifact’s guide to fake news websites and what they peddle.
- Gini, C. (1921). Measurement of inequality of incomes. *The Economic Journal*, 31(121):124–126.
- Graham, R. (2017). Google and advertising: digital capitalism in the context of post-fordism, the reification of language, and the rise of fake news. *Palgrave Communications*, 3(1):1–19.
- Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., and Lazer, D. (2019). Fake news on twitter during the 2016 us presidential election. *Science*, 363(6425):374–378.
- Hoewe, J. and Hatemi, P. K. (2017). Brand loyalty is influenced by the activation of political orientations. *Media Psychology*, 20(3):428–449.
- Iosifidis, P. and Nicoli, N. (2020). The battle to end fake news: A qualitative content analysis of facebook announcements on how it combats disinformation. *International Communication Gazette*, 82(1):60–81.
- Jagarlamudi, J., Daumé III, H., and Udupa, R. (2012). Incorporating lexical priors into topic models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 204–213. Association for Computational Linguistics.
- Jang, S. M. and Kim, J. K. (2018). Third person effects of fake news: Fake news regulation and media literacy interventions. *Computers in Human Behavior*, 80:295–302.
- Jones-Jang, S. M., Mortensen, T., and Liu, J. (2019). Does media literacy help identification of fake news? information literacy helps, but other literacies don’t. *American Behavioral Scientist*, page 0002764219869406.
- Kim, J. C., Park, B., and Dubois, D. (2018). How consumers’ political ideology and status-maintenance goals interact to shape their desire for luxury goods. *Journal of Marketing*, 82(6):132–149.

- Kraski, R. (2017). Combating fake news in social media: Us and german legal approaches. *. John's L. Rev.*, 91:923.
- Kshetri, N. and Voas, J. (2017). The economics of “fake news”. *IT Professional*, 19(6):8–12.
- Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., and Zittrain, J. L. (2018). The science of fake news. *Science*, 359(6380):1094–1096.
- Matos, G., Vinuales, G., and Sheinin, D. A. (2017). The power of politics in branding. *Journal of marketing theory and practice*, 25(2):125–140.
- Mele, N., Lazer, D., Baum, M., Grinberg, N., Friedland, L., Joseph, K., Hobbs, W., and Mattsson, C. (2017). Combating fake news: An agenda for research and action. *Retrieved on October*, 17:2018.
- Minocher, X. (2019). Online consumer activism: Challenging companies with change. org. *New Media & Society*, 21(3):620–638.
- Monti, F., Frasca, F., Eynard, D., Mannion, D., and Bronstein, M. M. (2019). Fake news detection on social media using geometric deep learning. *arXiv preprint arXiv:1902.06673*.
- Munger, K. (2020). All the news that’s fit to click: The economics of clickbait media. *Political Communication*, 37(3):376–397.
- Newman, K. M. (2004). *Radio active: Advertising and consumer activism, 1935-1947*, volume 13. Univ of California Press.
- Nguyen, V.-H., Sugiyama, K., Nakov, P., and Kan, M.-Y. (2020). Fang: Leveraging social context for fake news detection using graph representation. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1165–1174.
- Roozenbeek, J., van der Linden, S., and Nygren, T. (2020). Prebunking interventions based on “inoculation” theory can reduce susceptibility to misinformation across cultures. *Harvard Kennedy School Misinformation Review*, 1(2).

- Rubin, V., Conroy, N., Chen, Y., and Cornwell, S. (2016). Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News. In *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, pages 7–17, San Diego, California. Association for Computational Linguistics.
- Shao, C., Ciampaglia, G. L., Flammini, A., and Menczer, F. (2016). Hoaxy: A Platform for Tracking Online Misinformation. *Proceedings of the 25th International Conference Companion on World Wide Web - WWW '16 Companion*.
- Subramanian, S. (2017). Meet the Macedonian Teens Who Mastered Fake News and Corrupted the US Election. *Wired*.
- Swimberghe, K., Flurry, L. A., and Parker, J. M. (2011). Consumer religiosity: Consequences for consumer activism in the united states. *Journal of business ethics*, 103(3):453–467.
- Tamibini, D. (2017). How advertising fuels fake news. *Media Policy Blog*.
- van der Linden, S., Panagopoulos, C., and Roozenbeek, J. (2020). You are fake news: political bias in perceptions of fake news. *Media, Culture & Society*, 42(3):460–470.
- Van Zandt, D. (2018). Media bias/fact check (mbfc news) about.
- Vargo, C. J., Guo, L., and Amazeen, M. A. (2018). The agenda-setting power of fake news: A big data analysis of the online media landscape from 2014 to 2016. *new media and society*, 20(5):2028–2049.
- Verma, N., Fleischmann, K. R., and Koltai, K. S. (2018). Demographic factors and trust in different news sources. *Proceedings of the Association for Information Science and Technology*, 55(1):524–533.
- Wang, W. Y. (2017). "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. *arXiv:1705.00648 [cs]*. arXiv: 1705.00648.
- Wingfield, N., Isaac, M., and Benner, K. (2016). Google and facebook take aim at fake news sites. *The New York Times*, 11:12.

Yaqub, W., Kakhidze, O., Brockman, M. L., Memon, N., and Patil, S. (2020). Effects of credibility indicators on social media news sharing intent. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Zimdars, M. (2016). My “fake news list” went viral. but made-up stories are only part of the problem. *The Washington Post*.

Supplementary Materials

—An Analysis of the Partnership between Retailers and Low-credibility News Publishers

Lia Bozarth

University of Michigan, School of Information, U.S.A.

Ceren Budak

University of Michigan, School of Information, U.S.A.

Keywords: low-credibility news, fake news, misinformation, display ads, online advertisement, online retailers

1. Data

1.1 *Ads and retailer Data*

Data scraping and extraction is carried out in 3 steps:

Step1 (collect ad-related URLs): Here, we first use the Selenium WebDriver API (Avasarala, 2014), a tool that simulates the browser behavior of humans, to identify ad-related links present on low-credibility and traditional news sites. On a daily basis, for each news site,

our Selenium process starts a browser session in incognito mode with an empty user profile (note that an empty user profile ensures minimum ad targeting based on user browser history). Using the Selenium API, our automated script then navigates to the news site's homepage, scrolls through the page, and extracts ad-related URLs. A news site's homepage contains both ad-related URLs and non-ad URLs. We classify a URL as ad-related if it's served by any one of the 22.3K ad servers listed in *Easylist* and *Easylist Privacy* (<https://easylist.to>).

Step2 (separate ad-serving and ad-tracking URLs): Easylist and Easylist Privacy are the two most comprehensive and used sources for blocking unwanted web content such as ads. They include both ad trackers and ad servers. Ad trackers are only used to document and report user behavior and not to serve ads. Thus, We first filter out URLs that are from the top 50 most popular trackers. We also filter out URLs that are scripts (e.g., javascript and CSS), images (e.g., .png and .jpeg), videos (e.g., .svg, .mp4), and pixels. These URLs contain resources or are used for activity tracking, and they cannot be used to extract retailer information or navigate the consumer to the retailer page. We store the remaining URLs into files.

Step3 (extract retailers from ad-serving URLs): Next, in a separate process, we read the URLs from the files generated in the previous step and obtain retailers from these URLs using 2 approaches. We first use regular expression matching to extract retailers embedded in ad URLs. Specifically, we use the regex pattern:

“““=. {2,40} . (?:com|net|org|gov|edu|link|cn|uk|co%|ca)”””

That is, we match for the substring that starts with '=', has 2 to 40 characters in the middle before ending with '.com', '.net', et cetera. We ignore matches that start with 'ref=' or 'referral=', 'ex=', or 'source='¹.

For ads without a regular expression match, we then use the *requests-html* library

¹For example, our process extracts <https://www.nike.com> from <https://adclick.g.doubleclick.net/pcs/click?xai=AKA&urlfix=1&adurl=https://www.nike.com/&ref=nytimes.com> and ignore [nytimes.com](https://www.nike.com/) because the latter is the referral site.

to make *HTTP get requests* and obtain the landing pages and corresponding retailers². We note that this approach only works for URLs without embedded parameters³.

Ideally, the most effective way to collect ads and retailers is to simply use Selenium in *step2* to programmatically click on all the ad-related URLs, and then identify retailers based on the landing pages. However, this approach has significant memory requirements outside the scope of our project and resources.

Sector	school law		food ap- parel		auto jew- elry		comm- unity		fi- nance estate		real	farm	home	healthnews	com- puter
Model f1	0.86	0.85	0.85	0.84	0.84	0.80	0.78	0.77	0.76	0.74	0.71	0.71	0.68	0.64	
Human f1	0.70	0.80	0.90	0.76	0.88	0.68	0.75	0.75	0.72	0.76	0.70	0.78	0.62	0.66	

Table 1: Topic Modeling Model Performance. The row “Model f1” contains our algorithmic model’s F1 scores across the market sectors and the row “Human F1” contains an average crowd-source worker’s f1 scores across the market sectors.

1.2 Retailer Market Sector Classification

We use the following process to assign each retailer into its market sector.

Data Collection: First, we scrape the homepages of each of the 63.3K retailers and extract the main text content. We first use python’s native *requests* library to fetch each retailer’s homepage. Next, we extract the *title*, *navigation* (<nav> HTML tags), and *footer* text using *BeautifulSoup*, a commonly used HTML parser. We also extract the main content using the *trafilatura* (Barbaresi, 2019) library. Using *trafilatura* for main text extraction is better than directly concatenating all text from *body* together because it filters out noise text data (e.g. comments and invisible text).


Topic Modeling: We use the text content described above as input and build a guided-LDA model (Jagarlamudi et al., 2012) to perform the topic assignment. Past work (Budak et al., 2016) used vanilla LDA to perform a similar classification. Here, we use guided-

²As an example, our script will automatically request data from the ad URL <http://api.content.ad/Lib/TrackOutboundClick.aspx?wid=690055>, the request will be redirected to funnelwide.com, which is the retailer.

³For instance, the ad URL https://match.justpremium.com/match/spx?uid=SPOTX_ID contains the unknown parameter SPOTX_ID.

Website Categorization Instructions(Click to expand)

Which category (or categories) does the domain patrealty.ca belong to? (Please visit the domain homepage if image doesn't load)



Where would you like to look today?
Search a Street, City, Province, RP Number or MLS® Number

Price (Highest to Lowest) SEARCH / FILTER RESULTS

5224 results | Page 1 of 436 12 24 48

Please select the website's primary category and all applicable categories.

Primary Category	Relevant Categories	Category Description (we provide a few limited example subcategories for each broad category to help with the labeling task)
<input type="radio"/>	<input type="checkbox"/>	apparel: clothing, apparel, shoes, shirts, bag, dress, hat, jacket, shorts
<input type="radio"/>	<input type="checkbox"/>	banking&finance: bank, financial services, financial products, life insurance, health insurance, investment, mortgage loan, tax and accounting services
<input type="radio"/>	<input type="checkbox"/>	community: senior care, retirement homes, religion, church, funeral, memorial, assisted living, veteran care, family and child care, rehab, volunteer
<input type="radio"/>	<input type="checkbox"/>	computer,science&technology: computer, electronics, software, online services, large machinery, industrial systems, advanced technology and innovation
<input type="radio"/>	<input type="checkbox"/>	food&drink: food, grocery, wine, coffee, tea, organic, catering, cheese, meat
<input type="radio"/>	<input type="checkbox"/>	garden,farm&animals: garden,farm, plant, seed, tree, dog, animals, pets, horse, wildlife, fish, pest control
<input type="radio"/>	<input type="checkbox"/>	health&beauty: health, patient, skin care, dental, glasses, eye care, disease, hospital, surgery, pharmacy, hair, cosmetic, cannabis, vitamin, weight_loss, essential oil
<input type="radio"/>	<input type="checkbox"/>	home: home renovation, home maintenance, home accessories, furniture, appliances, kitchen, heating, ventilation, air conditioning, plumbing, roof repair, window installation, spa and pool
<input type="radio"/>	<input type="checkbox"/>	jewelry, arts&special occasions: jewelry&watch, precious metal and coins, holidays gifts, wedding, christmas, graduation, baby shower, fine arts, crafts, sewing, invitation, embroidery, stamps, ornament, bracelet, designer
<input type="radio"/>	<input type="checkbox"/>	legal services: lawyer, attorney, law firm, legal practice, personal injury, car accident, litigation, worker compensation
<input type="radio"/>	<input type="checkbox"/>	news&journalism: news media, journalism, blog, political commentaries, special interest media, entertainment news, science news and journal, radio station, tv station
<input type="radio"/>	<input type="checkbox"/>	non-profit: government agencies, non-profit organization, public services, public officials and candidates, employee associations
<input type="radio"/>	<input type="checkbox"/>	real estate: real estate, property management, commercial properties, home auction, condos, apartment complex
<input type="radio"/>	<input type="checkbox"/>	recreation,sports&travel: event tickets, sports, tours, travel, festival, concert, entertainment shows, film, theatre, live performance, recreational activities
<input type="radio"/>	<input type="checkbox"/>	school: college, campus, university, learning, high school, academic, education program, online education, training
<input type="radio"/>	<input type="checkbox"/>	vehicle: car, truck, motorcycle, small automotive vehicle, dealership, tire, vehicle repair, car racing, break, gear, engine
<input type="radio"/>	<input type="checkbox"/>	none of the above: if the website belongs to several categories but primarily belongs to one that's not listed, please select this as the primary category.
<input type="radio"/>	<input type="checkbox"/>	website inactive: use only if website is no longer available

Submit

Figure 1. Sample Amazon Mechanical Turk Task for Retailer Categorization.

LDA given past work that demonstrates better topic modeling performance for similar tasks (Bozarth et al., 2020). Guided-LDA uses sets of keywords to “nudge” document topic assignment. To generate the required sets of keywords, we first build base LDA models using *gensim*. We vary the topics numbers, $\{20, 21, 22 \dots 38, 39, 40\}$. Next, we pick the base model with the highest coherence score (O’callaghan et al., 2015), which is the model with 35 topics. We manually review the list of topics and the most weighted words in each topic; we then reassign these words into different sets according to coherent themes. This results in 24 sets of keywords. Finally, we run guided-LDA using the generated keywords. We again manually review the generated topics and identify 16 human-interpretable topics: $\{apparel, auto, community, computer, farm, finance, food, health, home, jewelry, law, news, other, public, recreation, real\ estate, school\}$. The remaining uninterpretable topics are collapsed into the *other* sector.

Evaluation: We use crowd-sourcing to assess the quality of the resulting sectors. We sample 50 retailers from each topic and 100 retailers from the *other* sector (a total of 900 retailers). We use crowd-sourcing, and Amazon Mechanical Turk (AMT) in particular, to assess the quality of the resulting sectors using these data. We assign 3 independent AMT workers to categorize each retailer. To ensure the quality of labels, we require workers to i) have completed more than 1000 tasks, ii) have an approval rate of at least 98%, and iii) reside in the United States. Workers are shown a screenshot of a retailer’s homepage, and are instructed to pick the primary market sector, as well as all applicable secondary market sectors for the retailer. See Figure 1 for a sample categorization task. If a majority vote is reached for the primary sector, the sector is then the ground truth label for the retailer. Our initial analysis of worker labels demonstrates that the 3 sectors *community*, *public*, and *recreation* are significantly interrelated. As such, we collapse them into a single *community* sector. We observe that 810 of the 900 retailers have a ground truth label, and the interrater reliability calculated using Krippendorff’s alpha is 0.63, suggesting considerable agreement.

We next compare the agreement between workers with the agreement between the classifier and workers, similar to (Rajadesingan et al., 2020) to evaluate our approach. We evaluate the performance using the F1-score measure through 3-fold cross-validation as follows: In each fold, we let one of the worker evaluations define the ground truth label. We compare the classifier assigned category label to the ground truth defined and, as a baseline

for comparison, check how well other workers predict the ground truth defined by the focal worker. The F1-score of the classifier and the baseline is 0.73. This suggests that our LDA model is performing as well as an average human. Performance for individual sectors is shown in Table 1. As shown, *school* has the highest f1 score of 0.86; *news* and *computer* has the lowest f1 scores at 0.68 and 0.64 respectively. The LDA model performs slightly better for the sectors *school*, *apparel*, *jewelry*, and an average human performs better for the sectors *food*, *auto*, and *health*.

2. Results

2.1 Retailer-centric Analysis

Robustness Check for Market Sector Analysis: We assign retailers in the *other* market sector (i.e., topics) to one of the 14 well-defined sectors using their 2nd most probable sectors if possible. To give an example, the domain *afreserve.com*’s most probable sector has a 0.55 probability, but this sector is not well-defined (i.e., topic modeling results lack coherence). If it also has a 0.25 probability of being in the sector *community* (the 2nd highest), then we assign it to *community* instead. However, if its 2nd most likely sector is also not well defined, it remains in the *other* sector. Using this approach, the number of retailers in *other* drops to 7.54K from the original 11.04K. We rerun market sector analysis using the updated sector assignments. As shown in Figure 2, we see that our initial results in the main paper are robust.

Disproportionate Low-credibility News Advertising Model: We run the following model to determine variables that are significant in predicting the z-scores of log-odds-ratios:

$$zscore_i = \beta_0 * sector_i + \beta_1 * rank_log_i + \beta_2 * ads_count_log_i + \beta_3 * domain_count + \epsilon \quad (1)$$

where $sector_i$ and $rank_log_i$ are retailer i ’s market sector and rank in log value; ads_count_i is i ’s total ad frequency on both low-credibility and traditional news sites, and $domain_count$ is the number of unique domains i advertised on. Finally, $days_fake_i$ and $days_traditional_i$ are the number of unique days that i advertised on low-credibility and traditional news sites respectively. Some of the independent variables are correlated, and therefore to address multicollinearity, we make sure that VIF values of all the independent variables are below 2.5 (Midi et al., 2010). Results are summarized in Table 2 (Model 1). We see that *apparel*

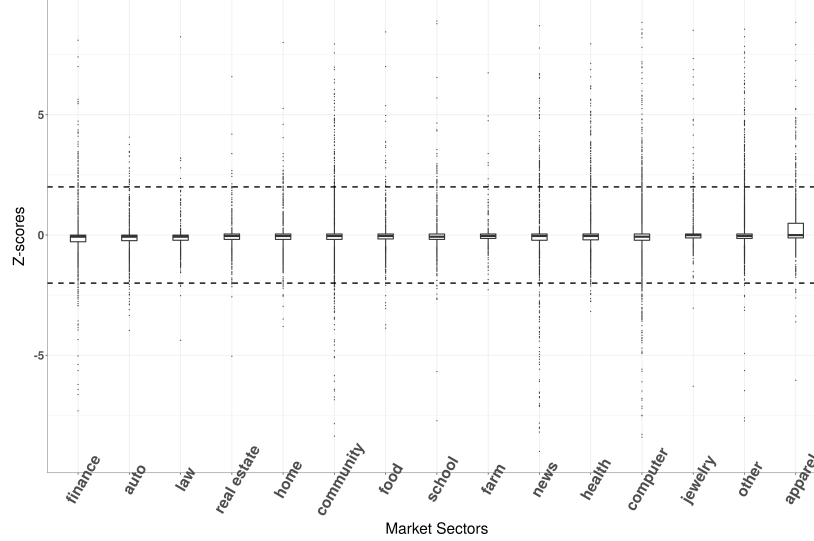


Figure 2. Updated results for disproportionate low-credibility news advertising across different market sectors. The x-axis denotes the sectors, and the y-axis indicates the z-scores, and the dashed lines mark $y = \pm 2$. Results here suggesting that a retailer’s odds of appearance on low-credibility news sites is not significantly correlated with its market sector

and *jewelry* sector retailer have significantly higher z-scores; however, the effect-sizes are small; and, a retailer’s popularity is insignificant in predicting its z-score.

Next, we redefine the dependent variable as $zscore_i^*$ with $zscore_i^* = 1$ if $zscore_i \geq 2$ (i.e., the retailer i is significantly more likely to be promoted on low-credibility news sites) else $zscore_i^* = 0$. In other words, model 2 measures a retailer’s likelihood of being significantly more likely to advertise on low-credibility news sites. We rerun the logistic regression model summarized in Equation 1. Results are depicted in Table 2 (Model 2) and are consistent with our prior observations.

2.2 Publisher-centric Analysis

Dependence on Top Retailers: The top-10 retailers with the highest weighted domain share or weighted traffic share are depicted in Table 3. Additionally, the cumulative fraction of weighted domain share and weighted traffic share accounted for by the top-10 retailers is 16% and 20% respectively.

	<i>Dependent variable:</i>	
	(Model 1) $zscore_i$	(Model 2) $zscore_i^*$
market sector		
apparel	0.239*** (0.046)	0.874*** (0.171)
auto	-0.191*** (0.046)	-1.102*** (0.337)
community	-0.162*** (0.024)	-0.797*** (0.144)
farm	-0.048 (0.068)	-0.058 (0.381)
finance	-0.236*** (0.036)	-0.682*** (0.207)
food	-0.134*** (0.035)	-0.467** (0.223)
health	-0.038 (0.031)	0.139 (0.148)
home	-0.112*** (0.038)	-0.571** (0.237)
jewelry	0.295*** (0.063)	0.598** (0.264)
law	-0.190*** (0.056)	-1.346*** (0.516)
news	-0.160*** (0.038)	0.363** (0.173)
other	0.043 (0.027)	0.316*** (0.121)
real estate	-0.119** (0.052)	-1.035*** (0.396)
school	-0.124*** (0.036)	-0.473** (0.216)
rank_log	0.003 (0.003)	0.001 (0.017)
domain_count	0.002*** (0.0002)	-0.001*** (0.0004)
ads_count_log	-0.089*** (0.005)	0.596*** (0.024)
Constant	0.187*** (0.052)	-4.932*** (0.260)
Observations	27,405	27,405
R^2	0.03	0.15

Note: *p<0.1; **p<0.05; ***p<0.01

Table 2: Regression Results for Disproportionate Advertising. The base category is *computer* which has the 2nd highest advertising frequency on low-credibility news sites (*other* being the highest). Model 1's dependent variable $zscore_i$ is the z-score of retailer i 's log-odds-ratios of advertising on low-credibility news sites over traditional news sites. In contrast, model 2's dependent variable $zscore_i^*$ is dichotomous: $zscore_i^* = 1$ if $zscore_i \geq +2$ and $zscore_i^* = 0$ otherwise. Conceptually, model 2 models a retailer's tendency to advertise significantly more often on low-credibility news sites. Focusing on Model 1, we see that on average *apparel* and *jewelry* retailers have a significantly higher z-score than average *computer* advertisers but the effect-size is small (e.g., *jewelry* retailers on average has 0.24 higher z-score). Finally, an advertiser's rank is insignificant in predicting its z-score. We observe similar results for Model 2. For instance, *apparel* retailers are 4.5% more likely to being disproportionately promoted on low-credibility news sites, and popularity rank is insignificant.

rank	retailer	weighted domain share	rank	retailer	weighted traffic share
1	donaldjtrump.com	0.027	1	amazon.com	0.033
2	zennioptical.com	0.024	2	ebay.com	0.029
3	amazon.com	0.020	3	urbanoutlit.com	0.025
4	usconcealedcarry.com	0.018	4	donaldjtrump.com	0.022
5	lowermybills.com	0.014	5	aarp.org	0.019
6	adremover.org	0.013	6	macys.com	0.019
7	ebay.com	0.011	7	yellowbook.com	0.015
8	duckduckgo.com	0.011	8	ponyo.com	0.013
9	wordads.co	0.009	9	selinc.com	0.012
10	mediaplayer10.com	0.008	10	zangdeal.com	0.010

Table 3: Top-10 Retailers Ranked by Weighted Domain Share or Weighted Traffic Share. For instance, we see that `donaldjtrump.com` has the highest weighted domain share of 2.7%.

Dependence Over Time Model Comparison: In this section, we determine the best fitting models for assessing low-credibility news sites’ dependence on retailers of different popularity tiers over time. Our data are collected through 2 separate time periods: i) 09/17/2019 to 12/02/2019, and ii) 03/13/2020 to 12/18/2020. We examine whether a single time-series approach (i.e., we assume that temporal trend is consistent across the two time periods) is better than a piecewise approach (i.e., we assume that trend differs between the two time periods). Here, we first write $f(k)_t$ as low-credibility news publishers’ dependence—measured using weighted domain share—on retailers from tier k at date t , where $k \in \{\leq 1M, 1M - 5M, 5M - 10M+\}$, and $t = 1$ on the date 09/17/2020 which is the start of the first time period. We then run the following single time-series model:

$$f(k)_t = \beta_0 * date + \beta_1 * k + \beta_2 * date * k + \varepsilon \quad (2)$$

Next, we use the following piecewise approach:

$$f(k)_t = \beta_0 * date + \beta_1 * k + \beta_2 * date * k + \beta_3 * is_dataset2 + \beta_4 * is_dataset2 * date + \varepsilon \quad (3)$$

where $is_dataset2 = 1$ for the second time period and $is_dataset2 = 0$ otherwise. Additionally $date2 = 1$ on the date 03/13/2020 which is the start of the second time period. Results for both approaches are summarized in Table 4 (Model 1 and 2). Here, we use AIC and BIC values to assess model fitness (Kuha, 2004). As shown, the AIC and BIC values

	Model 1 (weighted domain share -single time series)	Model 2 (weighted domain share -piecewise time series)	Model 3 (weighted traffic share -single time series)	Model 4 (weighted traffic share -piecewise time series)
is_dataset2		3.19e-02 (2.4e-02)		-5.57e-02 (8.45e-02)
date	-4.69e-04*** (1.79e-05)	-6.52e-04*** (1.66e-04)	-3.24e-04*** (6.29e-05)	2.21e-04 (5.84e-04)
tier_10M+	-6.01e-01*** (7.76e-03)	-6.01e-01*** (7.76e-03)	-4.9e-01*** (2.73e-02)	-4.9e-01*** (2.73e-02)
is_dataset2:date2		1.6e-04 (1.67e-04)		-6.32e-04 (5.89e-04)
date:tier_10M+	8.77e-04*** (2.53e-05)	8.77e-04*** (2.52e-05)	6.97e-04*** (8.9e-05)	6.97e-04*** (8.9e-05)
Constant	7.08e-01*** (5.49e-03)	7.13e-01*** (8.36e-03)	6.34e-01*** (1.93e-02)	6.08e-01*** (2.95e-02)
Observations	448	448	448	448
Log Likelihood	8.24e+02	8.25e+02	2.6e+02	2.61e+02
AIC	-1.64e+03	-1.64e+03	-5.12e+02	-5.1e+02
BIC	-1.62e+03	-1.61e+03	-4.91e+02	-4.8e+02
<i>Note:</i>	*p<0.1; **p<0.05; ***p<0.01			

Table 4: Low-credibility News Sites' Dependence Overtime Model Comparison. Model 1 (single time series) and 2 (piecewise time series) are results for weighted domain share. Models 3 (single time series) and 4 (piecewise time series) are results for weighted traffic share. Models are comparable in AIC and BIC values with the simpler single time-series models being a slightly better fit. Note that the base tier is $\leq 1M$ and results for the tiers $1M - 5M$ and $5M - 10M$ are omitted from the table due to their relative stationarity.

for both approaches are comparable. The simpler model has a slightly smaller AIC (and BIC), suggesting a better fit. Additionally, the coefficients for the terms *is_dataset2* and *is_dataset2*date2* in the piecewise approach are not significant. As a whole, results here suggest that the simpler single time-series approach is the better fit.

Similarly, we write $y(k)_t$ as low-credibility news sites' dependence—measured using weighted traffic share—on retailers from tier k at date t . We apply the regression equations 2 and 3 using $y(k)_t$ as the dependent variable. Results are summarized in Table 4 (Model 3 and 4). We again see that the AIC and BIC values are comparable for both models and that the coefficients for the terms *is_dataset2* and *is_dataset2*date2* are not significant. Additionally, we also see that the coefficient for *date* is no longer significant for the piecewise approach (Model 4), suggesting that there is no significant temporal change in weighted domain share for retailers in the $\leq 1M$ tier. This is not a significant deviation from the basic single time-series model, however, as the coefficient for the variable *date* suggests that the effect-size of *date* is small (Model 2).

References

- Avasarala, S. (2014). *Selenium WebDriver practical guide*. Packt Publishing Ltd.
- Barbarese, A. (2019). Generic web content extraction with open-source software. In *KONVENS 2019*, pages 267–268. GSCL.
- Bozarth, L., Saraf, A., and Budak, C. (2020). Higher ground? how groundtruth labeling impacts our understanding of fake news about the 2016 u.s. presidential nominees. *ICWSM*.
- Budak, C., Goel, S., Rao, J., and Zervas, G. (2016). Understanding emerging threats to online advertising. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, EC '16, pages 561–578, New York, NY, USA. ACM.
- Jagarlamudi, J., Daumé III, H., and Udupa, R. (2012). Incorporating lexical priors into topic models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 204–213. Association for Computational Linguistics.

- Kuha, J. (2004). Aic and bic: Comparisons of assumptions and performance. *Sociological methods & research*, 33(2):188–229.
- Midi, H., Sarkar, S. K., and Rana, S. (2010). Collinearity diagnostics of binary logistic regression model. *Journal of Interdisciplinary Mathematics*, 13(3):253–267.
- O’callaghan, D., Greene, D., Carthy, J., and Cunningham, P. (2015). An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications*, 42(13):5645–5657.
- Rajadesingan, A., Resnick, P., and Budak, C. (2020). Quick, community-specific learning: How distinctive toxicity norms are maintained in political subreddits. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 557–568.