Proceedings of the Fourteenth International AAAI Conference on Web and Social Media (ICWSM 2020)

Higher Ground? How Groundtruth Labeling Impacts Our Understanding of Fake News about the 2016 U.S. Presidential Nominees

Lia Bozarth, Aparajita Saraf, Ceren Budak
University of Michigan, School of Information
105 S State St
Ann Arbor, MI 48109
lbozarth,apsaraf,cbudak@umich.edu

Abstract

The spread of fake news on social media platforms has garnered much public attention and apprehension. Consequently, both the tech industry and academia alike are investing increased effort to understand, detect, and curb fake news. Yet, researchers differ in what they consider to be fake news sites. In this paper, we first aggregate 5 lists of fake and 3 of mainstream news sites published by experts and reputable organizations. Then, focusing on tweets about the democratic (Hillary Clinton) and republican (Donald Trump) nominees in the 2016 U.S. presidential election, we use each pair of fake and traditional news lists as an independent "groundtruth" to examine i) the prevalence, ii) temporal characteristics and iii) the agenda-setting differences between fake and traditional news sites. We observe that depending on the groundtruth, the prevalence of fake news varies significantly. However, the temporal trends and agenda-setting differences between fake and mainstream news sites remain moderately consistent across different groundtruth lists.

Introduction

Following the 2016 U.S. presidential election, fake news swiftly became a topic of interest and scrutiny for political pundits, media scholars, and the general public (Silverman 2016; Guo and Vargo 2018)—driving increased research efforts on fake news. The research community has been struggling to define fake news. While there is currently no consensus on the topic, leading scholars advocate "... focusing on the original sources—the publishers—rather than individual stories, because we view the defining element of fake news to be the intent and processes of the publisher." (Lazer, Baum, and others 2018). Yet, there is currently no agreement on which news producers are fake news producers either (Tandoc Jr, Lim, and Ling 2018).

Consequently, there are a number of lists with opaque generation processes (Zimdars 2016; Guo and Vargo 2018; Allcott and Gentzkow 2017; Van Zandt, Dave 2018; Politifact staff 2018; Shao et al. 2016) being used by studies with important implications such as examining fake news cascading behavior (Allcott and Gentzkow 2017; Allcott, Gentzkow, and Yu 2018), assessing agenda-setting powers of fake and traditional news sites (Vargo, Guo, and Amazeen 2018; Guo and Vargo 2018; Mukerji 2018) or characterizing changes in fake news trends (Allcott, Gentzkow, and Yu 2018). How robust are these studies, particularly the ones focused on the 2016 presidential elections, with respect to the choice of groundtruth lists that define which publishers are producers of fake or traditional news? We set out to answer this question through meta-analysis—a methodology used to overcome the limitations of any single study by consolidating multiple data sources or studies that aim to address the same research questions, and determining their similarities and differences (Boulianne 2015).

Here, we aggregate 5 lists of fake and 3 of mainstream news sites contributed by both the academia and other reputable sources (Poynter Institute 2019; Zimdars 2016; Wang 2017; White 2018; Leetaru and Schrodt 2013; Van Zandt, Dave 2018). We first review the labeling processes of these lists, assess their similarities and temporal changes. We then determine how selection impacts prevalence, temporal trends, and agenda-setting analysis of fake news about the 2016 presidential nominees.

We first examine prevalence given the divergent findings in recent work (Silverman 2016; Allcott and Gentzkow 2017; Bovet and Makse 2019) [1]. A careful analysis of prevalence can also help lawmakers/platforms in better prioritizing anti-misinformation actions (Lazer, Baum, and others 2018). Next, we investigate the robustness of trend analysis since having an accurate assessment of temporal patterns of fake news can assist lawmakers/platforms in evaluating whether their efforts to curtail fake news is successful (Allcott, Gentzkow, and Yu 2018). Finally, we turn to topic analysis. Agenda-setting theory (McCombs and Shaw 1972) postulates that the most frequently covered topics are what the general public considers the most important. Relatedly, fake news sites could have led voters to re-evaluate issue importance and nominee viability by prioritizing cer-

[1] For instance, Silverman (2016) suggests that fake news articles garnered "...sometimes more than twice as many as legitimate news scoops in major outlets". Whereas, Allcott and Gentzkow (2017) suggests that an average adult only saw and remembered 1.14 fake news articles during the 2016 presidential election.

tain topics over others (Guo and Vargo 2018). Determining the robustness of fake news agenda-setting effects is consequential to media effects research.

Our paper makes the following contributions:

- We demonstrate that existing fake news lists share very few domains in common. Additionally, popular fake news sites are more likely to be included (and included earlier) than unpopular ones. Further, domains in *hate*, *junksci*, *clickbait* subcategories are less likely to be included by lists compared to domains in the *fake* subcategory.
- Based on the groundtruth choice, the prevalence of fake news varies considerably (2%-to-40%). This discrepancy is mostly due to the inclusion or exclusion of domains with mixed factualness.
- We show that the time-series correlation between most lists is high, especially for the general election period where we observe an increase in fake news prevalence regardless of groundtruth choice. Further, we also show that scheduled events contribute to a temporary drop in fake news prevalence. Observations for scandals are not as robust and are dependent on selection.
- Studying the agenda-setting priority difference between fake and traditional news sites, we observe that whether a topic (e.g., immigration) was more central to the coverage from fake news outlets compared to the traditional news sites is robust to the choice of groundtruth.
- Finally, groundtruth selection of mainstream news lists has a very limited impact on all downstream analyses.

To summarize, through meta-analysis, we characterize what makes a domain a fake news producer to some but to not others. We show that the use of different groundtruth sets can account for diverging fake news prevalence findings. Further, despite the varied labeling and validation procedures used and domains listed by fake news annotators, the groundtruth selection has a limited to modest impact on studies reporting on the behaviors of fake news sites (e.g., agenda-setting).

Related Work

Researchers have extensively documented the negative impact fake news has on the quality of civic engagement, healthcare, markets, and disaster management (Rapoza 2017; Marcon, Murdoch, and Caulfield 2017; Palen and Hughes 2018), both within the United States (Silverman 2016; Main 2018; Starbird 2017) and internationally (Kucharski 2016; Alimonti and Veridiana 2018).

Many studies aim to distinguish false content from credible news articles at scale. Prior studies have identified differences in i) linguistic patterns such as punctuation and word choices (Potthast et al. 2017), ii) auxiliary data (Shu et al. 2017; Shu, Wang, and Liu 2018), iii) network cascading attributes such as depth, breadth, and speed (Shao et al. 2017; Allcott, Gentzkow, and Yu 2018; Vosoughi, Roy, and Aral 2018), and iv) agenda-setting priorities (Vargo, Guo, and Amazeen 2018). These differences are then used to build automated fake news detection platforms (Horne and Adali 2017) in an effort to curtail fake news.

However, efforts to study fake news and to diminish its spread are difficult (Budak, Agrawal, and El Abbadi 2011), partly because scholars do not have a consistent definition for fake news (Tandoc Jr, Lim, and Ling 2018; Wardle 2017). For instance, Tandoc et al. identify 2 primary dimensions of fake news: levels of facticity and deception. Wardle, on the other hand, conceptualizes fake news using 3 distinct dimensions: type of content, motivation, and dissemination method. Moreover, existing fake news labelsets (Politifact staff 2018; Zimdars 2016; Van Zandt, Dave 2018; White 2018; Leetaru and Schrodt 2013) have considerably different annotation and categorization procedures.

We first consolidate existing groundtruth labelsets of fake and mainstream news sites that have been generated by various groups. We then assess whether and to what extent differences in groundtruth selection affect downstream studies.

Data

We use 3 types of data: i) lists of fake and traditional news sites, ii) tweets about the two nominees during the 2016 U.S. presidential election, and iii) webpages, or news articles, corresponding to the URLs shared in those tweets.

Fake and Traditional News Site Lists: We collect 5 distinct fake news lists and 3 traditional news lists from both the academia and the press (Zimdars 2016; Guo and Vargo 2018; Allcott and Gentzkow 2017; Van Zandt, Dave 2018; Politifact staff 2018; Shao et al. 2016), resulting in 1884 aggregated fake news sites and 8238 traditional news sites. We describe and evaluate these lists in Section .

Twitter Data: The social media dataset is described in detail in Bode et al. (2020). The data collection was performed using Sysomos MAP - a social media search engine that includes access to all tweets (Twitter firehose) going back one year. For any given day between May 23, 2014, and January 1, 2017, our dataset includes i.) 5,000 tweets randomly sampled from all tweets that included the keyword “Trump”, and ii) 5,000 tweets similarly sampled from all that mentioned “Clinton”. The resulting dataset includes approximately 4.8 million tweets each about Donald Trump and Hillary Clinton respectively.

Webpages (News Articles): The webpages dataset (Budak 2019) includes the content of the webpages shared in the Twitter dataset described above. For each tweet with an external URL, the dataset includes a record with: i) the shortened URL, ii) the original URL, iii) domain name, iv) title of the document, v) body of the document, (vi) the date of the tweet, vii) Twitter account id of the user sharing the URL, and viii) a binary categorization that indicates whether this tweet is about Clinton or Trump. We remove the records with domains not listed in the aforementioned 10K+ news sites and filter out the tweets posted before 12/01/2015 or after 01/01/2017. We derive approximately 244K unique articles shared by 1M Tweets on Twitter.

Meta-review

In this section, we first examine the characteristics and applications of the available lists of fake and traditional news websites. Then, focusing on fake news lists, we assess their

commonalities and differences and explore the characteristics of websites that are correlated with them being included in or excluded from any given list. Finally, we explore fake news domains' likelihood of becoming defunct.

Lists of Fake News Sites: We collect 5 fake news lists.

1. ZDR: We refer to the set of fake news websites annotated by Zimdars et al. (2016) as ZDR. ZDR tags each website with at most 3 of the following 10 subcategories: *fake, satire, bias, conspiracy, rumor, state, junksci, hate, clickbait*, and *unreliable* [2]. Among these subcategories, *unreliable* and *clickbait* are noted to have "mixed" factualness.
2. MBFC: The set of sites labeled by *Media Bias/Fact Check*–an independent online media outlet maintained by a small team of researchers and journalists (Van Zandt, Dave 2018)–will be referred to as MBFC. Similar to ZDR, MBFC assigns domains to subcategories: *fake, conspiracy, satire*. Moreover, it also labels websites with political ideology (*extreme left, left, center, right, extreme right, unlabeled*) and rates websites by their factualness (*low, mixed, high*).
3. POLIT: The staff of PolitiFact, in collaboration with Facebook, identified the list of most-shared fake news sites on Facebook during the 2016 election (Politifact staff 2018). This list–referred to as POLIT–labels sites to *fake, imposter, some fake*, or *parody*.
4. DDOT: This list is shared by *the Daily Dot*, a mainstream online news site (White 2018). This list is largely created by referencing other pre-existing fake news lists and does not contain subcategories.
5. AGZ**:** Allcott et al. (2018) aggregated the following five lists: POLIT, Grinberg et al. (2018), Silverman (2016), Schaedel (2018), and Guess et al.(2018). This list is referred to as AGZ. The subcategorization process in AGZ is somewhat complex. For instance, POLIT subcategories were ignored and all the domains were relabeled as *fake*. However, the subcategories *black, red, orange* (black: completely false, red/orange: has unreliable claims) of Grinberg et al. (2018) were maintained. Finally, all domains from other referenced lists were labeled as *fake*.

A synthesis of these lists reveals that 4 out of the 5 lists share 2 common subcategories: i) a subcategory containing domains with *mixed* factualness, and ii) a *fake* subcategory (entirely fabricated). This consistency suggests that *mixed* or *fake* domains are conceptually distinct from others. Thus, studies should take this distinction into consideration.

Lists of Traditional News Sites: We consider the following 3 traditional news lists.

1. ALEXA: Alexa is an online domain directory owned by Amazon (Wikipedia contributors 2019). We crawl for all the websites listed under Alexa's *News* category.
2. MBFC(T): *Media Bias/Fact Check* also lists a large set of traditional news sites. We refer to this list as MBFC(T).
3. VARGO: This list contains fact-based news websites compiled through manual content analysis of the top news media websites found in GDELT's global knowledge graph (Vargo, Guo, and Amazeen 2018).

Considering fake news domain list quantities, DDOT has the fewest with 175 domains, followed by POLIT (327) and AGZ (673). ZDR (786) and MBFC (1183) are the largest lists. Traditional news site list quantities are MBFC(T) (1685), VARGO (2649) and ALEXA (5497). Table 1 provides a summary of the annotation processes and the uses of these lists. As is evident from the second column (*Annotation and Quality*), most lists do not have a transparent annotation and quality evaluation procedure. Perhaps due to the absence of such robust procedures, there is no consensus on which of these lists should be treated as the ultimate groundtruth. This is clear from the third column (*Applications*). More than 20 studies have used these lists of fake and traditional news sites. The lists are used for various important purposes such as building automated fake news classifiers or assessing the impact of fake news on the 2016 election. This highlights the importance of identifying similarities and differences between the lists.

Thus, we conduct downstream analysis using different groundtruth pairs (f,t) where $f \in \{\texttt{ZDR}, \texttt{MBFC}, \texttt{POLIT}, \texttt{DDOT}, \texttt{AGZ}\}$, and $t \in \{\texttt{ALEXA}, \texttt{MBFC(T)}, \texttt{VARGO}\}$.

List Overlap: Here, we identify the overlap among the 5 fake news lists using 2 metrics. We first calculate the fraction of websites being present in at least 2 of the 5 lists, then 3, then 4. We observe that close to 50% of all domains are only included in a single list. In fact, only 5.7% of the domains are included by all fake lists. Second, we also calculate the Jaccard similarity score (Goodall 1966) of each pair of lists. We observe that more than half of the 15 pairs of fake news lists have a similarity of $<= 0.1$. We note that MBFC and DDOT have the lowest Jaccard similarity score of 0.08, and AGZ and POLIT have the highest score of 0.48.

The extent of dissimilarity between the lists is surprising, and we identify four potential measures: i) *popularity*, defined as the number of times a URL from a given domain is shared in the Twitter dataset, ii) *age* (we collect data using whois.com, an online domain registration service), iii) *subcategory*, as defined by Zimdars et al. (2016)[3], and finally vi) *ideology*, as defined by *Media Bias/Fact Check* (Van Zandt, Dave 2018)[4]. The details of the regression model and analysis are provided in the Appendix. We observe that the popularity of a website is positively correlated with being included in lists (though the variable is not significant for

[2]Zimdars et al. also list a small subset of domains as *political, reliable* and *unidentified* which are not fake news sites and therefore removed from subsequent analyses.

[3]Zimdars et al. (2016) have the most comprehensive subcategories and a coherent labeling guideline. Subcategory is *unknown* if a domain is not listed by Zimdars et al. (2016).

[4]Ideology is *unknown* if the domain is not listed by *Media Bias/Fact Check* (Van Zandt, Dave 2018) or if *Media Bias/Fact Check* didn't mark it with an ideological label (approximately 18.6% domains). Here we collapse MBFC's *extreme left* and *left* categories into single *liberal* class. Same for *conservative*.

List	Annotation and Quality	Applications
DDOT	no information	build automated fake news trackers (Shao et al. 2016; Helmstetter and Paulheim 2018), assess agenda-setting powers of fake and traditional news sites (Vargo, Guo, and Amazeen 2018; Guo and Vargo 2018; Mukerji 2018)
AGZ	authors aggregate lists generated by others, and then use various combinations of these list for result robustness check	assess impact on election, examine fake news cascading behavior (Allcott and Gentzkow 2017); examining fake news trend (Allcott, Gentzkow, and Yu 2018)
MBFC	annotated by staff; authors examine wording, source, story selection, and political affiliation	studies of the Alt-right (Main 2018), globalism (Starbird 2017), the virality of fake news (Darwish, Magdy, and Zanouda 2017), information literacy (Farmer 2017), polarization (Croft and Moore 2017), and information quality (Nelimarkka, Laaksonen, and Semaan 2018)
POLIT	no information	study the diffusion of fake news on social media (Allcott, Gentzkow, and Yu 2018), information literacy (Mukerji 2018), automate fake news detection (Granskogen 2018)
ZDR	annotated by scholars and librarians; domain name, about us page, writing style, aesthetics, and social media accounts are among the examined characteristics	examine network cascading behavior difference between fake and real news articles during the 2016 Election (Allcott and Gentzkow 2017; Allcott, Gentzkow, and Yu 2018), build fake news classifiers (Shao et al. 2016; Horne and Adali 2017; Horne et al. 2018), assess agenda-setting powers of fake and real news sites (Vargo, Guo, and Amazeen 2018; Guo and Vargo 2018), impact assessment (Rini 2017; Figueira and Oliveira 2017; Doshi et al. 2018), ethics and policy (Farte and Obada 2018; Koulolias et al. 2018)
MBFC-T	see MBFC	see MBFC
ALEXA	no information	examine cascading behavior differences between fake and traditional news articles (Allcott and Gentzkow 2017; Allcott, Gentzkow, and Yu 2018), news sharing behavior in right-leaning echo chambers (Lima et al. 2018)
VARGO	annotated by authors; intercoder reliability of 0.988 Krippendorff's alpha.	assess agenda-setting power of fake and real news sites (Guo and Vargo 2018; Vargo, Guo, and Amazeen 2018)

Table 1: Traditional and Fake News Lists and Their Applications. Some studies below use multiple sources.

DDOT and POLIT). Further, ideology is not predictive of whether a domain will be included by lists except for AGZ (conservative-leaning domains are more likely to be listed). Finally, we observe that compared to domains subcategorized as *fake* by ZDR, domains that belong to other subcategories are uniformly less likely to be present in other lists.

Domain Addition and Removal through Time: We further examined how the lists changed over time and found the types of changes to be largely consistent. For the lists we have temporal information for (MBFC, ZDR, and DDOT), we observe the following: i) they include more popular domains earlier on–adding the less popular ones later, ii.) they include the sites that publish fake news earlier compared to sites that publish less problematic categories such as *clickbait* and *bias*, and iii.) interestingly, sites labeled as *satirical* are added early on to the lists, perhaps due to the ease of identification. For the regression model for temporal analysis, we refer the reader to the Appendix.

Besides the addition of domains through time, we also looked into i) domain removals and ii) domains with changed subcategories. We observe very few to no removals [5]; same for changes of subcategories.

Active and Defunct Domains: Once flagged as fake news websites, these publishers may aim to bypass fact-checking systems by using simple tricks such as abandoning their domains and migrating to new ones (Funke 2019). We observe that 68.9% of all websites listed under POLIT are no longer active—the highest defunct rate among all lists [6]. Further, AGZ and DDOT have comparable defunct rates of 64% and 62% respectively. In comparison, ZDR and MBFC have considerably lower rates of 40.6% and 30.9%. Similar to the previous section, we assess a domain's likelihood of being defunct as a function of its *popularity, age, subcategory*, and *ideology* (see Appendix).

We show that older, more popular, and ideologically conservative or ambiguous domains are less likely to be defunct. Further, compared to domains subcategorized by ZDR as *fake*, domains with other subcategories (e.g., *junksci*, *satire*) are less likely to be defunct. Thus, one possible explanation that ZDR and MBFC have lower defunct rates is that both sources include more domains that do not belong to the subcategory *fake* (e.g., *unreliable* and *conspiracy* websites), and these types of domains are targeted less frequently by fact-checking platforms and thus have less incentive to migrate.

GroundTruth Selection and Downstream Consequences

A meta-review of the fake news lists in the previous section demonstrates marked differences between these lists. How do these differences affect the downstream analysis? We aim to answer this question in this section. To that end, we first assess how groundtruth selection impacts the perceived prevalence of fake news during the 2016 election. Next, we measure the similarities or dissimilarities of fake news time-series generated using different groundtruth pairs (f,t). Finally, we determine whether there are any marked differences in agenda-setting priorities of fake and real news sites due to choice in groundtruth.

Prevalence

Here, we define *prevalence* as the fraction of tweets containing URLs that are from fake news sites. We examine to what extent groundtruth difference impacts perceived pervasiveness of fake news using 3 distinct boundary conditions (strictness in definition) for each fake news list:

[5] The exception being DDOT: in late 2016, DDOT contained 98 websites; it then removed a substantial number of sites and reduced its size to 25 in mid-2017; its latest version has a size of 175. No explanation was given for each change.

[6] We use *scrapy* (Mitchell 2018), a Python crawler library, to scrape website homepages. Domains timed-out during scraping, or returned 404 errors (Not Found), 502 (Bad Gateway), 503 (Service Unavailable), et cetera are labeled as defunct.

all, *all-except-mixed*, and *fake*. More specifically, given a groundtruth pair (f,t), we write f_{all} as the entire set of domains in f, f_{mixed} and f_{fake} as the set of domains in f_{all} that belong to subcategories with mixed factualness and the subcategory *fake* respectively. We then calculate *prevalence* as $\frac{|f_{all}|^s}{|f_{all}|^s+|t_{all}|^s}$, $\frac{|f_{all}|^s-|f_{mixed}|^s}{|f_{all}|^s+|t_{all}|^s}$, and $\frac{|f_{fake}|1|^s}{|f_{all}|^s+|t_{all}|^s}$ where $|f_{all}|^s$ is the number of tweets, or shares, contributed by f_{all}.

Results are shown in Figure 1. For the *all* condition, based on (f,t), fake news could amount to be more than 40% of total news shares or as low as less than 3%. Further, for robustness check (details in Section), we also redefine prevalence as the fraction of unique accounts that posted at least 1 fake news tweet and observe comparable results.

Additionally, if we discard all domains with mixed factualness, prevalence drops substantially to between 1.3% and 20.1%. Further, the fraction of fake news are comparable for the conditions *all-except-mixed* and *fake* except for ZDR. In other words, domains that are low in quality but not necessarily fake, (i.e. *mixed*), contribute to a large fraction of total articles shared, and domains that are neither *fake* nor *mixed* are not as popular on Twitter. To further illustrate this point, we calculate the average number of tweet shares per domain for each type of subcategories. We observe that *mixed* domains have an average of 0.7K to 2.4K tweet shares, 4 to 5 times that of the average of all subcategories for each list; in fact, *mixed* domains, on average, are considerably more popular than traditional news outlets which had an average tweet share of 0.15K to 0.66K.

Our analysis helps explain the divergent findings in the literature. While some studies raise significant concerns about the prevalence of fake news (Silverman 2016), others claimed limited prevalence (Allcott, Gentzkow, and Yu 2018) [7]. Here, similar to work by Grinberg et al. (2018) which showed that the analysis on fake news exposure is significantly dependent on whether domains of mixed factualness were included, we see that drastically divergent conclusions can be reached even with the same Twitter data as a function of the fake and traditional news lists and fake-ness definitions (e.g., fake, mixed) one chooses to use. In sum, the more comprehensive a fake news list is, the higher the fake news prevalence.

Time-Series Analysis

In this section, we first construct a time-series representing the fraction of fake news over all available news per day for each (f,t) from 3 different time periods (primary, general election, and after election) accounting for only Clinton tweets, only Trump tweets, and all tweets (for both nominee). Specifically, for each election phase i where $i \in \{primary, general\ election, after\ election\}$,

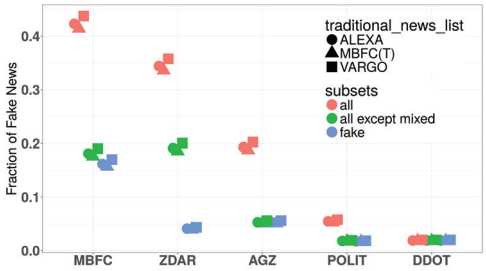


Figure 1: Fraction of Fake News. The x-axis indicates fake news lists. Each list is divided into subsets (marked by color) of *all*, *all-except-mixed* (not including domains in *mixed* subcategory), and *fake* (only domains in *fake* subcategory). The shape of each point denotes mainstream news lists, and y-axis is the fraction of tweets contain fake news.

given a groundtruth pair (f,t) and nominee n (where $n \in \{clinton, trump, both\}$), we write $|f|^s_{0,n}$, and $|t|^s_{0,n}$ as the total number of tweets, or shares, that mention n and contain URLs from f or t at day 0 [8]. We then derive the time-series $P^i(f,t,n) = \{\frac{|f|^s_{0,n}}{|f|^s_{0,n}+|t|^s_{0,n}}, \frac{|f|^s_{1,n}}{|f|^s_{1,n}+|t|^s_{1,n}} ...\}$.

Next, we then compare these time-series from 3 distinct dimensions: i) correlation, ii) trend, and iii) effects of external events. For example, one might be interested to know how consistent the fake news trend is over time for discussions about Clinton (n) during the primary (i) when using (MBFC, ALEXA) or (AGZ, ALEXA) as the groundtruth pair. For that, we can use $P^{primary}$(MBFC, ALEXA, *Clinton*) and $P^{primary}$(AGZ, ALEXA, *Clinton*). Furthermore, instead of comparing only 2 pairs, we can compute and contrast the findings for all 15 pairs to examine overall consistency of Clinton conversations during the primary season.

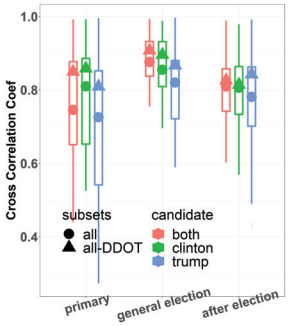
Time-series Correlation: We calculate correlation separately for each time period and nominee. For each n and i, given 2 groundtruth pairs $(f1,t1)$ and $(f2,t2)$ where $f_1 \neq f_2$ or $t_1 \neq t_2$, we compute the maximum normalized cross correlation coefficient and the corresponding time lag (Haugh 1976) of $P^i(f_1,t_1,n)$ and $P^i(f_2,t_2,n)$.

We observe that the highest correlation scores of all pairwise comparisons occur at 0 lag, indicating that no single time-series is "ahead" or "behind" others. Correlation scores are plotted in Figure 2a. Normalized coefficients have a range between $\{-1,1\}$. As shown, correlation for $P(f_1,t_1,n)$ and $P(f_2,t_2,n)$ is the highest when $f_1 \equiv f_2$ but $t_1 \neq t_2$, indicating traditional news list selection (choosing ALEXA, MBFC(T), or VARGO) has little impact here. Further, we also note that certain fake news lists have considerably high correlation (e.g., ZDR and MBFC have correlation consistently higher than 0.9). Yet, DDOT diverges significantly from others.

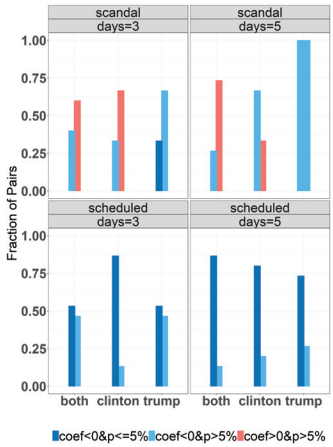
We further observe that the correlation is highest for the *general election* season (median correlation between the

[7] More specifically, Silverman (2016) selected the top 20 highest performing fake news stories from hundreds of known fake news sites and demonstrated, on aggregate, they had a larger number of tweet shares compared to the top 20 news stories selected from the top 13 traditional news sites. In comparison, Allcott et al. (Allcott, Gentzkow, and Yu 2018) aggregated 673 fake news sites and showed that an average adult saw and remembered a single fake news story.

[8] Here, we pick 2015-12-01, 2016-06-15, and 2016-11-09 as day 0 for primary, general election, and after election; and 2016-06-21, 2016-11-15, and 2017-01-01 as the last day.



(a) Cross correlation coefficient between pairs of lists for each time period. Circle and triangle points are the median coef scores when including or excluding DDOT.



(b) Level effect of scandals and scheduled events.

Figure 2: Time-series analysis results for *correlation* and *effects of external events*.

pairs for each nominee n are all above 0.8). Most efforts in fake news detection were motivated by the spread of fake news during the 2016 presidential election. This provides one potential explanation—fact-checkers and scholars could have had a stronger emphasis on the publishers that were active in this time frame, resulting in higher agreement.

period	nominee	majority trend	majority frac	median β_1	least.congruent
primary	trump	positive	0.67	0.04%	DDOT, POLTI
	clinton	stationary	0.60	NA	
	both	stationary, positive	0.47	NA, 0.02%	
general election	trump	positive	**1.00**	0.03%	NA
	clinton	positive	**1.00**	0.09%	
	both	positive	**1.00**	0.07%	
after election	trump	negative	0.67	-0.01%	AGZ, POLTI
	clinton	positive	0.80	0.07%	
	both	positive	0.67	0.05%	

Table 2: Fraction of Fake News Per Day Time-series Trend Using Different GroundTruth. Column *majority trend* denotes the trend observed by a majority of pairs, column *majority frac* is the fraction of pairs in majority.

Trend: Similar to prior work (Allcott, Gentzkow, and Yu 2018; Lazer, Baum, and others 2018), we are also interested in assessing whether there was an increase in fake news prevalence and to what degree findings would depend on the choice of groundtruth pairs. Here, we first employ seasonal decomposition using moving averages (Beveridge and Nelson 1981) to deconstruct each time-series $P^i(f,t,n)$ into its components *trend*, *seasonal*, and *residual*. This is to remove the seasonality and residuals from the original time-series. Next, we apply both Augmented Dickey-Fuller (ADF) and Kwiatkowski Phillips Schmidt Shin (KPSS), 2 commonly used methods to test for stationarity (Charemza and Syczewska 1998) on the *trend* component of $P^i(f,t,n)$. If any one of the tests show that unit root is non-stationary, we run the linear regression model $y^i(f,t,n) = \beta_0 + \beta_1 * T + \varepsilon$ (where $y^i(f,t,n)$ contains the values from *trend*, and T is the time elapsed since the start of the time-series). Here, a positive β_1 suggests a rise of fake news. Finally, we assess whether trend analysis results for each nominee n and time period i using all pairs (f,t) are consistent.

Results are on Table 2. Column "majority trend" shows the trend result shared by the largest fraction of groundtruth pairs and column "majority frac" is the size of that fraction. Additional, median β_1 scores indicate the median estimated percentage increase, over all congruent pairs, of fake news per day. As shown, conclusions for the *general election* are remarkably consistent: all lists pairs indicate an increase in fake news. In other words, regardless of whether groundtruth choice is (MBFC, ALEXA), (AGZ, VARGO), or any of the other combinations, we repeatedly see a positive trend for fake news in the general elections. Results for other election phases are less congruent (e.g., 80% pairs show a positive trend for Clinton-related fake news after the election, but the other 20% show stationarity or a negative trend). We observe that DDOT and POLIT disagree the most with other fake news lists (measured by the number of times a list diverges from "majority vote") in *primary*, whereas it's AGZ and POLIT in *after election*.

In sum, similar to time-series correlation analysis, we see a higher consistency for the *general election* period compared to *primary* and *after election*. Accurate trend analysis is vital given that it impacts platform owners and policymakers' decision-making. Facebook's fact-checking system targets domains listed in POLIT and AGZ, and consequently (Allcott, Gentzkow, and Yu 2018) shows significantly reduced content from these domains on Facebook over time. We do not have access to Facebook data and therefore cannot check the robustness of their curtailing efforts. Yet, we do demonstrate that caution must be taken when examining fake news spread outside of the general election period.

Effects of External Events: Many prior studies examined media coverage of i) unexpected political events such as scandals (Puglisi and Snyder Jr 2011) as well as ii) scheduled high-profile events such as the presidential debates (Scheufele, Kim, and Brossard 2007). Such events are shown to have important effects on campaign news coverage. Here, we examine whether these 2 distinct categories of events have a temporary effect on the prevalence of fake news, specifically in the *general election* period.

We first obtain a list of scandals and planned key events of Trump, Clinton, or both that occurred in the general election from *ABC News* and *The Guardian*. The list, ordered chronically, includes: Republican nomination (07/18), Democrat nomination (07/28), Clinton "deplorable" and "pneumonia" scandals (09/09), first debate (09/26), Clinton email involving Wikileaks and Trump Hollywood tape scandals (10/07), second debate (10/09), Clinton email scandals involving the FBI (10/28, 11/06), and finally election day (11/08). Here, nominations, debates, and election day are assigned to *scheduled* and others to *scandal*.

Next, we use the autoregressive integrated moving average (ARIMA) time-series model (Stock, Watson, and oth-

ers 2003) to run interrupted time-series analysis and identify whether *scandals* and *scheduled* events are associated with level changes in the fraction of fake news per day for x days where $x \in \{3,5,7\}$. In our paper, we use *auto.arima*, a common ARIMA model selection function (Makridakis, Wheelwright, and Hyndman 2008) from R's forecast library. Given a time-series, $P^i(f,t,n)$, and a set of external regressors (i.e., events), *auto.arima* selects the best ARIMA model based on the corrected Akaike information criterion (AIC). Here, we have 2 external regressors for each n. We denote $xreg^1_{n,T} = \{0,0,...1,1,1...\}$ where $xreg^1_{n,t} = 1$ if day t is within x days of the nearest *scandal* (after it has occurred) involving n. Similarly, we write $xreg^2_{n,T}$ for *scheduled* [9].

A positive coefficient returned by *auto.arima* for $xreg^1_{n,T}$ would mean that *scandals* temporarily increase the fraction of fake news per day. As shown in Figure 2b [10], regardless of the groundtruth selection, *scheduled* events generally contribute to a reduction of fake news. This does not mean planned events reduced the absolute volume of fake news. One possible explanation is mainstream media simply covered scheduled events much more, thus $\frac{|f|^s}{|f|^s+|t|^s}$ is smaller. Results for *scandal* are, however, more varied, suggesting that groundtruth pair selection has an impact on perceived effects of scandals. For instance, we see that *scandals* contributed to a short-term *increase* in the fraction of fake news shared per day when given groundtruth pair $(\texttt{ZDR}, \texttt{ALEXA})$, but a *decrease* if pair is $(\texttt{POLIT}, \texttt{ALEXA})$. This discrepancy is particularly important to studies that examine how scandals and negative media coverage diminish voter turnout in the 2016 election, particularly for Clinton (Faris et al. 2017).

Agenda-setting Priorities

In this section, we first use an iterative topic modeling process to extract issues, or topics, being covered by both fake and traditional news sites and assign each news article to its corresponding topic. Next, we examine whether the choice of groundtruth pairs impacts agenda-setting conclusions.

Topic Modeling of News Articles Using Guided LDA: We use Guided LDA for topic modeling. It is an extension of the base LDA that allows sets of keywords to guide document topic assignment by increasing their "confidence" or weights (Jagarlamudi, Daumé III, and Udupa 2012).

First, we use base LDA and manual labeling to extract seed words from news articles[11]. More specifically, we use *gensim* (Rehurek and Sojka 2011) to generate several base LDA models[12]. We then select the model which has the optimal coherence score[13]. From it, we obtain the top 30 most representative words for each topic. Next, we manually inspect words and categorize them into coherent sets (i.e., topics). Using this approach, we obtain 409 unique seed words divided into 33 different sets. Next, we run the guided using the derived seed word sets[14]. We filter out the subset of topics that lacked coherent themes and collapse topics that share the same human-interpretable theme into a single topic. This process results in 19 distinct topics. Finally, we assign each document into a single topic according to the maximum probability of its topic distribution. This topic is later referred to as the document's *predicted* topic label.

topic	doc frac	most weighted tokens	f1
abortion	0.96%	woman abort life plan_parenthood issu punish femal	0.87
benghazi	0.60%	attack benghazi libya committe report secretari secur	0.75
c-health	0.86%	medic doctor releas report mental suffer pneumonia	0.75
climate	1.40%	climat coal environment industri land administr regul	0.89
wst	0.30%	speech wall_street talk ask issu transcript releas	0.82
d&i	0.75%	commun lgbt issu equal woman discrimin anti marriag	0.78
economy	4.4%	trade job china deal compani manufactur econom	0.79
election	20.3%	sander berni primari voter percent poll voter cruz	0.77
email	5.76%	email depart investig server classifi comey secretari	0.84
border	2.28%	immigr border mexico wall illeg deport mexican build	0.85
mid-east	3.86%	muslim islam israel isi terror terrorist attack unit syria	0.76
religion	1.14%	christian evangel church faith religi leader pastor pope	0.78
russia	1.81%	russia russian putin intellig hack offici govern	0.76
security	1.70%	iran china nuclear polici foreign deal nato secur	0.78
sexual	1.93%	woman accus alleg rape husband sexual claim sexual_assault	0.82

Table 3: List of Topics, Fraction of Total Documents Accounted for, Most Weighted Keywords, and F1

Topic Modeling Quality Assessment and Selection: For each topic, we randomly sample 0.2% of its documents (or 10 if the size of a topic is small). This gives us 434 unique documents. We also sample 0.2% documents from the articles not included in the 19 topics. This results in 525 documents. Finally, we shuffle and publish the 1K $(434+525)$ documents on MTurk for crowdsourced labeling[15].

We assign 3 independent workers to categorize each document[16] and mark the *manual* topic of each article according

[9] If n is *both*, we only use events that involve both nominees.

[10] Trend results for when $x = 7$ is omitted due to space.

[11] We remove stop words, lemmatize and perform stemming. Finally, we remove all articles that have <100 or >800 word tokens.

[12] The number of topics are $\{50,75,100,125,150\}$ respectively for the models. In addition, we set all models to ignore words and bigrams that have a frequency of less than 100 or occur in more than 50% of total documents.

[13] Coherence score for a topic is the average of the pairwise word-similarity scores of its words (Newman et al. 2010). A model's coherence score is the sum over its topic coherence scores.

[14] We adjust model's seed confidence to 0.25 and set the number of total topics to 125. We use perplexity score (Misra, Cappé, and Yvon 2008) to determine the optimal number of topics given that *gensim* does not support coherence calculation for guided LDA.

[15] The success of a crowdsourcing task relies heavily on the right mechanisms to ensure worker qualifications. We require that workers: 1) reside in the U.S. 2) have successfully completed at least 1,000 HITs; and 3) have an approval rate of at least 98%.

[16] Workers are given a list of categories (19 topics listed in Table 3 + 1 *none of the above* option) to choose from and are instructed to select a single *primary* category of a given article. We use Krippendorff's alpha (Hayes and Krippendorff 2007) to measure interrater reliability. It is 0.62, which means a moderate agreement.

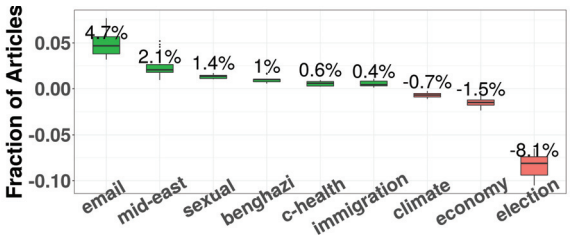


Figure 3: Relative agenda priority difference between fake and traditional news. Y-axis is fraction of fake news articles on topic i subtracted by the fraction of traditional news articles on i. Topics colored in green indicate a higher priority by fake news.

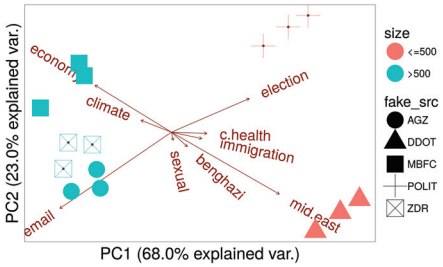


Figure 4: PCA plot for topic fractional difference distribution between fake and traditional news described in Section. Fake news lists are marked by shape.

to the majority vote [17]. Next, for each topic, we calculate its precision, recall, and f1 scores using the *manual* and *predicted* topic labels. We filter out the topics that have an f1 score of < 0.75. This process produces 15 distinct topics accounting for 49% of total articles. Table 3 provides this list of topics, their names, prevalence across domains that are listed by at least one fake or traditional news list, most weighted keywords, and f1 score. As shown, *election* is the most prevalent topic accounting for 20.3% of total news articles, followed by Clinton's email scandal, and the economy.

Agenda-setting Priorities: Next, we assess whether groundtruth choice affects the perceived agenda-setting difference between fake and mainstream news.

For a groundtruth pair (f,t), we derive the following topic distributions $K_{(f,t)} = \{k^1_{(f,t)}, k^2_{(f,t)} ... k^{16}_{(f,t)}\}$ and $L_{(f,t)} = \{l^1_{(f,t)}, l^2_{(f,t)} ... l^{16}_{(f,t)}\}$ where $k^i_{(f,t)}$ and $l^i_{(f,t)}$ are the fractions of fake and traditional news articles on topic i respectively, and $\sum_{i=1}^{16} k^i_{(f,t)} = 1$, $\sum_{i=1}^{16} l^i_{(f,t)} = 1$.

Then, for each topic i in $\mathscr{I}$ (where $\mathscr{I}$ is the entire set of topics), and all groundtruth pairs (F,T), we apply Student's T-test on $K^i(F,T)$ and $L^i(F,T)$ to determine whether the difference in mean is statistically significant between these 2 distributions (here, $K^i(F,T) = \{K^i(f1,t1), K^i(f2,t1) ... K^i(f5,t3)\}$). In other words, we assess whether fake news sites have published significantly more or fewer articles (measured using normalized fractions) on certain topics than traditional news sites and vice versa. We observe a significant difference in 9 topics. For instance, the average fraction of traditional news articles focusing on *election* is 22.5%, while the average is less than 15% for fake news articles. Traditional news sites are also more concentrated on topics including *economy* and *climate*. Fake news sites, on the other hand, spend a considerable fraction, approximately 10%, of all articles on Clinton's *email* scandal alone, twice that of traditional news sites. Fake news sites also place a stronger emphasis on topics such as *sexual* scandals (mostly related to Bill Clinton), and Hillary's pneumonia and claims of early onset dementia.

[17] Articles that do not have a majority is labeled as *unknown*. We observe 46, or 8.6% *unknown* documents. Note, *unknown* documents differ from *none of the above*.

For each pair (f,t), we calculate the difference distribution $D_{(f,t)} = \{d^1_{(f,t)}, d^2_{(f,t)} ... d^9_{(f,t)}\}$ where $d^1_{(f,t)} = k^1_{(f,t)} - l^1_{(f,t)}$. We plot $D^{\mathscr{I}}_{(F,T)}$ in Figure 3. Notably, the data points of $D^{\mathscr{I}}_{(f,t)}$ consistently stay above or below the horizontal $y = 0$ line. For instance, given groundtruth (AGZ, VARGO), 13.1% and 23.6% of fake and traditional news articles covered the election. The negative difference is statistically significant, suggesting that fake news places less priority on the horse-race coverage compared to traditional news. Further, the negative difference persists for all pairs (f,t). Similarly, for all pairs (f,t), we consistently see a higher fraction coverage of Clinton's email scandal by fake news outlets. In other words, the assessment as to whether a topic was more central to the coverage of fake vs. traditional news outlets is robust to the choice of groundtruth pairs. This is good news for studies that are focused on misinformation publishers' agenda-setting functions (Vargo, Guo, and Amazeen 2018): fake news domains commonly prioritize hyperpolarizing and hyperpartisan issues, and including more or fewer domains in a study is unlikely to change the overall results.

Groundtruth Difference Using Factor Analysis: Here, we provide an analysis of how topics contribute to the variance in agenda-setting across groundtruth pairs. We apply PCA (Wold, Esbensen, and Geladi 1987) on $D^{\mathscr{I}}_{F,T}$ and extract the first 2 principal components (the first and second component explains 68% and 23% of the total variance). The resulting biplot is shown in Figure 4. We see that MBFC, ZDR, and AGZ are more similar in their topic distributions. In comparison, fake sites in POLIT have a higher fraction of articles on *election*. One possible explanation is that this list is specifically created to reduce election-related fake news (Politifact staff 2018). Additionally, we also see that fake news sites in DDOT have a higher priority for scandals and controversial issues including *benghazi* and *sexual*, perhaps due to *Daily Dot* being a social news site focused on fake news sites that wrote entertaining false content.

Robustness Checks

In this section, we conduct additional analysis to ensure that our results on the prevalence, temporal attributes, and agenda-setting priorities of fake news with respect to groundtruth choice are robust.

Fake News Characteristics Measured Using User Participation: Thus far, we have approximated prevalence using the number of tweets. Yet, it's possible to have a few exceedingly active and concentrated accounts post a large amount of tweets containing fake news without gaining traction in the general population (Grinberg et al. 2018). Here, we re-examine fake news characteristics using the number of users. We observe comparable results.

First, we redefine prevalence as the fraction of accounts that posted at least 1 tweet containing fake news and observe that, depending on the groundtruth choice, the prevalence of fake news ranges from 3.9% to 55.7% (compared to from 1.3% to 43.7% when measured using tweets).

Focusing on temporal patterns, we again see a consistent positive trend on the fraction of users who shared fake news during the general election period. That is, regardless of groundtruth choice, we observe that the closer the time was to the general election date, the higher the fraction of users who shared fake news. Further, *scheduled* events are consistently associated with a short-term decrease in the fraction of users who shared fake news, whereas results for *scandals* are dependent on groundtruth choice (e.g., scandals are correlated with a short-term *decrease* in fake news when the groundtruth pair is $(\texttt{AGZ}, \texttt{ALEXA})$, but a short-term *increase* if pair is $(\texttt{MBFC}, \texttt{ALEXA})$). These observations are comparable to prior results obtained using tweets.

Finally, agenda-setting priority differences between fake and traditional news media measured by user participation (i.e., defining priority as how many unique accounts posted about a given topic versus how many tweets were posted about that topic) result in comparable conclusions. We observe that, for all combinations of f and t, topics including *email*, *mid-east*, and *sexual* have the highest priority in fake news, whereas *climate*, *economy*, and *election* have the highest priority in traditional news. In sum, we arrive at similar results when conducting analysis using user participation compared to when using tweets.

Addressing Potential Biases in Keywords-based Data Collection Another concern lies with data incompleteness leading to biased observations. Thus far, we only use keywords "trump" and "clinton" to collect tweets concerning each of the two presidential nominees respectively. Therefore, a tweet about Hillary Clinton that only includes the first name "hillary" is absent from our original data. Here, we expand our dataset to include the 2 additional random sample of tweets that contain the keywords "hillary" and "hillary clinton" respectively–collected using the Sysomos MAP pipeline (see "Data" section). We then repeat our prior analysis. While the additional data increases the total number of tweets for Clinton to 13.3M, 2.8 times the size of the original dataset, downstream results generally remain the same. For instance, fake news prevalence ranges from 2.2% to 47.7% when using the expanded dataset—similar to the range of 1.3% to 43.7% when using the original dataset. Further, time-series generated using the 2 datasets are also highly correlated (e.g., the median normalized cross-correlation for the time-series on Clinton is 0.94). In fact, the expanded dataset only resulted in 306 additional number of unique articles (a mere 0.13% increase from the total 244K).

Overall, the results suggest that our analysis are robust. However, we note that our dataset and assessments remain only focused on the two 2016 presidential nominees. Our data do not include other related subjects, or personalities, such as political parties and congressional candidates, and the study of these subjects is outside the scope of this paper.

Conclusion and Discussion

In this paper, we first provided a comprehensive overview of the publicly available lists of fake and traditional news sites. We showed that these lists have divergent labeling processes and very few domains in common. In addition, we illustrated that the perceived prevalence of fake news varies substantially based on groundtruth choice. Despite these initially discouraging results, we were able to reach several important robust conclusions. We noted an increase in fake news during the general election season regardless of the groundtruth selection and a temporary reduction of fake news due to scheduled events (conclusions for scandals were more mixed). Finally, after an iterative topic modeling process, we showed that agenda-setting priority differences between fake and mainstream news sites are relatively robust to the groundtruth pair choice. Overall, our results suggest groundtruth selection has a sizable impact on prevalence analysis and limited impact on downstream analysis in i) temporal characteristics, and ii) agenda-setting priorities.

There are several caveats to our study. First, our analysis of groundtruth difference and its impact is limited to domain-level labels. There are more granular datasets that annotate content at article–or even sentence–level. Second, while the focus of our meta-analysis—prevalence, temporal characteristics, and agenda-setting priorities—asks important research questions, future work should also review existing literature on similarly significant issues, such as fake news exposure in different demographics (Grinberg et al. 2018) or supervised fake news detection (Shao et al. 2016), identify similarities and potentially contradictory results, and determine whether groundtruth choice contributes to the observed differences (e.g., how groundtruth affect the performance of automated fake news classifiers).

Third, our dataset and analysis are only focused on the subset of fake news surroundings the two presidential nominees in the 2016 presidential election. Future work should address how the study of fake news in other fields (e.g., misinformation concerning vaccination) could also be potentially impacted by groundtruth choice.

Where do we go from here? How can we make progress as a research community despite the lack of agreement between fake news lists and domains with potential to be considered fake? Our findings can be leveraged to provide guidance.

Guidance on List Expansion and Maintenance for List Creators: Both fake news websites and groundtruth labels are indeed changing through time. List creators should include methods that track and evaluate these changes.

For efficient and timely list expansion, one key road-

	Model 1					Model 2		Model 3
	(DDOT)	(POLIT)	(AGZ)	(MBFC)	(ZDR)	(ZDR)	(MBFC)	(defunct)
independent variables								
ideology_conservative	−0.007	0.049	0.125*		0.006	0.124**	0.033	−0.130*
ideology_unknown	−0.006	0.027	0.170**		0.126	0.064	−0.066	−0.142**
subtype_bias	−0.664***	−0.480***	−0.375***	0.008		0.094**		−0.229***
subtype_clickbait	−0.677***	−0.462***	−0.373***	−0.128*		0.124*		−0.150*
subtype_conspiracy	−0.686***	−0.442***	−0.370***	0.021		−0.047	−0.001	−0.219***
subtype_hate	−0.703***	−0.452***	−0.348***	−0.251***		0.172**		−0.034
subtype_junksci	−0.705***	−0.460***	−0.388***	0.055		−0.032		−0.297***
subtype_rumor	−0.704***	−0.420***	−0.408***	−0.148		0.012		−0.058
subtype_satire	−0.704***	−0.345***	−0.284***	0.047		−0.136***	−0.115*	−0.271***
subtype_unknown	−0.703***	−0.412***	−0.268***	0.295***				−0.186***
subtype_unreliable	−0.703***	−0.512***	−0.511***	−0.178***		0.068		−0.205***
popularity	−0.0004	−0.002	0.023***	0.048***	0.042***	−0.018***	−0.046***	−0.021***
age_in_year	−0.0001	−0.009***	−0.024***	0.008***	−0.003	0.001	0.005*	−0.021***
Observations	1,644	1,644	1,644	1,644	1,644	695	724	1,644
p-value	0.62	0.21	0.18	0.17	0.053	0.057	0.067	0.173
Note:								*p<0.1; **p<0.05; ***p<0.01

Table 4: Model 1 assesses a domain's likelihood of being listed by a source (DDOT, POLIT, AGZ, MBFC, ZDR) given its i) ideology, ii) subcategory, iii) age, and iv) popularity. Model 2 examines characteristics that contribute to a domain's time of inclusion in sources ZDR and MBFC. Model 3 analyzes attributes correlated with the likelihood of a domain being defunct.

block is the amount of manual labor required[18]. List creators can reduce workload by using supervised machine learning models to classify unlabeled news domains into fake or mainstream provided that potential model biases are examined and understood [19].

Second, for list maintenance, we urge researchers to undertake the following tasks. First, it's valuable to i) document the exact timestamp when a domain is added, removed, updated (e.g., change of subcategory), or defunct in the list. Further, if a change is unusual (e.g., subcategory modification), creators should also ii) underline reasons for the change. Next, if the annotation process is updated (e.g., ZDR introduced more subcategories as the list expands), it's also integral to iii) keep both the initial and updated procedures separate, highlight the differences, and note the time of change. These tasks not only generate useful metadata that is required by various studies, they also make the maintenance process much more transparent, which can enhance the list's credibility and help researchers identify potential discrepancies or errors early on.

Lastly, given that top fake news stories in 2016 ostensibly target white, older conservative men and favor Trump over Clinton (Grinberg et al. 2018; Allcott and Gentzkow 2017), we posit that the ideological-leaning of fake news sites will be undoubtedly valuable to future work in this field, and propose that creators also include the meta-data and the relevant annotation process in the lists.

Guidance on Groundtruth Selection for List Users: First, researchers need to consider whether an analysis is directly affected by list size, as in the case of prevalence. Other types of analysis that depend on the nature of the fake news domain (as opposed to counts) are more robust to the choice (e.g., temporal and topical analysis).

The second consideration relates to which lists one should use for evaluation. We first observe that the choice of traditional news lists seems to not matter, thus reducing the effort to carry out research. Second, we also see consistent clustering of fake news lists across different analyses and we recommend selecting a list from each cluster. MBFC, AGZ, and ZDR are commonly clustered together (e.g., topic analysis latent space and prevalence). POLIT and DDOT are rather distinct from the rest. By selecting a list from each (e.g., MBFC and POLIT), researchers can determine informative bounds on their analyses. Finally, if the findings diverge, expanding the set of lists used as a function of (i) annotation and quality measure described in the meta-analysis and (ii) list clustering, i.e., considering the next most distinct list, can help explore this data space systematically.

[18]For instance, for MBFC, unaffiliated individuals first submit questionable websites which automatically go into *pending* status, then the staff will review each pending domain and reach a decision using existing annotation procedure. Given that mediabiasfactcheck.com currently has a backlog of 500+ domains in pending, it suffices to say that the process is painstakingly slow.

[19]For instance, creators can assess whether a model is biased against domains' i) ideology-leaning, ii) popularity, iii) age and iv) subcategory. Biases in a model may not automatically disqualify it from being employed, but documenting these biases can help future scholars using these lists and models better conceptualize how potential limitations may impact the validity of their studies.

Acknowledgement

This research was partly supported by Michigan Institute for Data Science (MIDAS) at the University of Michigan and the National Science Foundation (Grant IIS-1815875).

Appendix

Regression Model for Domain Inclusion For a given domain i that's listed by at least one f where $f \in \{\texttt{ZDR}, \texttt{MBFC}, \texttt{AGZ}, \texttt{DDOT}, \texttt{POLIT}\}$, let the binary variable $y_{i,s} =$

$\{0,1\}$ denote whether domain i exists in the list of fake news sites f. We fit model for each f using *ideology*, *subcategory*, *popularity*, and *age* as the explanatory variables.

$$y_f = \beta_0 + \beta_1 ideology + \beta_2 subtype + \beta_3 popularity + \beta_4 age + \varepsilon_i$$

Results are summarized on Table 4 (Model 1).

Regression Model for the Time of Addition We first use web.archive.org and authors' websites to obtain 3 timestamped snapshots [20] of ZDR, MBFC, and DDOT. Let i be a website that was added to ZDR in one of its 3 snapshots and remained on the list thereafter, we determine i's preferred *ideology*, *subcategory*, *popularity*, and *age*. Let the variable $y_{i,zdr} = \{0,1,2\}$ denote whether domain i was added in the 1st, 2nd, or 3rd version of ZDR, we fit the following:

$$y_{zdr} = \beta_0 + \beta_1 ideology + \beta_2 subtype + \beta_3 popularity + \beta_4 age + \varepsilon_i$$

We repeat the same procedure for DDOT and MBFC. Regression results are summarized on Table 4 (Model 2) [21].

Regression Model for Active and Defunct Domains For a given domain i that's listed by at least one f where $f \in \{\texttt{ZDR}, \texttt{MBFC}, \texttt{AGZ}, \texttt{DDOT}, \texttt{POLIT}\}$, let the binary variable $y_{i,s} = \{0,1\}$ denote whether domain i is defunct (i.e. $y = 1$ when i is no longer active). We fit model:

$$y_f = \beta_0 + \beta_1 ideology + \beta_2 subtype + \beta_3 popularity + \beta_4 age + \varepsilon_i$$

Results are summarized on Table 4 (Model 3).

References

Alimonti, K. R., and Veridiana. 2018. "fake news" offers latin american consolidated powers an opportunity to censor opponents.

Allcott, H., and Gentzkow, M. 2017. Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives* 31(2):211–236.

Allcott, H.; Gentzkow, M.; and Yu, C. 2018. Trends in the diffusion of misinformation on social media. *arXiv preprint arXiv:1809.05901*.

Beveridge, S., and Nelson, C. R. 1981. A new approach to decomposition of economic time series into permanent and transitory components. *Journal of Monetary economics* 7(2):151–174.

Bode, L.; Budak, C.; Ladd, J. M.; Newport, F.; Pasek, J.; Singh, L. O.; Soroka, S. N.; and Traugott, M. W. 2020. *Words That Matter: How the News and Social Media Shaped the 2016 Presidential Campaign*. Washington, D.C.: Brookings Institution Press.

Boulianne, S. 2015. Social media use and participation: A meta-analysis of current research. *Information, communication & society* 18(5):524–538.

Bovet, A., and Makse, H. A. 2019. Influence of fake news in twitter during the 2016 us presidential election. *Nature communications* 10(1):7.

[20] December 2016, June 2017, and December 2017 with each separated by 6 months

[21] Results for DDOT are removed given none of the variables are significant.

Budak, C.; Agrawal, D.; and El Abbadi, A. 2011. Limiting the spread of misinformation in social networks. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, 665–674. New York, NY, USA: ACM.

Budak, C. 2019. What happened? the spread of fake news publisher content during the 2016 u.s. presidential election. WWW '19.

Charemza, W. W., and Syczewska, E. M. 1998. Joint application of the dickey-fuller and kpss tests. *Economics Letters* 61(1):17–21.

Croft, M., and Moore, R. 2017. Checking what students know about checking the news.

Darwish, K.; Magdy, W.; and Zanouda, T. 2017. Trump vs. hillary: What went viral during the 2016 us presidential election. In *International Conference on Social Informatics*.

Doshi, A. R.; Raghavan, S.; Weiss, R.; and Petitt, E. 2018. The impact of the supply of fake news on consumer behavior during the 2016 us election.

Faris, R.; Roberts, H.; Etling, B.; Bourassa, N.; Zuckerman, E.; and Benkler, Y. 2017. Partisanship, propaganda, and disinformation: Online media and the 2016 us presidential election.

Farmer, L. S. 2017. Don't get faked out by the news: Becoming an informed citizen. In *The Fifth European Conference on Information Literacy (ECIL)*, 174.

Farte, G.-I., and Obada, D.-R. 2018. Reactive public relations strategies for managing fake news in the online environment.

Figueira, Á., and Oliveira, L. 2017. The current state of fake news: challenges and opportunities. *Procedia Computer Science* 121:817–825.

Funke, D. 2019. Want to get away with posting fake news on facebook? just change your website domain.

Goodall, D. W. 1966. A new similarity index based on probability. *Biometrics* 882–907.

Granskogen, T. 2018. Automatic detection of fake news in social media using contextual information. Master's thesis, NTNU.

Grinberg, N.; Joseph, K.; Friedland, L.; Swire-Thompson, B.; and Lazer, D. 2018. Fake news on twitter during the 2016 us presidential election. Technical report.

Guess, A.; Nyhan, B.; and Reifler, J. 2018. Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 us presidential campaign. *European Research Council* 9.

Guo, L., and Vargo, C. 2018. "Fake News" and Emerging Online Media Ecosystem. *Communication Research* 009365021877717.

Haugh, L. D. 1976. Checking the independence of two covariance-stationary time series: a univariate residual cross-correlation approach. *Journal of ASA* 71.

Hayes, A. F., and Krippendorff, K. 2007. Answering the call for a standard reliability measure for coding data. *Communication methods and measures* 1(1):77–89.

Helmstetter, S., and Paulheim, H. 2018. Weakly supervised learning for fake news detection on twitter. 274–277. IEEE.

Horne, B. D., and Adali, S. 2017. This Just In. *arXiv:1703.09398 [cs]*. arXiv: 1703.09398.

Horne, B. D.; Dron, W.; Khedr, S.; and Adali, S. 2018. Assessing the News Landscape. Lyon, France: ACM Press.

Jagarlamudi, J.; Daumé III, H.; and Udupa, R. 2012. Incorporating lexical priors into topic models. 204–213. Association for Computational Linguistics.

Koulolias, V.; Jonathan, G. M.; Fernandez, M.; and Sotirchos, D. 2018. Combating misinformation: An ecosystem in co-creation.

Kucharski, A. 2016. Post-truth: Study epidemiology of fake news. *Nature* 540(7634):525.

Lazer, D. M.; Baum, M. A.; et al. 2018. The science of fake news. *Science* 359(6380):1094–1096.

Leetaru, K., and Schrodt, P. A. 2013. Gdelt: Global data on events, location, and tone, 1979–2012. In *ISA annual convention*, 1–49. Citeseer.

Lima, L.; Reis, J.; Melo, P.; Murai, F.; Araújo, L.; Vikatos, P.; and Benevenuto, F. 2018. Inside the right-leaning echo chambers.

Main, T. J. 2018. *The Rise of the Alt-Right*. Brookings Institution Press.

Makridakis, S.; Wheelwright, S.; and Hyndman, R. 2008. *Forecasting methods and applications*. John wiley & sons.

Marcon, A. R.; Murdoch, B.; and Caulfield, T. 2017. Fake news portrayals of stem cells and stem cell research. *Regenerative medicine* 765–775.

McCombs, M. E., and Shaw, D. L. 1972. The agenda-setting function of mass media. *Public opinion quarterly* 36(2).

Misra, H.; Cappé, O.; and Yvon, F. 2008. Using lda to detect semantically incoherent documents. 41–48. Association for Computational Linguistics.

Mitchell, R. 2018. *Web Scraping with Python: Collecting More Data from the Modern Web*. " O'Reilly Media, Inc.".

Mukerji, N. S. 2018. A conceptual analysis of fake news.

Nelimarkka, M.; Laaksonen, S.-M.; and Semaan, B. 2018. Social media is polarized. 957–970. ACM.

Newman, D.; Lau, J. H.; Grieser, K.; and Baldwin, T. 2010. Automatic evaluation of topic coherence. 100–108. Association for Computational Linguistics.

Palen, L., and Hughes, A. L. 2018. Social media in disaster communication. In *Handbook of Disaster Research*. Springer.

Politifact staff. 2018. Politifact guide to fake news websites and what they peddle.

Potthast, M.; Kiesel, J.; Reinartz, K.; Bevendorff, J.; and Stein, B. 2017. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638*.

Poynter Institute. 2019. International Fact Checking Network. [Online; accessed -04-November-2019].

Puglisi, R., and Snyder Jr, J. M. 2011. Newspaper coverage of political scandals. *The Journal of Politics* 73(3):931–950.

Rapoza, K. 2017. Can 'fake news' impact the stock market?

Rehurek, R., and Sojka, P. 2011. Gensim–python framework for vector space modelling.

Rini, R. 2017. Fake news and partisan epistemology. *Kennedy Institute of Ethics Journal* 27(2):E–43.

Schaedel, S. 2018. Websites that post fake and satirical stories. factcheck.

Scheufele, D. A.; Kim, E.; and Brossard, D. 2007. My friend's enemy: How split-screen debate coverage influences evaluation of presidential debates. *Communication Research* 34(1):3–24.

Shao, C.; Ciampaglia, G. L.; Flammini, A.; and Menczer, F. 2016. Hoaxy. *WWW '16 Companion*. arXiv: 1603.01511.

Shao, C.; Ciampaglia, G. L.; Varol, O.; Flammini, A.; and Menczer, F. 2017. The spread of fake news by social bots. 96–104.

Shu, K.; Sliva, A.; Wang, S.; Tang, J.; and Liu, H. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter* 19(1):22–36.

Shu, K.; Wang, S.; and Liu, H. 2018. Understanding user profiles on social media for fake news detection. In *2018 IEEE Conference on (MIPR)*, 430–435. IEEE.

Silverman, C. 2016. Here are 50 of the biggest fake news hits on facebook from 2016.

Starbird, K. 2017. Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on twitter. In *ICWSM*, 230–239.

Stock, J. H.; Watson, M. W.; et al. 2003. *Introduction to econometrics*, volume 104. Addison Wesley Boston.

Tandoc Jr, E. C.; Lim, Z. W.; and Ling, R. 2018. Defining "fake news" a typology of scholarly definitions. *Digital Journalism* 6(2):137–153.

Van Zandt, Dave. 2018. Media bias/fact check (mbfc news).

Vargo, C. J.; Guo, L.; and Amazeen, M. A. 2018. The agenda-setting power of fake news. *new media & society* 20(5):2028–2049.

Vosoughi, S.; Roy, D.; and Aral, S. 2018. The spread of true and false news online. *Science* 359(6380):1146–1151.

Wang, W. 2017. "Liar, Liar Pants on Fire". *arXiv:1705.00648*.

Wardle, C. 2017. Fake news. it's complicated. *First Draft*.

White, N. 2018. The daily dot.

Wikipedia contributors. 2019. Alexa internet — Wikipedia, the free encyclopedia. [Online; accessed 4-May-2019].

Wold, S.; Esbensen, K.; and Geladi, P. 1987. Principal component analysis. *Chemom. Intell. Lab. Syst.* 2(1-3):37–52.

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