

Detection and Discovery of Misinformation Sources Using Attributed Webgraphs

Peter Carragher, Evan M. Williams, Kathleen M. Carley

Carnegie Mellon University
4665 Forbes Ave, Pittsburgh, PA 15213
{pcarragh, emwillia, carley}@andrew.cmu.edu

Abstract

Website reliability labels underpin almost all research in misinformation detection. However, misinformation sources often exhibit transient behavior, which makes many such labeled lists obsolete over time. We demonstrate that Search Engine Optimization (SEO) attributes provide strong signals for predicting news site reliability. We introduce a novel attributed webgraph dataset with labeled news domains and their connections to outlinking and backlinking domains. We demonstrate the success of graph neural networks in detecting news site reliability using these attributed webgraphs, and show that our baseline news site reliability classifier outperforms current SOTA methods on the PoliticalNews dataset, achieving an F1 score of 0.96. Finally, we introduce and evaluate a novel graph-based algorithm for discovering previously unknown misinformation news sources.

Introduction

In 2014, Sinclair Treadway and Sean Adl-Tabatabai founded YourNewsWire, which over time was identified by many fact-checkers as a purveyor of misinformation. In 2019, the co-founders, blaming loss of revenue on Facebook’s fact-checking system, rebranded their website by simply migrating the domain name from “yournewswire.com” to “newspunch.com” (Binder 2019).

The detection of misinformation sources among news domains relies heavily on preexisting domain reliability labels. However, these curated domain lists are subject to two core limitations. Firstly, as illustrated above, building automated systems to detect websites that consistently spread misinformation is made challenging by the ease of the evasion and exit tactics available to bad actors. As researchers and fact-checking organizations identify and release misinformation sources, the owners can shut down the site, migrate to a new domain, or simply start over with a new website. Consequently, we observe that over 50% of domains on unmaintained blocklists published between 2017 and 2019 are dead. This presents a major obstacle for misinformation source detection and discovery tasks.

Secondly, the vast majority of misinformation source research relies on signals mined from public social media data.

As misinformation becomes increasingly prevalent in developing countries where social media usage remains uncommon or where English is not the primary language, detection and discovery methodologies that are not dependent on website content or social media are needed (Okereke et al. 2021).

We attempt to address these obstacles by demonstrating the predictive power of webgraph and Search Engine Optimization (SEO) data in detecting news site reliability and bias. This paper does not rely on social media data to identify unreliable domains, but rather relies on pointers between websites. We provide three primary contributions in this work. Firstly, we introduce a novel attributed webgraph dataset with labeled news domains and their connections to top outlinking and backlinking domains, i.e. domains to which a target domain links and domains that link to the target domain. Secondly, with extensive experiments, we demonstrate the predictive power present in the data by using graph neural network (GNN) models to predict the reliability and bias of reliable and unreliable news domains. We show that these models outperform existing work on the PoliticalNews dataset (Castelo et al. 2019). We make all data¹ and code public².

Finally, we present a novel webgraph-based discovery approach that is not dependent on language, content, or social media data, and we demonstrate the success of the approach in identifying previously-unknown unreliable news domains. The content-agnostic and social media-agnostic nature of this work differentiates our approach from all previous work on this topic and provides several key advantages. Firstly, this allows our approach to be implemented in non-English settings. Secondly, our approach is not reliant on API access to social media data. Finally, our content agnostic system will not be impacted by increasing use of generative AI in SEO settings, which has been identified as an emerging threat (Petrou et al. 2023).

¹<https://doi.org/10.1184/R1/25174193.v1>

²<https://github.com/CASOS-IDeAS-CMU/Detection-and-Discovery-of-Misinformation-Sources>

Related Work

Misinformation Detection

Classifying individual texts or articles is a core problem addressed by research in misinformation detection. Castelo et al. (2019) address misinformation page detection using article content which Chen and Freire (2020) extend to unreliable domain discovery using social media context. Building on that work, Silva et al. (2021) combine website content and social context resulting in a misinformation page detection method that leverages multi-modal data.

We depart from the article and content-centered approaches to misinformation detection in that our detection systems are at the domain-level. While a single article may be considered misinformation, the collection of articles published by a single news domain has an associated probability distribution that captures how likely an article is to contain misinformation, given the source. In content-centered approaches, reliability scores are assigned to domains based on this distribution, labeling domains that frequently publish articles containing misinformation as “unreliable”. As we are concerned with misinformation *sources*, we discuss unreliable domain detection rather than misinformation *page* detection, using ‘misinformation source’ and ‘unreliable domain’ interchangeably.

Motivating our domain-level and content-agnostic approach, Petrou et al. (2023) employed change-point detection to reveal that the linguistic features outlined in Castelo et al. (2019) evolve over time. They find significant variance in the psychological features that distinguish reliable articles from those peddling misinformation, making unreliable page detection difficult over long periods of time. Our webgraph methods present a novel approach to unreliable domain detection and discovery which avoids common pitfalls of the content-based approach.

Webgraphs

Several studies have explored the predictive power of webgraphs on tasks related to unreliable domain detection. Using community detection methods, Aires, G. Nakamura, and F. Nakamura (2019) scrape links from US and UK news sites and find clusters based on political bias using labels from **Media Bias Fact Check (MBFC)** (Zandt 2022). Similarly, Hrckova et al. (2021) found that partisan news domains in central Europe form clusters in webgraph data. Sehgal et al. (2021) present a case study on hyperlink and social media usage for peddling misinformation. They explore coordinated use of hyperlinks and social media on a dataset of 1.4k URLs with reliability labels drawn from various sources, including MBFC. Lee et al. used the linkage patterns of labeled news domains shared by Twitter users to visualize connections of domains to unreliable domains on Twitter (Lee et al. 2022). These analyses provide theoretical justifications for using webgraph data to detect website reliability and bias.

Search Engine Optimization

SEO is an 80 billion dollar industry built to increase traffic to websites (McCue 2018). High-ranking pages are far

more likely to be seen. In a 2013 analysis of 300 million search engine clicks: 92% were on first page of search results with the majority of clicks going to just the top two (Chitika 2013). Furthermore, users were found to be 140% more likely to click the last result on the first page than the first result on the second page (Chitika 2013). Consequently, SEO has become ubiquitous. On a study of over 250k search results, Lewandowski, Sünkler, and Yagci (2021) found that only 12% were likely not to have been optimized, and 58% were certainly optimized. Zhang and Cabage (2017) contrast the effectiveness of SEO and social media-sharing for media promotion. They find that while social media-sharing results in more immediate boosts to traffic in the short term, link-building is superior in the long term. In this work we elect to focus on link building, leaving the role that social-sharing strategies plays in boosting unreliable new sources to future work.

Black-hat SEO While many SEO practices are endorsed by search engines, those aimed at gaming recommendation algorithms are forbidden by search engine webmaster guidelines³—these tactics are typically referred to as “black-hat” SEO. For a detailed overview of white-hat and black-hat SEO methods, we refer the reader to the work done by Malaga (2010). Common black-hat SEO techniques of relevance here are: (1) attracting search engine indexers through networks of blogs (‘Blog Ping’), (2) automated user engagement in forums or comment sections (‘User Generated Content’), and (3) link building by spamming (‘Link Scheming’). More recently, Aswani et al. (2018) analysed the sentiment of SEO service customers and found that 32% of respondents believed that the widespread use of black-hat SEO methods by SEO service providers was damaging their ranking. Aswani et al. (2021) suggest that clustering domains based on high level SEO attributes such as Pagerank score and Domain Authority can identify black-hat link scheme sites. This motivates our approach to identifying link scheme sites, which is the first step in our network-based unreliable domain discovery process (Algorithm 1).

SEO and Misinformation Search engine rankings have broad impacts even outside of e-commerce; it has been shown that Search engine rankings can have a substantial impact on the political beliefs and voting patterns of users (Epstein and Robertson 2015). Three laboratory experiments with double-blind control group design found that relatively minor changes in search engine rankings could influence decisions of undecided voters⁴. Search engine audit studies—studies where researchers query terms on multiple search engines and compare the rankings—have revealed a link between SEO and misinformation. For example, Urman et al. (2022) executed conspiratorial searches on Google, Bing, DuckDuckGo, and Yandex and found that

³<https://developers.google.com/search/docs/essentials/spam-policies>

⁴The magnitude of the effects in this work are disputed by Katharina Zweig, who finds that the cumulative effect is likely on the order of 2-4% rather than the 20% reported in the study (Zweig 2017).

Yandex and DuckDuckGo consistently return misinformation from unreliable sources in their top search results. Bradshaw (2019) analyzed how a set of 29 junk news sites optimize keywords to increase web-traffic and spread disinformation over a three year period. More recently, Williams and Carley (2023) demonstrated the usefulness of webgraph data in identifying large-scale link scheme activities around Kremlin-aligned propaganda domains. We build on these efforts in an attempt to link specific black-hat SEO tactics to unreliable news sources.

Domain Monitoring

Weinberg et al. (2017) find that older website blocklists have a high rate of parked domains—domains that are up for sale and not currently in use. Domain survival rates are estimated using a classifier that is trained to flag parked domains (Visser, Joosen, and Nikiforakis 2015; Szurdi et al. 2014). We note that this limitation of blocklists has *not* been recognized by previous work on unreliable domain lists.

Tasks

We evaluate our approach using three classification tasks. 1) *Reliability Prediction*: a binary classification task determining domain reliability. 2) *Relative Political Bias*: binary “left”–“right” partisan lean classification. 3) *Absolute Political Bias*: a binary “center”–“extreme” task determining whether the site publishes extremist or centrist content.

Previous research and government reporting has demonstrated how IRA trolls amplify both extreme left and extreme right sources to sow discord and increase polarization (Bradshaw, DiResta, and Miller 2022). Training separate classifiers for left-right and extreme-center dimensions of political bias enables a more fine-grained discovery approach.

Given the transient nature of many misinformation domains, classification of existing labeled domains may be of limited value. We therefore introduce a final task, 4) *Misinformation Domain Discovery*, to allow us to identify new misinformation domains. We split this process into two conceptual steps: 1) Develop a set of criteria for identifying suspicious backlinking patterns and 2) Extract unreliable websites to which suspicious websites link.

Data

News Site Identification

To identify a starting set of news sources for our tasks, we leverage information from two different sources. The bulk of our domain list is made up of domain labels scraped from MBFC⁵ (Zandt 2022). MBFC has six grades of reliability, ranging from very low to very high. MBFC also provides political bias labels ranging from extreme left (-2) to extreme right (+2), the distribution of which is given in figure 1.

Domains collected by Grinberg et al. (2019) were part of an investigation into the role of social media in spreading misinformation during the 2016 US Presidential Elections. As part of their analysis, they publish lists of unreliable sites from various investigative sources; Buzzfeed, Politico,

and Snopes.com, as well as domains identified in their own analysis. They also draw on a blocklist built by Allcott and Gentzkow (2017), which published a dataset of 125 URLs extracted from articles scraped from Politico, Snopes.com, and Facebook. URLs are extracted based on keywords pertinent to the 2016 election. We combine the MBFC data with these previous lists (Grinberg et al. 2019) for a total of 4,206 domains. A recent study on the correlation of news reliability ratings constructs a similar list (Lin et al. 2022). Table 1 gives a count of domains per source.

Domain List Analysis

We investigate the survival rates of domains from previously-published unreliable domain blocklists, and filter out dead domains. For each domain we determine whether the domain responds to GET requests and whether the domains are parked, i.e., are not being held for sale by a parked domain registrar. If a GET request returns a 404, the site is dead. If the domain is parked or dead, then it should no longer be included in the dataset.

To determine whether domains were parked, we build a boosted decision tree classifier following the methodology proposed in Vissers, Joosen, and Nikiforakis (2015). For more information on the Parked Domain classifier methodology, see the appendix. When applied to our list of domains, the model identified 248 parked domains with 96% precision (based on a manual evaluation on a 10% sample).

In Table 1, we show that the older labeled sources from Grinberg et al. (2019) have low survival rates. MBFC is continuously-maintained, so it has high survival rates. With the exception of the Snopes.com list (which is drawn from the dedicated fact-checking site), the non-maintained lists have survival rates of less than 50%. The number of links that point to a site is also an indication of the status of a site (Ahrefs 2022). We observe that simply dropping news websites with less than 10k backlinks excludes the majority of the dead domains: it filters out 63% of domains returning 404s and 87% of parked domains. We postulate that this makes filtering by backlinks a reasonable heuristic for the parked domain classifier and the 404 classifier in unreliable domain discovery. This further simplifies our method such that parked domain and 404 classifiers are not needed.

In summary, we filter out any sites that have less than 10k backlinks as those are unlikely to be influential news sites. Our final domain reliability dataset contains 3,211 labeled domains which we will refer to as **MBFC*** (Table 1).

Source	URLs	!404	!Parked	Both	>10k
MBFC	3227	97%	99%	96%	87%
MBFC-Q	426	90%	94%	84%	67%
Snopes	55	80%	84%	67.5%	69%
Blocklist	359	74%	59%	43.5%	11%
Politico	83	65%	68%	44%	18%
Buzzfeed	55	70%	50%	35%	11%
Total	4205	3910	3952	3658	3211

Table 1: Unreliable Source Survival Rates

⁵<https://mediabiasfactcheck.com/methodology/>

MBFC* Site Labels

To unify labels from each source, we needed to binarize labels. There are some sites that MBFC classifies as ‘questionable’, but it does not assign these sites a reliability score⁶. After inspecting a sample of these domains, we chose to assign these questionable websites an unreliable label. We handle overlap between lists by taking the most recent label available. This results in current MBFC labels having precedence over the labels assigned by earlier works.

The blocklist domains label unreliable websites as black (very low reliability), red (low), and orange/yellow (mixed), which we map to the corresponding labels used in MBFC. Remaining consistent with Chen and Freire (2020), any domain with a reliability rating of mixed or lower is assigned an unreliable label, the rest are assigned reliable labels.

For the bias tasks we use only the labels from MBFC as the blocklist sites do not have bias labels (Grinberg et al. 2019)). Using the >10k backlink criteria for filtering out dead domains, we are left with 2,629 bias labels. For the relative political bias task, we label the domain left if the MBFC bias score is less than 0, right if the bias score greater than 0. Labels of value zero, i.e., centrist labels, are dropped for this task. For the absolute bias task, bias scores of -1, 0, and 1 are assigned the centrist label; otherwise we label it extreme.

Our two final datasets contain 3,211 reliability-labeled domains and 2,629 bias-labeled domains. In Figure 1 we visualize the distributions of political bias labels for each news site reliability label. Aside from a 2:1 ratio of reliable to unreliable domains, we note reliable domains tend to skew centrist, whereas unreliable domains tend to skew towards political extremes. Exploratory data analysis shows that SEO attributes align with our intuition with respect to website reliability (see Figure 9 in the appendix). For example, we find a positive correlation between the binary reliability labels and the proportion of backlinks that come from .edu domains ($r = 0.09$), as well as from .gov domains ($r = 0.08$).

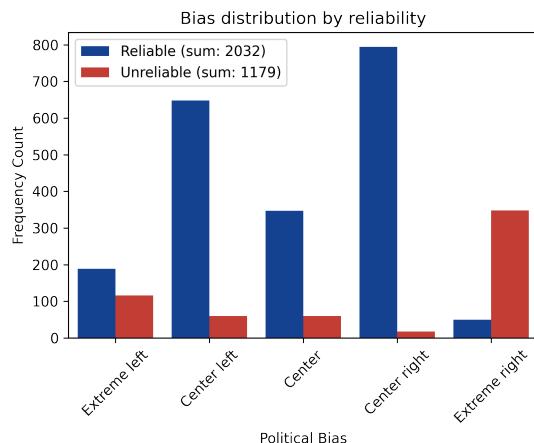


Figure 1: Bias label counts grouped by reliability.

⁶<https://mediabiasfactcheck.com/fake-news/>

PoliticalNews

The PoliticalNews dataset is a popular evaluation dataset in fake news detection studies. It was proposed by Castelo et al. (2019) alongside the Topic-AGnostic (TAG) classifier for use in misinformation page detection. It has since been adopted by Chen and Freire (2020) to evaluate unreliable domain detection methods. It consists of 137 domains, including 58 news domains and 79 unreliable domains collected between 2013 and 2018. As other detection methods have published results on this evaluation set (Castelo et al. 2019; Chen and Freire 2020), it serves as a useful benchmark for our SEO-based detection method.

Site Attributes

We pull SEO attributes for both MBFC* and Political-News using Ahrefs—a proprietary search engine optimization toolkit that purports to possess a 13 trillion link database and the third most active commercial webcrawler after Google⁷. We pull 24 different attributes for each domain in our dataset, including the number of .edu domains that link to the website, total number of backlinks, and whether or not the domain has links from user-generated content (Figure 5).

Constructing a Webgraph

We construct webgraphs for the MBFC* dataset using top backlink and outlink domain data from Ahrefs. As a result of API and monetary constraints, we limit the webgraph to a single hop. We extract three networks for each labeled domain: a backlink network (top 10 backlinks per news site), an outlink network (top 10 outlinks per news site), and a combined network (the union of the top 10 backlink and outlink networks). Once data were pulled, we constructed weighted and attributed Graphs (webgraphs), $\mathcal{G} = V, E$ where V are websites and E are represent either backlinks or outlinks of domains. We visualize the backlink network colored by news-site reliability in Figure 2. Visually, unreliable and reliable domains seem to largely cluster together. This visual separability suggests that we should be able to use network signal to classify reliability with reasonable accuracy. For summary statistics on the three networks, see Table 7 in the appendix.

Misinformation Source Detection

Flat Baselines For our flat baseline experiments we use only the SEO attributes for each URL, ignoring network structure as well as the top backlinking and outlinking domains. We then run standard flat ML baselines: gradient boosted decision tree (GBDT), random forest, decision tree, MLP, and SVM. For parameters we use 50 estimators for the random forest classifier, 2 hidden layers of dimension 200 for the MLP and the linear kernel for the SVM. Scikit-learn defaults are used for the remaining parameters. We repeat these baselines for the reliability, absolute political bias, and relative political bias tasks on the labelled dataset. Five-fold cross validation is used with an 80:20 training-test set split.

⁷<https://ahrefs.com/robot>

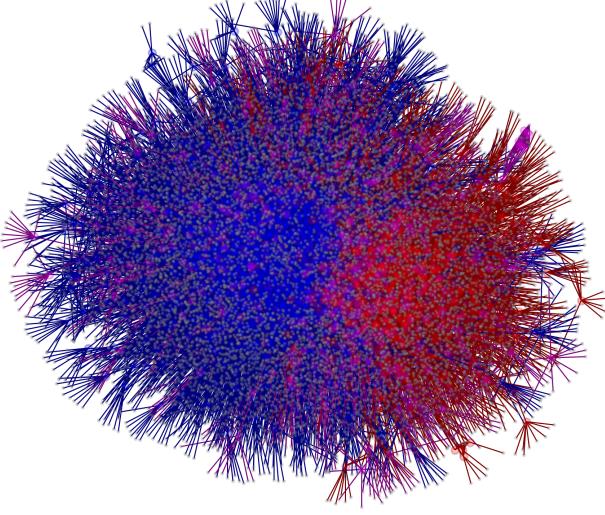


Figure 2: Backlink network where node colors show reliability labels; red are low reliability, blue are high reliability, pink are mixed reliability, grey are backlinking domains.

GNNs For our GNN experiments, our goal is to leverage local homophily present in our partially-labeled SEO network in order to better classify unreliable and biased domains. Formally, for a given task, for website u , our goal is to create node embeddings $z_u \in \mathbb{R}^d$ that map u to its corresponding one-hot-encoded label $y_u \in \mathbb{Z}^c$. The layer-wise propagation of a Graph Convolutional Neural Network (GCN) (Kipf and Welling 2016) can be written as:

$$H^{(l+1)} = \sigma(\tilde{D}^{-1/2} \tilde{A} \tilde{D}^{-1/2} H^{(l)} W^{(l)}) \quad (1)$$

where σ is an activation function, \tilde{A} is the sum of the Adjacency matrix A and its identity matrix I , \tilde{D} is the diagonal degree matrix of \tilde{A} , $H^{(l)}$ is the l^{th} layer activation matrix, and $W^{(l)}$ is a trainable weight matrix.

For each task, we implement a homogeneous two layer GCN with 64 output channels that feed into a final linear layer with relu activation and a dropout of 0.5. We found that using GCN layers outperformed GraphSage (Hamilton, Ying, and Leskovec 2017) and GAT (Veličković et al. 2017) layers on our baseline architecture, so we chose to use GCN layers for all subsequent experiments. We use crossentropy loss function and and an Adam optimizer. A log softmax activation is applied to the final layer.

The graph is partially labeled as many backlinks and outlinks do not have reliability or bias labeled. Consequently, unlabeled nodes are masked, but their features are still available to labeled neighbors during the propagation step. To train, we use a transductive random 80:10:10 split and 0.05 for the learning rate. We select models using early stopping with a patience of 30 and a minimum delta of 1e-4.

Link Normalization We investigate several link normalization schemes, as GNN-based models have been shown to work better when edge weights are normalized. The link data can be represented as an Adjacency Matrix A where

$A_{i,j}$ represents the number of links from domain i to j , and the (diagonal) *backlink* degree matrix D_b , and *outlink* degree matrix D_o , where $D_{i,i}$ represents the backlink and outlink total for domain i respectively. As we only observe links from the 10 highest-volume backlinking and outlinking domains, we also have the sampled degree matrices \hat{D}_b and \hat{D}_o , with diagonal elements representing the sum of all incoming and outgoing links observed *in the graph*. We explore 7 different edge-weighting approaches:

- *links*: $A_{i,j}$
- *log links*: $\log(A_{i,j})$ iff $A_{i,j} > 0$
- *backlink*: $D_b^{-1} A$ — i.e. the % of j 's backlinks from i
- *outlink*: $A D_o^{-1}$ — i.e. the % of i 's outlinks that go to j
- *graph-backlink*: $\hat{D}_b^{-1} A$
- *graph-outlink*: $A \hat{D}_o^{-1}$
- *page*: % of j 's reference pages from i 's domain

Backlink and outlink edge weightings of the form AD^{-1} and $D^{-1}A$ follow the random-walk normalization method. We also include *page*, an edge weight calculated on reference page count—the number of unique pages on a domain that link to a target domain. Intuitively, when $page_{i,j}$ is near 0, links from domain i to j come from a narrow subset of domain i subpages relative to the total number of webpages that reference domain j .

Detection Results

Given the label imbalance on the MBFC* unreliable classification task, we evaluate models using Binary F1 where “unreliable” is the positive label (Table 2).

On the PoliticalNews domain list, we replicate the TAG classifier using data released by the authors (Castelo et al. 2019) and show that models trained on SEO features consistently outperform those trained on TAG features. The SEO GBDT model outperforms TAG GBDT accuracy and recall by 14%, achieving 96% accuracy, 0.96 F1. Furthermore, the TAG paper reports an accuracy of 83% for a linear SVM classifier (Castelo et al. 2019), which our linear SVM outperforms by 10%.

The TAG dataset assigns domain-level labels to article-level, content-based features, so every article from a given domain has the same label, but different features. On the other hand, SEO features exist at the domain level—each article is assigned the SEO features of its domain, and so every article from a given domain has the same label and the same SEO features. As the sample of articles in the PoliticalNews dataset is drawn such that it is representative of the domain labels (Castelo et al. 2019), it is not surprising that domain-level SEO features outperform TAG features on these labels.

Of the flat models trained on the MBFC* dataset, GBDT consistently yielded the highest F1 on the reliability and absolute bias tasks and was comparable to Random Forest on the relative bias task. Performance on MBFC* is understandably lower than for PoliticalNews, as it’s domain list is

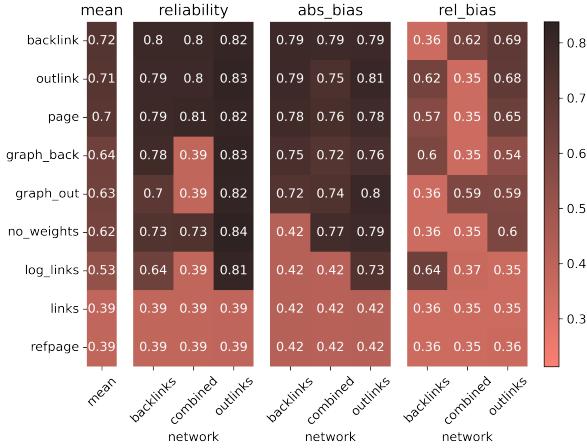


Figure 3: F1 scores for webgraph design space exploration, using GCN with various network structure and link weighting approaches.

an order of magnitude larger and so the task is more challenging. This is also shown by SVM’s success on PoliticalNews (SEO F1 = 0.93) vs. its failure across all three tasks on MBFC* ($F1 \leq 0.01$).

The GNN models significantly outperform the flat classifiers on every task across a range of link networks and weight schemes. GNN_{back} and GNN_{out} are the GNNs trained with *backlink* and *outlink* weight normalization and schemes. Due to the smaller dataset size, the GNN methods are not suitable for the PoliticalNews dataset. However, similar GBDT feature importances are seen for both datasets on the reliability tasks, suggesting that classifiers trained on SEO features have generality (Figure 5).

Webgraph Experiments Figure 3 provides F1 scores for each combination of network structure and link weighting approaches on the established detection tasks. We observe that the outlink network achieves the strongest overall performance. The combined network often results in the lowest performance, particularly on the relative bias task. Additionally, we observe that link normalization approaches have a substantial impact on model performance. The weighting approaches in Figure 3 are sorted in descending order from highest (top) to lowest (bottom) by the mean F1 score across all experiments. The backlink weighting achieves the highest mean F1 score. Interestingly, the highest F1 score on the reliability task is obtained by the outlink network with no link weights. However, in general the decision to normalize is more important than the exact normalization method employed, and further performance increases with more appropriate weighting approaches have smaller gains. It is also worth noting that the graph normalized variants, with sampled degree matrices \hat{D} , underperform compared to the backlink and outlink weightings.

We find the inclusion of webgraph and SEO context is strongly related to the overall performance of our models. Figure 4 shows the effects of including a varying number of backlinks in the backlink network across each task (the top

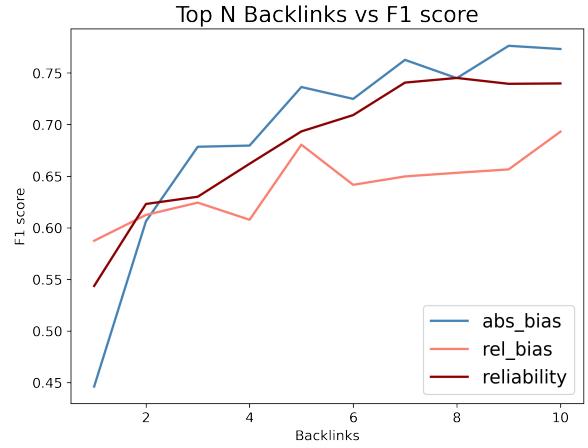


Figure 4: Increasing backlink context improves the performance of our models, to a point.

N backlinks, varying between 1 and 10). We use log weights, as we found those weightings to be the most intuitive as we vary N. We observe that, for all tasks, as the number of backlinks increases, performance tends to increase as well. Additionally, as we increase the number of backlinks used we observe diminishing returns. The F1 score plateaus at 0.76 with top 7 backlinks for the reliability task, 0.78 with 9 backlinks for the absolute bias task, and 0.68 with 10 backlinks for the relative bias task.

Black-Hat SEO We find evidence that classic black-hat SEO techniques disproportionately support unreliable news sites. Figure 8 in the appendix shows the negative correlation between reliability and percentage of backlinks coming from blogspot domains ($R = -0.25$). This is indicative of “Blog Ping”, where link farming—an SEO tool to generate backlinks—is done on domains hosted by top blogging sites, because such sites are known to quickly attract web crawlers for indexing (Malaga 2010). Similarly, black-hat link building is often done through user generated content—namely forums and comment sections (Malaga 2010). We observe similar negative correlations between the ‘ugc’ user-generated content attribute & reliability ($R = -0.1$).

Misinformation Source Discovery

Given the webgraph of a labeled seed list, we seek to capture a subset of nodes which maximizes the likelihood of nodes within that subset being unreliable. As described in Figure 6, our method is a two-part process based upon the hypothesis that backlink sites that heavily link to known misinformation sources also link to unknown sources.

Link Scheme Identification In the first stage of the discovery process, we attempt to capture backlinking domains with suspicious linking patterns. We extract the subset of backlinking sites that link to multiple sites labeled as unreliable (breadth constraint, β). We further parameterize by the total number of backlinks targeting unreliable sites (depth constraint, α). Algorithm 1 details this process.

	PoliticalNews				MBFC*					
	SEO		TAG		Reliability		Abs. Bias		Rel. Bias	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
GNN _{to}	-	-	-	-	84	.82	85	.79	70	.69
GNN _{sb}	-	-	-	-	82	.79	83	.79	66	.62
GBDT	96.2	0.96	82.5	0.82	83.5	.76	83.7	.65	71.4	.65
RF	91.3	0.92	78.3	0.76	82.2	.74	82.9	.63	71.5	.66
DT	93.6	0.94	80.7	0.8	77.0	.69	76.5	.56	63.5	.59
MLP	81.0	0.81	82.1	0.81	61.4	.35	62.5	.14	54.6	.22
SVM	92.7	0.93	80.1	0.8	63.3	.01	73.2	.00	55.7	.00

Table 2: Comparison of flat and graph-based detection methods on each of the MBFC* and PoliticalNews (Castelo et al. 2019) domain lists. For MBFC* we report F1 and accuracy on three sets of labels. For the PoliticalNews reliability labels, we compare the performance of models trained on the original TAG dataset (article-level) and our SEO dataset (domain-level).

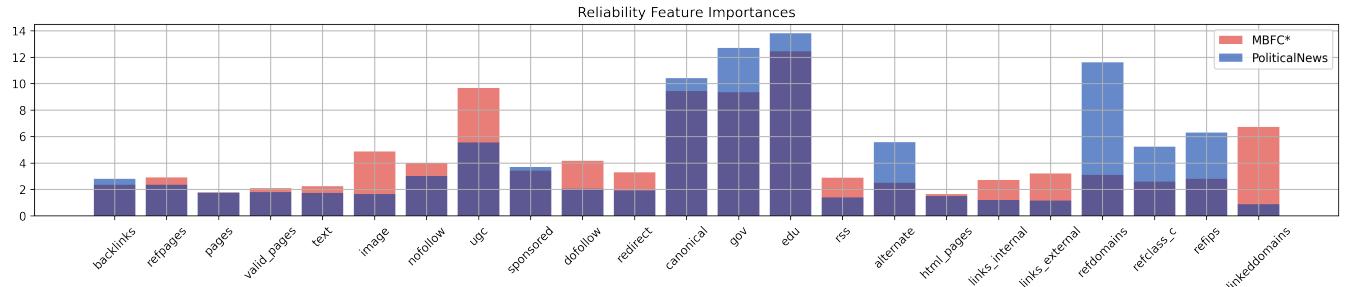


Figure 5: SEO attribute importances from Ahrefs on predicting reliability labels on the PoliticalNews and MBFC* datasets. For a full description of features used, see <https://ahrefs.com/api/documentation/metrics-extended>.

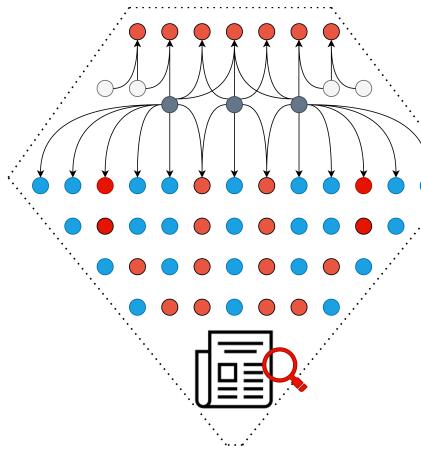


Figure 6: Diagram of the multi-step domain discovery process. We begin with a list of known unreliable domains (red)(1), then analyze the backlink network (white) to uncover link schemes (grey)(2), and pull the outlinks (blue) for these link schemes (3). To filter out noise, we apply a news classifier (4), a bias classifier (5), and a reliability classifier (6). Finally, we validate the results.

Algorithm 1: The link scheme identification algorithm

```

Require:  $\beta_{min} \geq 0$                                 ▷ Breath criteria
Require:  $\alpha_{min} \geq 0$                                 ▷ Depth criteria
Input:  $G \leftarrow (V, E)$                             ▷ Backlink Network
 $\mathcal{C} \leftarrow []$                                      ▷ Link Scheme Candidates
for each  $(s, t) \in G.edges()$  do
    if  $t.reliability = 0$  then ▷ Unreliable News Sources
         $C.append(s)$ 
    end if
end for
 $L \leftarrow []$                                          ▷ Link Schemes
for each  $c \in \mathcal{C}$  do
     $S \leftarrow G.successors(c)$ 
     $\beta \leftarrow \sum_s^S 1$ 
     $\alpha \leftarrow \sum_s^S s.backlinks$ 
    if  $\beta \geq \beta_{min} \wedge \alpha \geq \alpha_{min}$  then
         $L.append(c)$ 
    end if
end for

```

News Classification To identify candidate news domains, we attempt to classify news domains in link scheme outlinks. Link scheme outlinks are noisy and can contain link farms, businesses, and other irrelevant sources. We train a classifier on SEO features to determine which link scheme outlinks are news sources. We draw positive samples from our MBFC* dataset and we draw negative samples randomly from the Common Crawl dataset⁸. We found that a GBDT with 50:50 negative to positive sample ratio obtained the best results, obtaining 89% accuracy and 0.87 binary F1 score. Increasing the number of negative samples from common crawl had little effect.

Reliability Classification To identify which news site candidates are unreliable, we use our best-performing flat classifiers. Specifically, we apply both the reliability and absolute bias classifiers. Although we expect the GNNs to outperform the flat baselines on the reliability task, inference with a GNN requires building a link network for the URLs on which we run inference. As the reliability classifier is trained only on domains with >10k backlinks, we use the same criteria to filter candidate domains because we expect the classifier’s performance to be inconsistent outside this range. Finally, we manually evaluate the reliability of the candidates, using translation tools where necessary.

Discovery Results

We run our discovery process for two datasets, the PoliticalNews evaluation set (Castelo et al. 2019) and our MBFC* dataset. Parameters along with resulting domain counts are provided in Table 3.

Evaluation on the PoliticalNews dataset To evaluate the performance of our discovery system, we employ the partial F1 score metric as proposed by Chen and Freire (2020). We find that the partial F1 of our webgraph-based system (0.28) is on par with the best reported partial F1 (0.29) of the content and social media based system outlined in Chen and Freire (2020), with both systems achieving a partial F1 of 0.25 for a large number of system configurations. The partial F1 metric defines the set of true positive unreliable domains as those labeled mixed reliability or worse by MBFC. We acknowledge that these results are not directly comparable as the current work uses a more up-to-date list of domain labels from MBFC. Rather, these results are meant to give an indication of performance for a system that is both challenging to evaluate and novel in terms of the methods used (i.e. webgraph vs content and social media based).

Figure 7 shows results from a design space exploration of the discovery algorithm. We find that optimal parameters with respect to partial F1 on the PoliticalNews dataset are $\alpha_{min} = 2500$, $\beta_{min} = 2$, and 100 outlinks per link scheme. Most notably, we find that our system is recall-bounded. That is to say that as we increase recall, by lowering the β_{min} criteria (finding more link schemes) and increasing the number of outlinks used per link scheme, we increase the partial F1 score. In doing so, we observe diminishing returns, analogous to our GNN experiments showing

how F1 increases with the number of backlinks used in the webgraph for reliability classification (figure 4).

Evaluation of PoliticalNews Webgraph Discovery

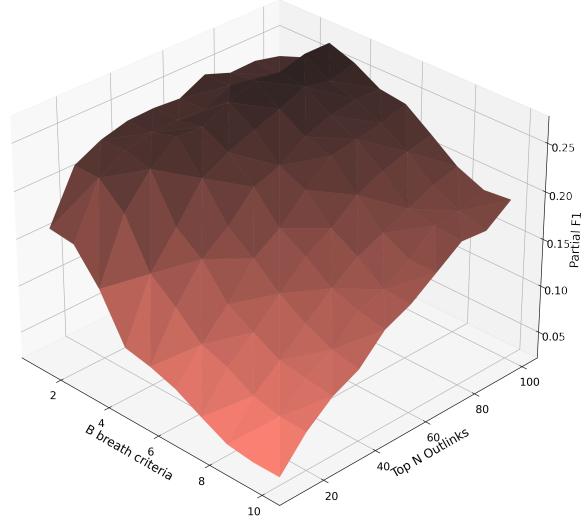


Figure 7: Link Scheme Identification; the x-axis represents the # of outlinks used for each link scheme site, the z-axis shows the # of misinfo sites linked to (β_{min}), and the y-axis shows partial F1 score (Chen and Freire 2020).

System Demonstration on our MBFC* dataset In order to run the discovery process for the full list of domains in table 1, we select parameters based on the optimal parameters from the design space exploration on the PoliticalNews dataset in figure 7. However, certain parameters are dependent on the size of the seed list; the depth constraint α_{min} represents the minimum number of links to unreliable domains required for a link scheme. Since our MBFC* dataset has an order of magnitude more seed domains, our criteria must increase to reflect this. In Table 3 we provide the parameters used in our discovery process. From left to right, we follow the order of execution in Algorithm 1: number of backlinks per known unreliable domain, α_{min} and β_{min} link scheme criteria, L link schemes, number of outlinks per link scheme, and MC misinformation source candidates. Included are resulting domain counts after each stage in the discovery process from the size of the initial seed lists, to the final set of MC misinformation source candidates that are classified as politically extreme, unreliable news domains. Brackets imply the range of parameters used in experimentation.

We observe evidence supporting the hypothesis of the discovery method: that link scheme sites which link to known unreliable sites also link to unknown unreliable sites. Table 4 provides the rate of unreliable domains in each explored list, as approximated by the reliability classifier. We find that link scheme outlinks contain a higher proportion of domains classified as unreliable than a random sample of 3241 domains from Common Crawl and the combined outlink and

⁸<https://commoncrawl.org/>

	Seed	Backlinks	α_{min}	β_{min}	L	Outlinks	MC	& News	& Misinfo	& Biased
PNews	79	100	2.5k [1-10k]	2 [1-10]	132	100 [10-100]	4.7k	3.1k	1.7k	1.4k
MBFC*	1179	10	400k	3	270	100	11.6k	6.9k	3.4k	2k

Table 3: Parameters used for link scheme identification in algorithm 1 and number of candidates discovered using both the PoliticalNews evaluation dataset and our MBFC* dataset as seed lists in the discovery process.

Source	Misinfo Rate
Common Crawl	0.1%
Outlinks & Backlinks	10%
Link Scheme Outlinks	29%
Link Scheme Outlinks & News Sources	49%

Table 4: Unreliable domain rates across the various domain lists as given by the reliability classifier

backlink networks. Further improvements are gained by filtering out sites that are not news according to our News Classifier.

Ideally, the final stage of unreliable domain discovery would involve the manual validation of all candidate domains. However, due to time and budget constraints, we instead sample 10% (200 domains) from the candidate list to estimate the accuracy of our discovery method. All labels were independently annotated by two researchers. Krippendorff’s alpha is computed to determine inter-annotator agreement. We find a strong agreement for reliability labels ($\alpha = 0.86$), and some disagreement for bias labels ($\alpha = 0.7$). As summarized in Table 5, we find that 47% of domains frequently feature misinformation (mixed or worse), 37% predominantly feature misinformation articles (low or very low), 46% are frequently biased (not center), and 38% predominantly feature politically polarized articles (extreme left or right). We note that 42% of discovered domains are not news sources, which greatly impacts our results. However, given that a discovered website is a news source, it is highly likely to be unreliable (83%), peddle known misinformation (65%) or have extreme right political bias (65%). The distribution of reliability and political bias labels for manually labeled news sources is given in Table 5.

Additionally, we profile the distribution of countries in our domain lists according to their IP addresses. For the MBFC* dataset, we find that 90% of domain IPs are US-based. In comparison, we find that 80% of the newly discovered domain IPs are US-based, which is confirmed by our evaluation (Table 5). This reiterates our concern that misinformation source discovery systems are heavily biased based on their seed list of domains.

We further find evidence of evasion and exit behavior in the operation of unreliable news domains. We observe domain switching through redirects and context switching through content changes. From our sample of candidate misinformation sources, 8% are dead URLs, 4% redirect to newer domains, and 2% were once misinformation sources but have since ‘context switched’ so that the domain is no

Reliability		Bias		Locale	
Label	#	Label	#	Country	#
V. Low	63	Ext. Left	7	US	80%
Low	10	Left	9	EU	15%
Mixed	18	Center	14	Middle East	2%
High	14	Right	6	South Asia	1%
V. High	6	Ext. Right	67	Misc.	2%

Table 5: Evaluation of the discovered domains sample

longer a news site⁹. Together, 15% of domains in our sample were previously news sites. In our annotations, we do not classify these domains as misinformation sources.

Discussion

Survival Rates Given the ease of evasion and exit tactics like domain swapping, pivoting from news to satire, or shutting down altogether, more dynamic approaches for detecting and discovering misinformation sources are needed. Once unreliable news sites are exposed, they can be boycotted or penalized by search engines or social media companies. We find that 3 of 4 unmaintained domain lists published between 2019 and 2017 have domain survivability rates of 35-44%.

Webgraph GNNs The outlink network outperforms both the backlink and combined networks across the majority of our experiments (see Figure 3). Intuitively we had expected backlinks to perform better, as they contain the link scheming behaviour that we use identify for site discovery. However, the better performance of outlinks can be understood in two ways; firstly, the information gained from the outlink network is denser due to there being fewer outlink sites than backlinks in general. If we were to compare results on a GNN trained on all backlinks vs. all outlinks (rather than 10 of each), we would likely get very different results. Unfortunately, such an experiment is not feasible given data limitations. Secondly, websites have more control over their outlinks than their backlinks, which may better provide a better proxy for site intent.

We find that performance on the absolute bias task, as measured by F1 score, is higher than for the relative bias task (Figure 3). While this is in line with our hypothesis that link structure is more predictive of absolute bias, we note that it may also be due to the apparent skew in the unreliable labels towards the extreme right (Figure 1). This imbalance likely impacts performance on the relative bias task.

⁹We used archive.org to verify domains had context-switched.

Unsurprisingly, we find diminishing returns when adding more backlinking domains to the backlink network. As shown in Figure 4, F1 score begins to increase more slowly as we add successive backlinks to the network. This is further supported by the design space exploration of the discovery process (figure 7), where increasing the number of outlinks used per link scheme improves recall, to a point.

Discovery The core motivation for the development of the discovery process is that lists of unreliable domains will become quickly outdated. However the discovery process is still heavily biased towards the initial set of seed domains; Table 5 reveals that the known biases of locale and political orientation from the MBFC dataset (Figure 1) are reflected in the set of discovered domains; although slightly improved, the discovered domains are still predominantly US-based and skew extreme-right. We also note the potential political bias in using the partial F1 evaluation metric as it relies on the MBFC list.

We find that a small change in location distributions occurs during link scheme identification. This could allow researchers investigating unreliable news domains in a given region to set a starting locale distribution using both an initial seed list as well as by appropriately filtering the link scheme sites computed by Algorithm 1. While the danger of misuse is clear, we posit that our proposed methodology is extendable to other novel country and language contexts to better identify sources of misinformation. We further stress the language-, content-, and social media-agnostic aspects of the system in this regard.

We note that using archive.org in combination with our SEO classifiers is an effective way to investigate site history. As seen in our evaluation, 7% of our sample is composed of non-news domains with scopes that have changed over time. However, backlink profiles are ‘sticky’ and in many cases live longer than the actual content of the domain. By analysing mismatches between a domains link profile and its current content, we can uncover evasion and exit strategies.

Limitations and Future Work

We observe that blocklists of unreliable news domains have poor survival rates, with more than 50% of such domains being parked or returning 404s to get requests within three years of publication. We employ a backlink heuristic to filter out 87% of parked domains and 63% of 404 domains that have less than 10k backlinks. This heuristic also filters out >500 domains that are still live, albeit with a minimal backlink profile. While the rationale here is to avoid training a model to predict dead domains as unreliable, it is still a limitation of the proposed method. As a result, domains should have a substantial history and a well-developed backlink profile for there to be confidence in the predictions of this detection method. Additionally, as with competing detection and discovery methods, a seed list of labeled domains is required.

Data collection through proprietary APIs can be costly. This cost constrains the validation of our discovery algorithm, which is why we use flat classifiers for the discovery process. However, we have shown that the GNN method is

superior for reliability and bias prediction and so we expect that the accuracy of the discovery method would be higher using the GNNs for inference. We note that these methods are not ‘data-hungry’ and achieve impressive performance with limited backlinks and outlinks (figure 4, 7).

In our evaluation of the unreliable domain discovery process, we find that the lines between extreme political bias and reliability become blurred. While our proposed method is successful in creating a new sample of domains with a high prevalence of unreliable sites among them, the final evaluation is not a simple process. We encounter a broad range of domains in the link network, with news sites from all over the world appearing in the sample. In many countries and contexts, the spectrum of political bias may not exist on a clear left-right scale, so it is natural that inter-annotator agreement for bias labels is lower. We also encountered numerous satire sites which can be difficult to classify. Our aim is not to ignore the subjectivity of reliability and we acknowledge that the evaluation of candidate unreliable domains requires the development of thorough labeling protocols.

Finally, despite much of the related work taking advantage of social networks to determine news source reliability and to discover new unreliable domains, we exclusively look at webgraph and SEO attributes. In future work, we plan to combine our current webgraph approach with social network data and website content.

Conclusion

Progress in the misinformation space is typically realized by the publication of lists of domains that curate unreliable content and produce misinformation. However, due to the transient nature of many of these sites, we demonstrate that such lists do not stay relevant for long. With this in mind, we present a novel dataset crafted from a combination of webgraph data and SEO attributes. We provide strong GNN baselines for predicting news reliability and political bias labels on our MBFC* dataset, with F1 scores of 0.84 & 0.81 respectively. We achieve SoTA results with an F1 score of 0.96 on the PoliticalNews dataset.

Additionally, we propose an algorithm for discovering new unreliable news domains. By leveraging signals of blackhat link-building methods and the natural homophily within webgraphs, we introduce a graph-based methodology to generate sets of candidate news sources. Applying the flat classifiers trained on our SEO dataset to these candidates, we effectively filter out false positives. Based on partial F1 scores, our content and social media agnostic algorithm is en-par with Chen and Freire (2020). Evaluating this algorithm on the MBFC* dataset, 47% of discovered domains are unreliable and 46% are biased. We provide annotations for a sample of these newly discovered unreliable domains.

We demonstrate that webgraphs provide a strong signal for numerous tasks in the misinformation ecosystem. This methodology can allow online misinformation research in new country-contexts where social media usage may be low and English may not be the predominant language.

Ethical Statement

We acknowledge a growing distrust in misinformation detection systems as many feel that their advancement coincides with the erosion of free speech. While our unreliable domain discovery process is a crucial line of research to enable informed discussion about media reliability, it could also lead to misuse if applied to reliable or opposition domains by government censors.

A limitation of unreliable news lists is that they tend to contain more extreme-right domains than extreme-left domains (figure 1). While we do not know the true underlying left-right distributions of unreliable new sources, discovery algorithms trained on these imbalanced lists may bias the discovery algorithm disproportionately towards unreliable conservative domains. This is problematic as psychological research has found that both extreme-right and extreme-left individuals are susceptible to misinformation (Harper and Baguley 2019).

To validate the partisan concern, we show that a classifier can be trained to distinguish left-wing domains from right-wing domains. Consequently, discovery algorithms biased towards one side of the political spectrum may miss harmful domains that target the other. As a result, we propose the use of an ‘absolute’ bias classifier for use in the discovery process, which is trained to identify politically extreme domains targeting both left and right.

Acknowledgements

We thank Isabel Murdock for developing the MBFC scraper. This work was supported in part by the Office of Naval Research grant (N000142112229) and the US Army grant (W911NF20D0002). Additional support was provided by the Center for Computational Analysis of Social and Organizational Systems (CASOS) at Carnegie Mellon University. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Knight Foundation, Office of Naval Research, or the U.S. Government.

References

- Ahrefs. 2022. SEO tools & resources to grow your traffic.
- Aires, V. P.; G. Nakamura, F.; and F. Nakamura, E. 2019. A link-based approach to detect media bias in news websites. In *Companion Proceedings of The 2019 World Wide Web Conference*, 742–745.
- Allcott, H.; and Gentzkow, M. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2): 211–36.
- Aswani, R.; Ghrera, S.; Chandra, S.; and Kar, A. 2021. A hybrid evolutionary approach for identifying spam websites for search engine marketing. *Evolutionary Intelligence*, 14(4): 1803–1815.
- Aswani, R.; Kar, A. K.; Ilavarasan, P. V.; and Dwivedi, Y. K. 2018. Search engine marketing is not all gold: Insights from Twitter and SEO Clerks. *International Journal of Information Management*, 38(1): 107–116.
- Binder, M. 2019. Fake news sites are simply changing their domain name to get around Facebook fact-checkers.
- Bradshaw, S. 2019. Disinformation optimised: gaming search engine algorithms to amplify junk news. *Internet policy review*, 8(4): 1–24.
- Bradshaw, S.; DiResta, R.; and Miller, C. 2022. Playing Both Sides: Russian State-Backed Media Coverage of the# BlackLivesMatter Movement. *The International Journal of Press/Politics*, 19401612221082052.
- Castelo, S.; Almeida, T.; Elghafari, A.; Santos, A.; Pham, K.; Nakamura, E.; and Freire, J. 2019. A topic-agnostic approach for identifying fake news pages. In *Companion proceedings of the 2019 World Wide Web conference*, 975–980.
- Chen, Z.; and Freire, J. 2020. Proactive discovery of fake news domains from real-time social media feeds. In *Companion Proceedings of the Web Conference 2020*, 584–592.
- Chitika. 2013. The value of Google result positioning. *Westborough: Chitika Insights*.
- Epstein, R.; and Robertson, R. E. 2015. The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences*, 112(33): E4512–E4521. Publisher: Proceedings of the National Academy of Sciences.
- 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.
- Hamilton, W.; Ying, Z.; and Leskovec, J. 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30.
- Harper, C. A.; and Baguley, T. 2019. “You are fake news”: Ideological (A) symmetries in perceptions of media legitimacy.
- Hrckova, A.; Moro, R.; Srba, I.; and Bielikova, M. 2021. Quantitative and qualitative analysis of linking patterns of mainstream and partisan online news media in Central Europe. *Online Information Review*.
- Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.
- Lee, S.; Afroz, S.; Park, H.; Wang, Z. J.; Shaikh, O.; Sehgal, V.; Peshin, A.; and Chau, D. H. 2022. Explaining Website Reliability by Visualizing Hyperlink Connectivity. In *2022 IEEE Visualization and Visual Analytics (VIS)*, 26–30. IEEE.
- Lewandowski, D.; Sünklar, S.; and Yagci, N. 2021. The influence of search engine optimization on Google’s results:: A multi-dimensional approach for detecting SEO. 12–20. ISBN 978-1-4503-8330-1.
- Lin, H.; Lasser, J.; Lewandowsky, S.; Cole, R.; Gully, A.; Rand, D.; and Pennycook, G. 2022. High level of agreement across different news domain quality ratings.
- Malaga, R. A. 2010. Chapter 1 - Search Engine Optimization—Black and White Hat Approaches. In *Advances in Computers*, volume 78 of *Advances in Computers: Improving the Web*, 1–39. Elsevier.

- McCue, T. 2018. SEO Industry Approaching \$80 Billion But All You Want Is More Web Traffic.
- Okereke, M.; Ukor, N. A.; Ngaruiya, L. M.; Mwansa, C.; Al-haj, S. M.; Ogunkola, I. O.; Jaber, H. M.; Isa, M. A.; Ekpenyong, A.; and Lucero-Prisno III, D. E. 2021. COVID-19 misinformation and infodemic in rural Africa. *The American Journal of Tropical Medicine and Hygiene*, 104(2): 453.
- Petrou, N.; Christodoulou, C.; Anastasiou, A.; Pallis, G.; and Dikaiakos, M. D. 2023. A Multiple change-point detection framework on linguistic characteristics of real versus fake news articles. *Scientific Reports*, 13(1): 6086.
- Sehgal, V.; Peshin, A.; Afroz, S.; and Farid, H. 2021. Mutual Hyperlinking Among Misinformation Peddlers. ArXiv:2104.11694 [cs].
- Silva, A.; Luo, L.; Karunasekera, S.; and Leckie, C. 2021. Embracing domain differences in fake news: Cross-domain fake news detection using multi-modal data. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 557–565.
- Szurdi, J.; Kocso, B.; Cseh, G.; Spring, J.; Felegyhazi, M.; and Kanich, C. 2014. The Long {"Taile"} of Typosquatting Domain Names. In *23rd USENIX Security Symposium (USENIX Security 14)*, 191–206.
- Urman, A.; Makhortykh, M.; Ulloa, R.; and Kulshrestha, J. 2022. Where the earth is flat and 9/11 is an inside job: A comparative algorithm audit of conspiratorial information in web search results. *Telematics and informatics*, 72: 101860.
- Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; and Bengio, Y. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.
- Vissers, T.; Joosen, W.; and Nikiforakis, N. 2015. Parking Sensors: Analyzing and Detecting Parked Domains. In *Proceedings of the ISOC Network and Distributed System Security Symposium (NDSS'15)*.
- Weinberg, Z.; Sharif, M.; Szurdi, J.; and Christin, N. 2017. Topics of Controversy: An Empirical Analysis of Web Censorship Lists. *Proc. Priv. Enhancing Technol.*, 2017(1): 42–61.
- Williams, E. M.; and Carley, K. M. 2023. Search engine manipulation to spread pro-Kremlin propaganda. *Harvard Kennedy School Misinformation Review*.
- Zandt, D. V. 2022. Media Bias Fact Check: A Comprehensive Media Bias Resource.
- Zhang, S.; and Cabage, N. 2017. Search Engine Optimization: Comparison of Link Building and Social Sharing. *Journal of Computer Information Systems*, 57(2): 148–159.
- Zweig, K. A. 2017. Watching the watchers: Epstein and Robertson's „Search Engine Manipulation Effect“.
2. Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
 3. Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes
 4. Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes
 5. Did you describe the limitations of your work? Yes
 6. Did you discuss any potential negative societal impacts of your work? Yes
 7. Did you discuss any potential misuse of your work? Yes
 8. Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes
 9. Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
- Additionally, if your study involves hypotheses testing...
- Did you clearly state the assumptions underlying all theoretical results? Yes
 - Have you provided justifications for all theoretical results? Yes
 - Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Yes
 - Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Yes
 - Did you address potential biases or limitations in your theoretical framework? Yes
 - Have you related your theoretical results to the existing literature in social science? Yes
 - Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes
- Additionally, if you are including theoretical proofs...
- Did you state the full set of assumptions of all theoretical results? N/A
 - Did you include complete proofs of all theoretical results? N/A
- Additionally, if you ran machine learning experiments...
- Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes
 - Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes
 - Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? We report 5-fold cross-validation results.
 - Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? No - the resources required are negligible.

Paper Checklist

- Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes

5. Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes
6. Do you discuss what is “the cost“ of misclassification and fault (in)tolerance? Yes

Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

1. If your work uses existing assets, did you cite the creators? Yes
2. Did you mention the license of the assets? No
3. Did you include any new assets in the supplemental material or as a URL? Yes
4. Did you discuss whether and how consent was obtained from people whose data you’re using/curating? N/A, no personal data
5. Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? N/A, no personally identifiable information or offensive content
6. If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? Yes, see <https://doi.org/10.1184/R1/25174193.v1>
7. If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? Yes, see <https://doi.org/10.1184/R1/25174193.v1>

Additionally, if you used crowdsourcing or conducted research with human subjects... N/A

Parked Domain Classifiers

We evaluate two methods for classifying parked domains; a classifier trained on HTML & HTTP response features (Vissers, Joosen, and Nikiforakis 2015) and a set of regular expressions (Regexp) that match known HTML patterns for parked domains (Szurdi et al. 2014). The regular expression method does not require training. For the classifier, we construct positive labels by scraping 3k domains from sedo.com, a popular parking domain registrar. We construct negative labels by extracting random sample of 3k domains from Common Crawl.

Contrary to previous findings that a set of regular expression designed to catch parked domain HTML templates performs better than the trained classifier (Weinberg et al. 2017), we find that the regular expression method has low recall over our evaluation set of 3000 parked domains from sedo.com. Regexp obtains just 26% recall on this list, with the trained classifier obtaining over 90% recall. Performance between the two methods is compared in terms of accuracy and F1 score in Table 6.

We further find that Regexp have much lower recall over our news lists, with Regexp finding only 38% of the parked domains that the trained classifier found. The status of these domains, once identified by the classifiers, was manually verified. We attribute the gap in recall to the regular expressions being outdated.

Method	Acc	F1	# in News List	Precision
Regexp	100%	0.43	98	100%
Classifier	95%	0.92	253	96%

Table 6: Evaluation of Parked Domain Classifiers

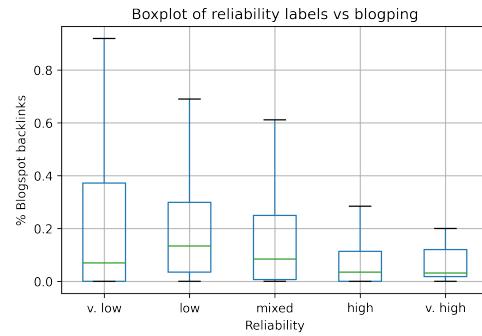


Figure 8: Percentage of backlinks from blogspot.com sites

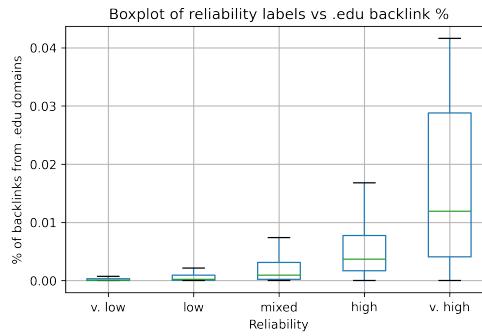


Figure 9: Percentage of backlinks from .edu domains

Statistic	Backlinks	Outlinks	Combined
Nodes	14959	9827	20817
Edges	32110	30397	61917
Avg Degree	4.29	6.18	5.95
Transitivity	0.016	0.040	0.027
CPL	0.401	0.140	0.334
Density	1.4e-4	3.2e-4	1.4e-4
Assortativity	-0.104	-0.033	0.100

Table 7: Summary Statistics for Webgraphs