NETWORK ANALYSIS OF COVID-19 NEWS SOURCES

Felix Parker
Department of Civil and Systems Engineering
Johns Hopkins University
fparker9@jhu.edu

Tianmang Chen
Department of Civil and Systems Engineering
Johns Hopkins University
tchen95@jhu.edu

December 7, 2020

ABSTRACT

Receiving correct and informative news related to COVID-19 is critical for people, who must make significant decisions about what activities are safe, what precautions to take, and other aspects of their lives using the information they have. Since news related to COVID-19 is so important, it is vital that we understand the news landscape and sources for this information. To this end we collected and studied a network of websites that serve as sources of news about the pandemic. We find that we can successfully extract information about source importance and other network properties; however, we did not find significant community structure in the network as it was constructed.

Keywords network analysis, news, COVID-19

1 Intro

Since COVID-19 was first identified in December 2019 it has evolved into a massive global pandemic. As of December 2020, more than 67 million cases have been confirmed, resulting in more than 1.53 million deaths [Dong et al., 2020]. Nearly 15 million cases and 300,000 deaths have been reported in the United States alone [Dong et al., 2020]. The response to the pandemic has fundamentally impacted the lives of billions more people around the world, with many being forced to work or study from home, social distance, or wear masks in public. Communication about the disease, responses to the pandemic, and plans for the future has been critical to ensuring that the public knows what measures they are required to take and have the information required to make safe and responsible decisions. With 73.6% of Americans reporting that they get information on COVID from social media and 68.4% reporting that they got information on COVID from a non-governmental website [Ali et al., 2020] it is important to understand online sources for this information.

In this work we aim to further the understanding of news sources related to COVID-19 on the internet using standard tools from network analysis. The structure of the sub-graph of the internet induced by limiting the nodes to potential news sources for COVID-19 encodes information about which sources are influential, which sources are more closely related to others, and how information flows between sources to readers. In order to extract this information we approximate the network by sampling articles known to contain references to COVID-19 and following all links from these articles. Given this network we employ centrality measures, clustering algorithms, and other tools to glean insights from it.

Our primary contributions in this work are as follows:

1. We collect a large (n=63,514) dataset of news articles related to COVID-19 with links to other articles and sources.
2. We collect a large dataset of Tweets related to COVID-19 that contain links to news sources.
3. We construct a series of networks using articles, direct hyperlinks, and article content.
4. We employ standard methods from network analysis to study the network of news sources pertaining to COVID-19.

2 Related Work

In this section we highlight some of the existing literature that is related to our work. In particular, we report on work that studies COVID-19 information in social networks, online news, and other sources.

2.1 COVID-19 in Social Media

There has been extensive research using social media data to study the COVID-19 pandemic. There have been numerous papers on collecting posts and other user behavior from popular social media websites including Twitter, Facebook, and Instagram. One such work is Banda et al. [2020], which collects over 800,000,000 tweets related to COVID-19 for use in other research. Hung et al. [2020] collect a large dataset of COVID-related tweets as well and perform sentiment analysis to better understand discussion of COVID-19 online.

2.2 COVID-19 in Online News

In addition to studying social media, other papers have studied COVID information in news sources published online. Much of this work, like studies of online news from before the pandemic, has focused on misinformation. The CoAID dataset [Cui and Lee, 2020] collects news articles on COVID-19 and associated interactions, and classifies each article based on whether it contains misinformation. The ReCOVery dataset [Zhou et al., 2020] manually identifies news sources for COVID-19 and assigns them a credibility score. They then collect articles from these sources and related tweets for further research.

Another topic in this area is information about COVID-19 in academic papers published online. The CORD-19 dataset [Wang et al., 2020a] collects over 140,000 such papers for automated information retrieval and has been used for a variety of tasks.

Finally, Ali et al. [2020] has also studied which sources people report getting their news on COVID-19 from and how much they trust these sources.

2.3 Network Analysis of News Sources

There is some existing literature on using network analysis to study online news. Shu et al. [2018] constructs a network of news sources and interactions with articles on social media over time to detect misinformation propagation. Weber and Monge [2011] does a detailed analysis of news sources to study how information flows through the network from news sources to news authorities to news hubs.

3 Methods

In this section we describe how we compiled a dataset of news articles with the relevant metadata for our analysis, how we constructed networks from this dataset, and finally how we analyzed those networks.

3.1 Data Collection

A large database of articles with the relevant attributes for each article was crucial to being able to construct and study the networks that are the focus of this work. The attributes that we made use of included: article URL, the URL of all links on the page, article keywords, and article content. In order to avoid manually selecting websites ourselves, we used larger datasets of COVID-19 news articles, specifically the CoAID dataset [Cui and Lee, 2020] and the Aylien COVID news dataset[1], to compile a list of articles. We then used automatic web scraping to collect the content and links of each article. Given the content of an article we then counted the occurrences of specific keywords that we defined.

In order to estimate how popular and influential an article was, we also collected Twitter data. We used an existing dataset[2] of tweets that came from Twitter's COVID-19 data stream, which filters for tweets that are likely related to COVID. We then extracted all links to websites in our articles database and all uses of our defined keywords.

[1] https://aylien.com/blog/free-coronavirus-news-dataset

[2] https://www.kaggle.com/gpreda/covid19-tweets

3.2 Network Construction

In addition to compiling the database of articles, we also constructed a series of networks using the same underlying data. The networks constructed can be divided into two categories: article-level networks and website-level networks.

We start with article-level networks, in which each node represents an individual news article or webpage. The simplest construction involved adding a directed link from each node to all other nodes in the network that the corresponding article contained a direct link to. However, this network, by nature, is very sparse because most news articles contain few direct links to other news articles. In order to make the resulting network more meaningful and increase its density, we then added links from each article to all articles with sufficiently overlapping keywords.

Due to the sparsity and size of the article-level networks, they were not the primary focus of our work. Instead, we focused on the website-level networks. In the website-level networks a node represents an entire website, which may contain many individual articles. In this network construction, edges between nodes have weights that correspond to the frequency that an article in the source node had a link to an article in the target node. As in the article-level network a link can either be a direct hyperlink or some connection involving content or keywords. For our analysis we only use direct hyperlinks in this type of network.

3.3 Network Analysis

The primary aim of this work was to analyze the network of sources for news related to COVID-19. In particular, we wanted to investigate the importance of each source and determine what communities, if any, existed in the network. To analyze the relative importance of each node we employed PageRank and degree centrality. PageRank was selected because it was designed for a very similar application – quantifying importance in a network of websites based on links between websites. Community structure in the networks was investigated using the Ravasz algorithm. We also considered other metrics, such as density and assortativity, which can be seen in Table 1.

3.4 Implementation

The methods described in this section were implemented in Python. The source code is available at: `https://github.com/flixpar/covid-news-network-analysis`. Website scraping was performed using the requests and BeautifulSoup packages. Network construction and analysis was done using the NetworkX package.

4 Results

4.1 Dataset

The dataset we collected contained 63,514 articles from 154 websites, with 9789 direct hyperlinks between articles. Not all of the 154 websites are typical news websites, but the specific webpages all contain some information about COVID-19. Each website has between 5 and 1300 associated articles. Attributes collected were: article title, date, URL, domain, links, shares on Twitter, shares on Facebook, shares on Reddit, keywords, hastags, and content. To collect shares on Twitter 179,108 tweets were processed as well, yielding 18,479 links to websites in our dataset.

4.2 Network Construction

The primary network used in our analysis is the domain-level network with weighted directed edges corresponding to the frequency that articles on one site link to articles on another site. This network is summarized in Table 1 and is visualized in Figure 1. It contains 143 nodes and 4821 links. The number of nodes present is less than the number of domains in our dataset because some domains did not have any outgoing links, and those domains were excluded from the network. This was likely due to errors in scraping links from those websites as some had measures to prevent automated content scraping.

Article-level networks were also constructed. However, we found that connecting articles just using direct links produced a graph that was very sparse. It contained 63,514 vertices and just 9789 arcs, which means that most articles were not connected to others, making the network less interesting to analyze. We were able to increase the density by adding arcs corresponding to keyword similarity or direct mentions, but ultimately we decided to focus on the domain-level network instead.

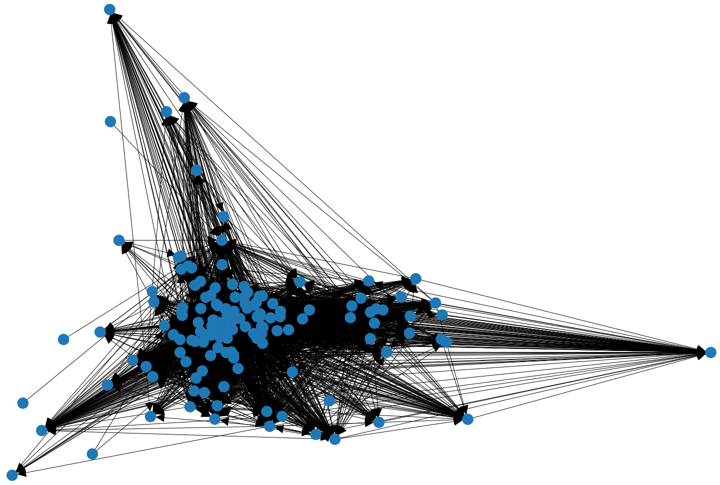

Figure 1: Website-level network visualization, with two nodes excluded to improve the layout.

4.3 Network Metrics

We analyzed our constructed networks using a variety of summary statistics, as well as per-node metrics including centrality and PageRank. We also applied community detection algorithms.

In Table 1 we see that the network has little assortativity in terms of in-degree, out-degree, shares on Facebook, and shares on Twitter. This means that there is little correspondence between which other nodes a node is connected to and its properties.

On the other hand, in Table 2 we see that there is generally a positive correlation between the degree of a node in the network and how frequently it is shared on social media. This means that if many websites link to a particular source then people on social media are more likely to link to that source as well. It suggests that there is a similarity between importance or popularity in the news source network and in social media.

Metric Name	Metric Value
Number of Nodes	143
Number of Links	4821
Density	0.237
In-Degree Assortativity	-0.0902
Out-Degree Assortativity	0.0548
Assortativity (Shares on Facebook)	-0.0091
Assortativity (Shares on Twitter)	-0.0185

Table 1: Network analysis metrics for the domain-level network.

Variable A	Variable B	corr(A,B)
Total Degree	Facebook shares	0.516
In Degree	Facebook shares	0.350
Out Degree	Facebook shares	0.461
Total Degree	Twitter shares	0.144
In Degree	Twitter shares	0.341
Out Degree	Twitter shares	-0.080

Table 2: Pearson correlation metrics for the domain-level network.

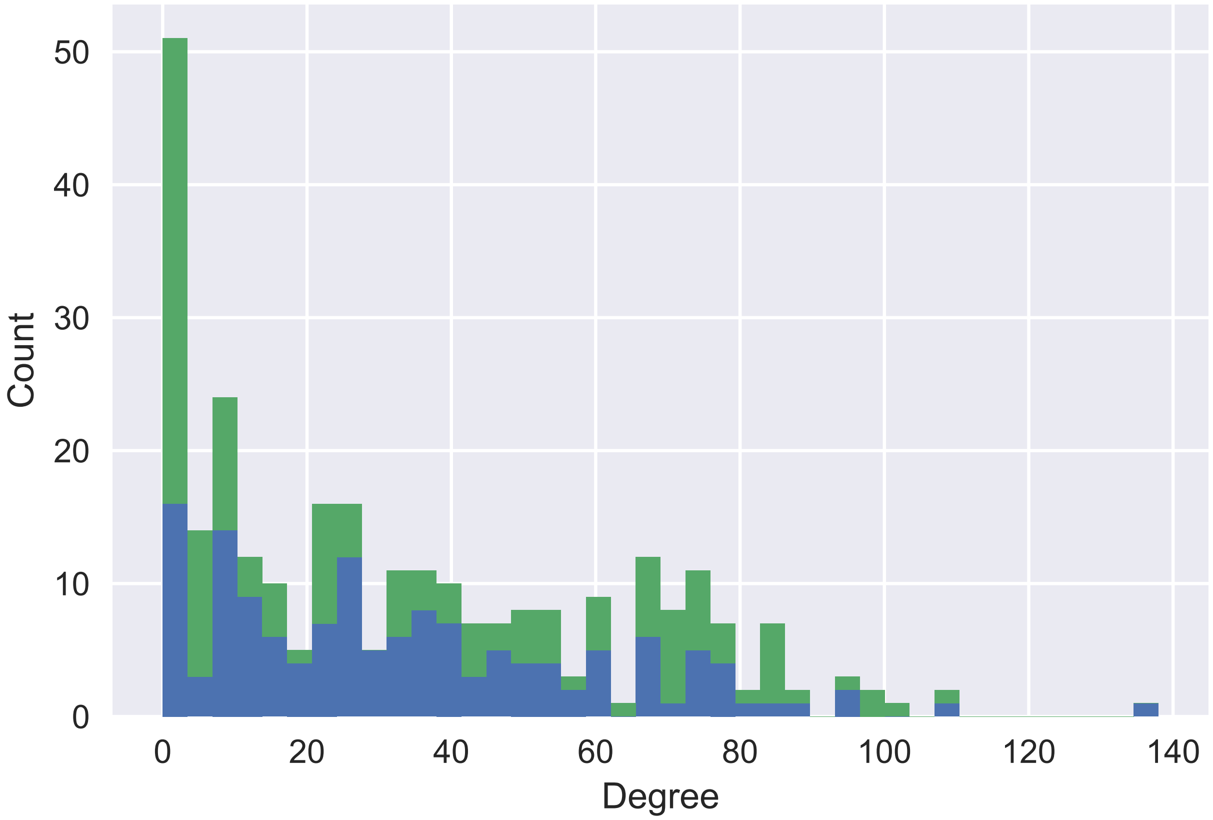

Figure 2: In-degree distribution (blue) and out-degree distribution (green) of the domain-level network.

In Figure 2, we can see that many nodes have out-degree of 5 or less, which indicates that pages on these websites have few links to others. The distribution above an out-degrees of 5 and below 100 looks relatively uniform, while almost no nodes have out-degree more than 100. As for in-degree, the frequency seems to decrease as the in-degree increases. Since the node count is not that high, the degree distribution is noisy, and therefore it is unclear whether it is following a power-law distribution that is indicative of a small-world network.

Figure 2 shows both the in-degree (blue) and out-degree (green) centrality. Higher in-degree centrality generally corresponds with higher node importance in the network because it means that many other websites link to the given website. By this metric, the most important nodes in the network, in descending order are: Twitter, Google News, the Centers for Disease Control (CDC), Apple press releases, the New York Times, the World Health Organization (WHO), CNN, and Reuters. This list largely agrees with our expectations about which sites would be important sources for news about COVID. An exception to this is that it was surprising that Apple press releases appeared so high on the list, but there were few of them related to COVID and they are likely to be cited in buisiness or economic news. Out-degree centrality (green) corresponds to linking to many other sources, so nodes with a high out-degree may be hubs for collecting news, which can also be important. Sites with a high out-degree may also cite sources for information more frequently, which makes them more reliable sources for news. Vox, Wired, Slate, Forbes, and The Atlantic were the top ranked websites by out-degree centrality.

PageRank is another metric for quantifying the importance of each node in the network. It is used in practice to assign a ranking to websites for use in search results, so it is well suited to this domain. Based on PageRank the most important sources for COVID news are Twitter, Google News, Medium, The Atlantic, Apple press releases, LinkedIn, Yahoo News, and the New York Times. Interestingly, the primary sources present in the top sites by degree centrality, namely the CDC and the WHO, are missing here even though we expect them to be among the most influential. This may be in part because those websites are often referred to by name rather than linked to directly. Otherwise, these top websites by PageRank are expected.

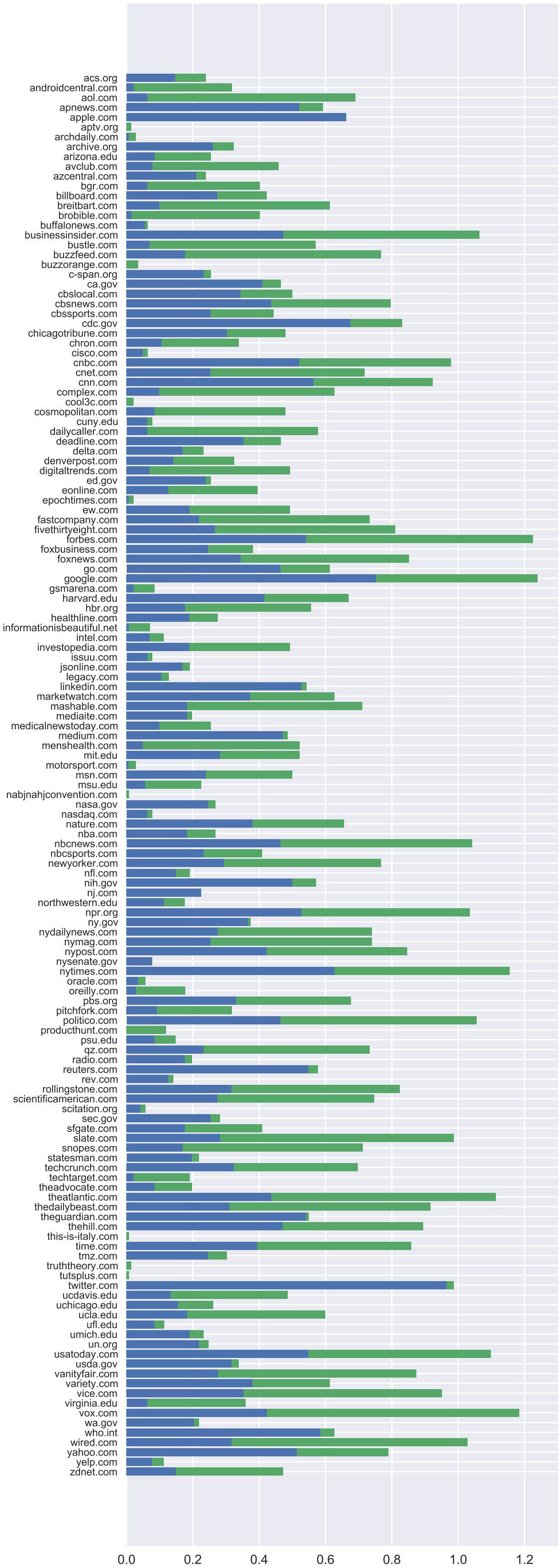

Figure 3: In-degree centrality (blue) and out-degree centrality (green) for each website included in our analysis.

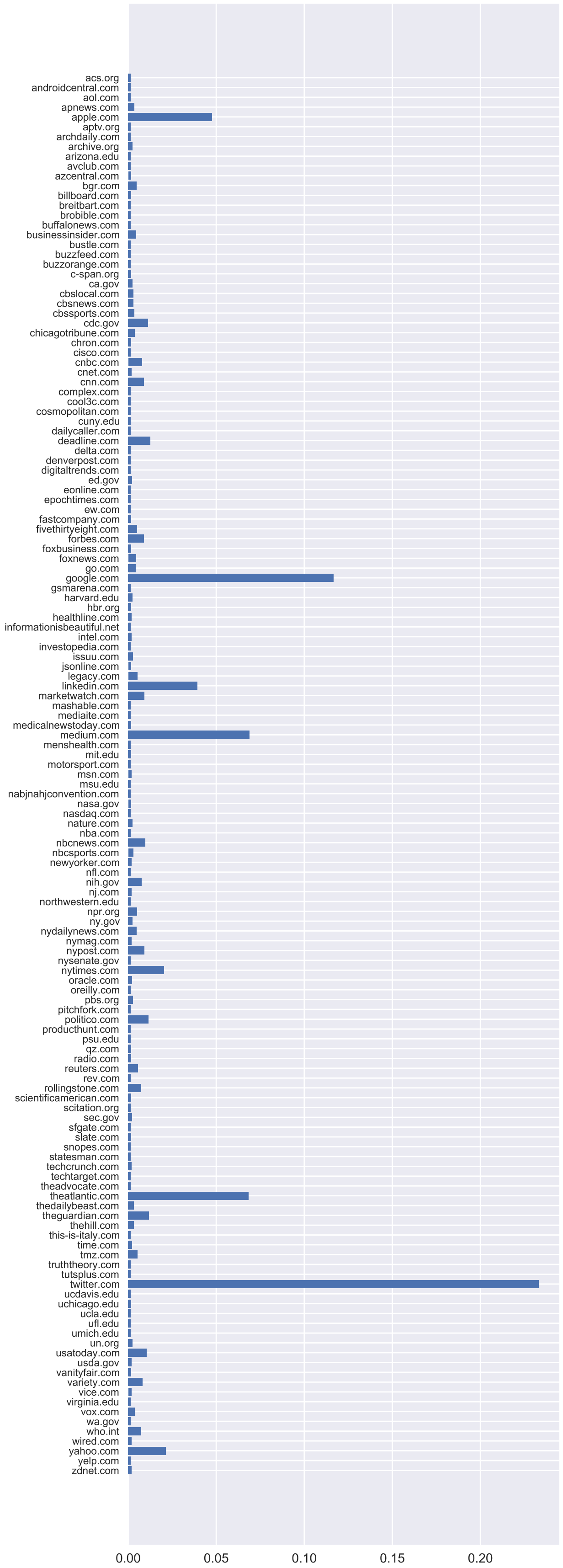

Figure 4: PageRank for each website included in our network.

4.4 Community Detection

We used the Ravasz Algorithm to perform hierarchal community detection, resulting in the dendrogram in Figure 5. The Ravasz Algorithm, as it is used in this work, iteratively merges communities that have a high average overlap (intersection over union) in neighborhoods. The results of this algorithm appear to be mixed based on our expectations of which clusters might emerge. Some clusters, including the first in the dendrogram (Vox, NPR, Politico, The Atlantic), or (acs.org, healthline.com, medicalnewstoday.com) found below, are straightforwards to explain and could be expected. However, there are also many clusters which are not expected, such as usada.gov with delta.com.

A more quantitative approach to evaluating the communities found is to measure the modularity of the communities returned at each iteration of the algorithm. In Figure 6 we plot modularity vs the number of communities, and we measure a maximum modularity of 0.00378 with 28 communities. This is quite a poor modularity score, which may indicate that there is no significant community structure in the constructed network.

This result is also suggested by the results of other community detection algorithms applied to this network. The Girvan-Newman algorithm generates a dendrogram that at each iteration simply removes a single node from the primary cluster and makes it into a singleton cluster. This means that it does not build any meaningful communities. Similarly, the greedy modularity algorithm returns a community that contains the entire network.

This lack of community structure is likely specific to the way the network was constructed from the dataset, and may be due in part to the high density of the network. In any case, community structure may still be present in the network of news sources for COVID-19, and other network constructions may make this structure clear. However, for this analysis, it appears that there is no significant community structure in the network.

The optimal community structure found by the Ravasz algorithm is given below:

```
{’archdaily.com’},
{’buffalonews.com’},
{’cool3c.com’},
{’cuny.edu’},
{’epochtimes.com’},
{’gsmarena.com’},
{’informationisbeautiful.net’},
{’intel.com’},
{’issuu.com’},
{’legacy.com’},
{’nasdaq.com’},
{’nfl.com’},
{’nysenate.gov’},
{’oreilly.com’},
{’rev.com’},
{’scitation.org’},
{’statesman.com’},
{’ufl.edu’},
{’yelp.com’},
{’motorsport.com’, ’buzzorange.com’},
{’nabjnahjconvention.com’, ’this-is-italy.com’, ’truththeory.com’,
 ’aptv.org’, ’tutsplus.com’},
{’cisco.com’, ’oracle.com’},
{’northwestern.edu’, ’uchicago.edu’, ’un.org’, ’umich.edu’},
{’acs.org’, ’healthline.com’, ’medicalnewstoday.com’},
{’denverpost.com’, ’theadvocate.com’, ’radio.com’, ’tmz.com’, ’jsonline.com’,
 ’c-span.org’, ’nj.com’, ’mediaite.com’, ’azcentral.com’},
{’producthunt.com’, ’psu.edu’},
{’msu.edu’, ’techtarget.com’, ’arizona.edu’, ’androidcentral.com’},
{’digitaltrends.com’, ’npr.org’, ’rollingstone.com’, ’thehill.com’, ’nytimes.com’,
 ’avclub.com’, ’sfgate.com’, ’cdc.gov’, ’linkedin.com’, ’vice.com’, ’thedailybeast.com’,
 ’usatoday.com’, ’cbssports.com’, ’wired.com’, ’cbslocal.com’, ’mit.edu’, ’mashable.com’,
 ’deadline.com’, ’variety.com’, ’cbsnews.com’, ’google.com’, ’foxnews.com’, ’twitter.com’,
 ’qz.com’, ’chron.com’, ’go.com’, ’apple.com’, ’billboard.com’, ’aol.com’, ’forbes.com’,
 ’yahoo.com’, ’nymag.com’, ’sec.gov’, ’slate.com’, ’fivethirtyeight.com’, ’usda.gov’,
 ’businessinsider.com’, ’nasa.gov’, ’brobible.com’, ’cosmopolitan.com’, ’delta.com’,
 ’cnn.com’, ’ew.com’, ’apnews.com’, ’bustle.com’, ’nydailynews.com’, ’ny.gov’,
 ’newyorker.com’, ’time.com’, ’msn.com’, ’dailycaller.com’, ’complex.com’, ’ca.gov’,
 ’ucdavis.edu’, ’pbs.org’, ’nbcsports.com’, ’cnbc.com’, ’nbcnews.com’, ’nature.com’,
 ’techcrunch.com’, ’pitchfork.com’, ’vanityfair.com’, ’nypost.com’, ’marketwatch.com’,
 ’buzzfeed.com’, ’politico.com’, ’ucla.edu’, ’wa.gov’, ’bgr.com’, ’archive.org’,
 ’scientificamerican.com’, ’hbr.org’, ’zdnet.com’, ’virginia.edu’, ’theatlantic.com’,
 ’nba.com’, ’cnet.com’, ’eonline.com’, ’menshealth.com’, ’ed.gov’, ’reuters.com’,
 ’snopes.com’, ’harvard.edu’, ’theguardian.com’, ’medium.com’, ’nih.gov’, ’foxbusiness.com’,
 ’breitbart.com’, ’chicagotribune.com’, ’investopedia.com’, ’who.int’, ’fastcompany.com’,
 ’vox.com’}
```

5 Discussion

Understanding how information about COVID-19 is spread online is an important task in ensuring that correct information is flowing from authorities to individuals reliably and correctly. Existing work has studied information flow on social media, collected databases of COVID news articles for analysis, and has examined where people get news on COVID and how much they trust those sources. This paper fits into that line of research and does a preliminary analysis of a large database of COVID news articles using tools from network analysis to gain insights into node influence and other network properties.

5.1 Limitations and Shortcomings

We have done an initial analysis of the network structure of COVID news sources. However, our work has a number of limitations that we are aware of along with possible future directions for further research in this area. A major shortcoming of our work at this point is the lack of further analysis of the article-level network. The version of the article-level network that just uses direct links was too sparse to be useful, and while we made progress on using article content to increase density in a meaningful way, we were not able to complete the analysis in time for the paper. Similarly, we had plans and some progress to use social media data more in our analysis.

Some additional shortcomings of our work include relying on existing datasets of article links which may be biased, difficultly in properly scraping all websites automatically to get consistent output, and focusing only on articles written in English. Future work could address these issues, or explore other paths such as using natural language processing to extract information and construct links, and analyzing the difference between real information and misinformation spread online.

Despite these issues, we think we have demonstrated how network analysis can be used to study spread of news and information related to COVID online and found interesting insights into the network structure of COVID news sources.

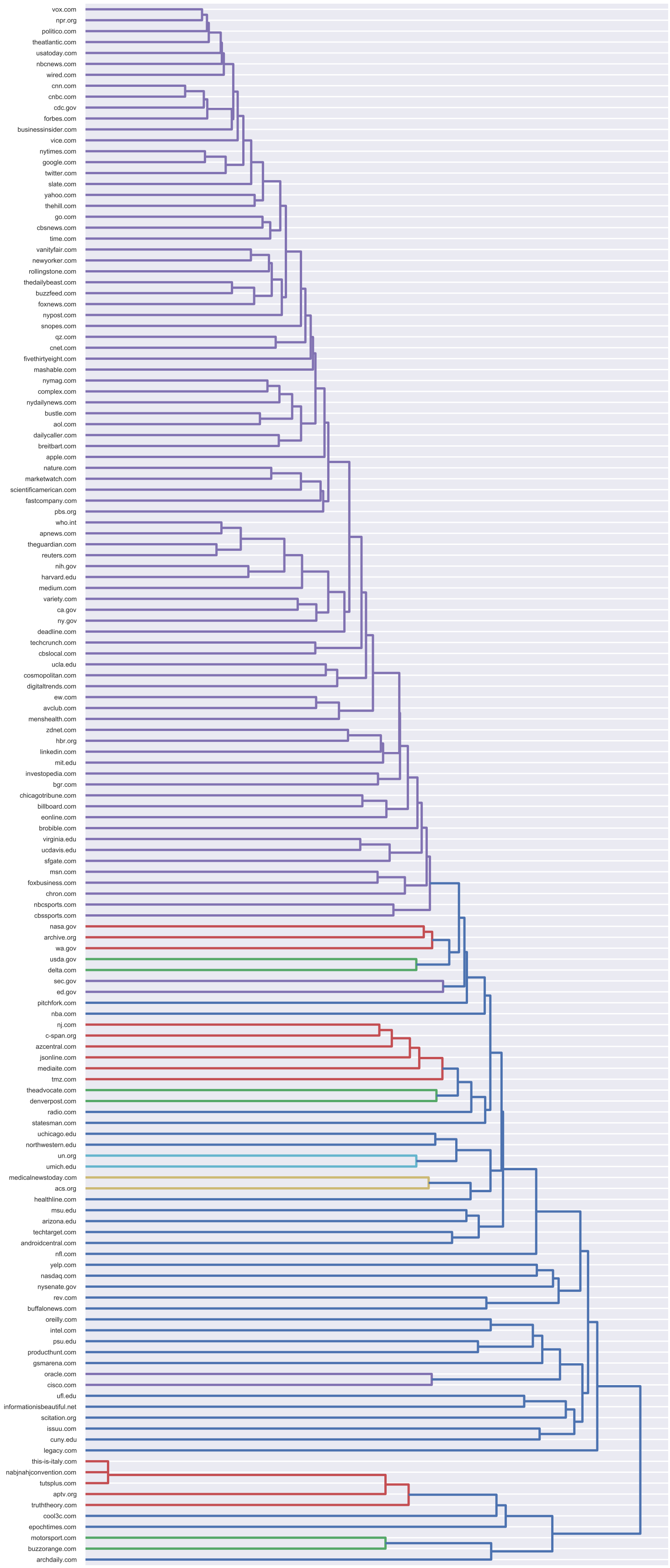

Figure 5: Denrogram generated by the Ravasz Algorithm.

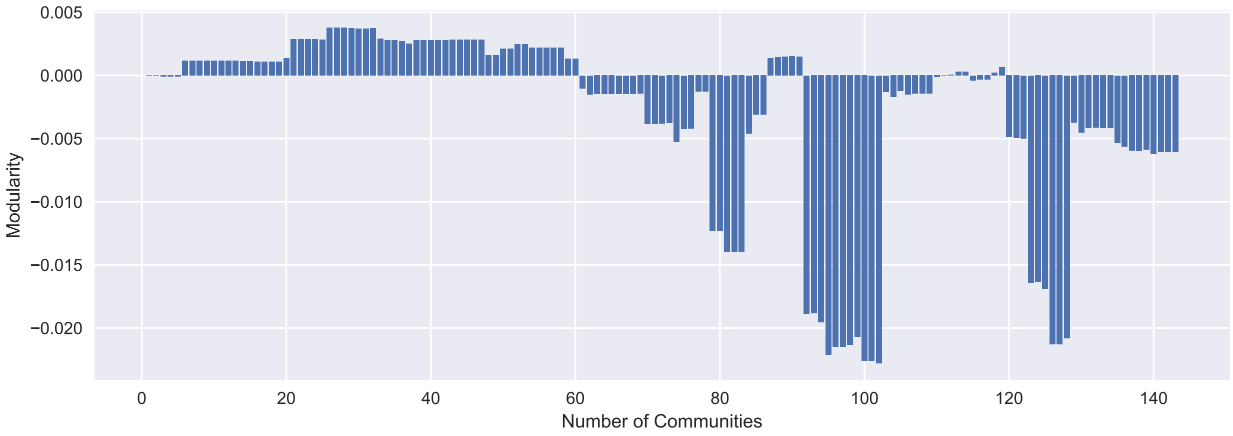

Figure 6: The modularity calculated for each iteration of the Ravasz Algorithm.

References

Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*, 20(5):533–534, May 2020. ISSN 1473-3099, 1474-4457. doi: 10.1016/S1473-3099(20)30120-1. URL https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(20)30120-1/abstract.

Shahmir H. Ali, Joshua Foreman, Yesim Tozan, Ariadna Capasso, Abbey M. Jones, and Ralph J. DiClemente. Trends and Predictors of COVID-19 Information Sources and Their Relationship With Knowledge and Beliefs Related to the Pandemic: Nationwide Cross-Sectional Study. *JMIR Public Health and Surveillance*, 6(4):e21071, 2020. doi: 10.2196/21071. URL https://publichealth.jmir.org/2020/4/e21071/.

Juan M. Banda, Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, Katya Artemova, Elena Tutubalina, and Gerardo Chowell. A large-scale COVID-19 Twitter chatter dataset for open scientific research - an international collaboration, December 2020. URL https://zenodo.org/record/4308491#.X873MMtKg-Q. type: dataset.

Man Hung, Evelyn Lauren, Eric S. Hon, Wendy C. Birmingham, Julie Xu, Sharon Su, Shirley D. Hon, Jungweon Park, Peter Dang, and Martin S. Lipsky. Social Network Analysis of COVID-19 Sentiments: Application of Artificial Intelligence. *Journal of Medical Internet Research*, 22(8):e22590, 2020. doi: 10.2196/22590. URL https://www.jmir.org/2020/8/e22590/.

Limeng Cui and Dongwon Lee. CoAID: COVID-19 Healthcare Misinformation Dataset. *arXiv:2006.00885 [cs]*, November 2020. URL http://arxiv.org/abs/2006.00885. arXiv: 2006.00885.

Xinyi Zhou, Apurva Mulay, Emilio Ferrara, and Reza Zafarani. ReCOVery: A Multimodal Repository for COVID-19 News Credibility Research. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 3205–3212, October 2020. doi: 10.1145/3340531.3412880. URL http://arxiv.org/abs/2006.05557. arXiv: 2006.05557.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex Wade, Kuansan Wang, Nancy Xin Ru Wang, Chris Wilhelm, Boya Xie, Douglas Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. CORD-19: The COVID-19 Open Research Dataset. *arXiv:2004.10706 [cs]*, July 2020a. URL http://arxiv.org/abs/2004.10706. arXiv: 2004.10706.

Kai Shu, H. Russell Bernard, and Huan Liu. Studying Fake News via Network Analysis: Detection and Mitigation. *arXiv:1804.10233 [cs]*, April 2018. URL http://arxiv.org/abs/1804.10233. arXiv: 1804.10233.

Matthew S. Weber and Peter Monge. The Flow of Digital News in a Network of Sources, Authorities, and Hubs. *Journal of Communication*, 61(6):1062–1081, 2011. ISSN 1460-2466. doi: https://doi.org/10.1111/j.1460-2466.2011.01596.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1460-2466.2011.01596.x.

Shahan Ali Memon and Kathleen M. Carley. Characterizing COVID-19 Misinformation Communities Using a Novel Twitter Dataset. *arXiv:2008.00791 [cs]*, September 2020. URL `http://arxiv.org/abs/2008.00791`. arXiv: 2008.00791.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex Wade, Kuansan Wang, Nancy Xin Ru Wang, Chris Wilhelm, Boya Xie, Douglas Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. CORD-19: The COVID-19 Open Research Dataset. *arXiv:2004.10706 [cs]*, July 2020b. URL `http://arxiv.org/abs/2004.10706`. arXiv: 2004.10706.