# Pandemic Pulse: Unraveling and Modeling Social Signals During the COVID-19 Pandemic

STEVEN J. KRIEG, JENNIFER J. SCHNUR, JERMAINE D. MARSHALL, MATTHEW M. SCHOENBAUER, and NITESH V. CHAWLA, University of Notre Dame

COVID-19 has presented society with a unique set of challenges, including seeking a scientific understanding of the novel coronavirus, modeling its epidemiology, and inferring appropriate societal response. In this article, we posit that fighting a pandemic is as much a social endeavor as a medicinal and scientific one and focus on developing a platform for understand the social pulse of the United States during the COVID-19 crisis. We collected a multitude of data that includes longitudinal trends of news topics, social distancing behaviors, community mobility changes, web searches, and other descriptors of the COVID-19 pandemic's effects on the United States. Our preliminary results show that the number of COVID-19-related news articles published immediately after the World Health Organization declared the pandemic on March 11 have steadily decreased—regardless of changes in the number of cases or public policies. Additionally, we found that politically moderate and scientifically grounded sources have, relative to baselines measured before the beginning of the pandemic, published a lower proportion of COVID-19 news articles than more politically extreme sources—a fact that has implications for the spread and consequences of misinformation during the pandemic. We suggest that further analysis of these multi-modal signals could produce meaningful social insights and present an interactive dashboard to aid further exploration.[1]

19

CCS Concepts: • **Computing methodologies** → *Model development and analysis*;

Additional Key Words and Phrases: Covid-19, coronavirus, sars-cov-2, social signals, news data, political bias, social distancing, mobility trends

**ACM Reference format:**
Steven J. Krieg, Jennifer J. Schnur, Jermaine D. Marshall, Matthew M. Schoenbauer, and Nitesh V. Chawla. 2020. Pandemic Pulse: Unraveling and Modeling Social Signals During the COVID-19 Pandemic. *Digit. Gov.: Res. Pract.* 2, 2, Article 19 (December 2020), 9 pages.
https://doi.org/10.1145/3431805

## 1 INTRODUCTION

The COVID-19 pandemic has disrupted the rhythm of global society in unprecedented ways and at an unparalleled scale. A concerted response to the challenge presented by this pandemic requires not only rapid scientific and medicinal advances but also social adaptations in behavior and norms to control the spread of the epidemic.

[1] http://cnds.nd.edu/pandemicpulse.

Authors' address: S. J. Krieg, J. J. Schnur, J. D. Marshall, M. M. Schoenbauer, and N. V. Chawla, University of Notre Dame, Center for Network and Data Science, Notre Dame, Indiana 46556; emails: {skrieg, jschnur, jmarsha5, mschoenb, nchawla}@nd.edu.


2639-0175/2020/12-ART19
https://doi.org/10.1145/3431805

Societies are at the front lines of the fight with the pandemic, and we ask whether the narratives forming the societal response are consistent and scientifically informed. A potent response to the pandemic, for example, requires an alignment of actors and institutions in a society to follow procedures and policies that might collectively slow down the process. How disconnected are the media narratives that form our opinions? How disconnected are our actions whether in mask wearing or social distancing? It is important to be able to answer questions such as these to gauge the pulse of a society and develop appropriate responses.

In this work, we present a collection of social signals that are representative part of the social pulse of the United States. These signals include COVID-19 case data, demographic data, longitudinal news and web search trends, media bias data, and mobility reports. As a doctor studies a patient's vitals to aid in identifying a diagnosis and prescribing treatment, we aim to unravel a society's vitals and model these vitals to inform our understanding of broad effects of the COVID-19 pandemic on the spread of information, social behaviors, and more. To aid in further exploration, we published an interactive dashboard alongside this article.[1]

The rest of the article proceeds as follows: In Section 2, we describe data collection and preprocessing; in Section 3, we present the results of preliminary analysis of news signals in conversation with misinformation and its consequences; and in Section 4, we discuss opportunities for future work.

## 2 DATA

We collected COVID-19 case data from Johns Hopkins Univerisity [13], news data from the Global Database of Events, Language, and Tone (GDELT) [16], web search data from Google trends, media bias labels from Media Bias/Fact Check [8] and AllSides [5], social distancing data from Unacast [20], and demographic data from the Center for Disease Control and Prevention [1–4, 9, 11]. In the following sections, we detail our methods for data collection and analysis.

### 2.1 COVID-19 Case Data

Johns Hopkins University (JHU) has created a repository for COVID-19 case data that combines information from the World Health Organization (WHO) and a number of other global and national sources [13]. We use this data from JHU to report the number of new cases and new deaths by location and date.

### 2.2 United States Demographics

To represent demographic information as well as risk factors based on individual states, we collected data from various sources, including the Center for Disease Control, United States Census Bureau, and the Bureau of Labor Statistics. These data enable us to explore correlations between demographic information for locations and other data, such as searching for relationships between locations with higher rates of COVID-19 deaths. The demographic data include heart disease hospitalization rate, cancer rate, population age, hypertension and stroke rates, obesity, walk scores, eating habits (i.e., veggie intake), ethnicity, and smoking habits. After collecting all variables for each state, we performed typical data preprocessing and cleaning steps: noise removal, aggregation, and conversion to percentages.

### 2.3 News Data

*2.3.1 COVID-19 Articles.* The GDELT monitors worldwide print, broadcast, and online news in over 100 languages [16]. For each article published, GDELT adds to its Global Knowledge Graph (GKG) a record that contains a variety of metadata, including geographical references, textual themes, and sentiment scores.[2] The GKG processes several terabytes of data every year, making it a rich source of longitudinal news data. We created a corpus of COVID-19 news by extracting from the GKG any record that met at least one of the criteria listed in Table 1.

[2]While the GKG monitors other news formats, the vast majority of its COVID-19-related records represent textual pieces. We therefore use the term "article" to refer generally to any record stored in the GKG.

Table 1. Criteria Used to Determine Whether an Article from the GKG Should Be Included in the COVID-19 Corpus

| Description | GKG Column(s) | Possible Values |
|---|---|---|
| Article title or URL contains | Title<br>DocumentIdentifier | coronavirus<br>covid<br>2019-ncov<br>ncov-2019<br>ncov2019<br>sars-cov-2 |
| Article text includes a reference to the COVID-19 virus, COVID-19 cases, pandemic, or a related term | Themes | WB_2167_PANDEMICS<br>HEALTH_PANDEMIC<br>*_CORONAVIRUS<br>*_CORONAVIRUSES<br>*_CORONAVIRUS_INFECTIONS |

The * character represents the prefix “TAX_DISEASE”.

Table 2. List of Possible Ratings Assigned to News Sources by Media Bias/Fact Check and AllSides

| Possible Bias Ratings | |
|---|---|
| **Media Bias/Fact Check** | **AllSides** |
| Left | Left |
| Left-Center | Left-Center |
| Least Biased | Center |
| Right-Center | Right-center |
| Right | Right |
| Scientific | Mixed |
| Questionable Sources | |
| Conspiracy-Pseudoscience | |

We also removed duplicate articles, which we defined as those with a non-unique combination of publisher and title.

*2.3.2 Media Bias Data.* We used two independent sources for labeling the political bias of news sources: Media Bias/Fact Check (MBFC) and AllSides. MBFC is an independent online media outlet that evaluates news sources on their political bias and the factuality of their publications [8]. AllSides [5] takes a similar task but incorporates surveys, reviews, and additional data into their evaluation process. Both have been utilized in recent works on media bias detection [14, 17, 22]. Table 2 lists the possible ratings given by each organizaztion.

We utilize MBFC as our primary source and AllSides as supplementary. We prefer MBFC for the following reasons:

(1) MBFC’s evaluation methodology is explained in more detail and thus more transparent.
(2) MBFC includes a “Scientific” category, which we found to be a helpful addition. Most of MBFC’s Scientific sources were labeled “Least Biased” by AllSides.
(3) MBFC includes a “Questionable Sources” category. While this is comprised largely of extreme right sources, it also contains many extreme left sources. We found it helpful to separate these extremes from regular right and left-leaning sources.

Table 3. The 10 Keywords that Appear Most Frequently in the Titles of COVID-19-related News Articles

| Keyword | News Mentions |
| --- | --- |
| coronavirus | 2,575,769 |
| covid-19 | 1,765,810 |
| news | 744,754 |
| cases | 705,753 |
| virus | 562,247 |
| pandemic | 474,974 |
| death | 405,571 |
| trump | 379,232 |
| lockdown | 293,885 |
| china | 279,150 |

(4) MBFC rates the factuality of each source, which we found to be useful in studying the implications of COVID-19 news trends (Section 3).

## 2.4 Social Distancing and Mobility Data

*2.4.1 Unacast Social Distancing Data.* Unacast provides social distancing scores for U.S. states and counties based on cell phone GPS data [20]. From this dataset, retrieved the Daily Distance Reduction score for all states since February 24. This feature measures the change between the average distance traveled per device for each day and the average distance traveled on the same weekday during the four weeks prior to the COVID-19 outbreak in the U.S. (February 10–March 8). Based on this percentage change, each state is given a letter grade on each day according to the following rules [21]:

- **A**: > 70% decrease
- **B**: 55–70% decrease
- **C**: 40–55% decrease
- **D**: 25–40% decrease
- **F**: < 25% decrease or increase.

*2.4.2 Google Community Mobility Reports.* The publicly available global Google Mobility Report [6] describes longitudinal changes in population movement trends over the course of the COVID-19 outbreak. These movement trends are divided into categories for retail and recreation, groceries and pharmacies, parks, transit stations, workplaces, and residential. For each category, the report provides the percentage change in visitation or time spent in places of that category relative to a baseline, which is computed as the median value for each weekday from the 5-week period from January 3, 2020, to February 6, 2020. The data are aggregated from anonymized users who have opted in to sharing their location history in Google Maps.

## 2.5 Google Search Trends

Using our collection of COVID-19-related news, we first extracted a set of keywords by tokenizing and lemmatizing the titles of each news article. Next, we retrieved the 1,000 most frequently mentioned terms, the first 10 of which are reported in Table 3.

We then scraped Google Trends [7] for the longitudinal "Interest over Time" of each keyword in each U.S. state. For each keyword, Trends measures web search popularity by taking an anonymized sample of Google

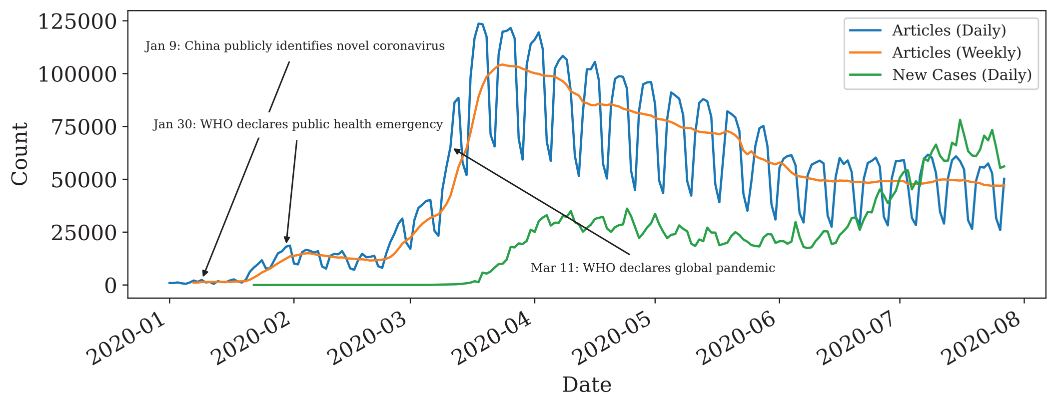


Fig. 1. The number of COVID-19 related articles extracted from the GKG, measured daily and weekly from January 1 through July 27, 2020, plotted alongside the number of new COVID-19 cases reported in the U.S.

searches and dividing the total count of searches containing the given keyword by the total searches associated with a particular location and time range. This value is normalized between 0 and 100 to represent search interest relative to the given state and time, where 100 represents peak popularity for the term and 0 represents a lack of available data for the given term.

## 3 PRELIMARY ANALYSIS

### 3.1 Quantity of News

Through July 27, 2020, we have extracted data on over 10.4 million news articles related to the COVID-19 pandemic. Figure 1 shows the daily and weekly article counts from January 1 through July 27, 2020. The daily oscillation represents a consistent pattern that fewer articles are published on Saturdays and Sundays. The weekly coverage increased at the end of January, around when the first case was confirmed in the United States (January 20) and the Chinese authorities quarantined the city of Wuhan (January 23). A local peak of 18,636 articles were published on January 31, the day after the WHO declared a public health emergency. However, average weekly coverage slowly declined until the last week of February, when cases surged in Italy and Iran. At this point news coverage surged through the first reported death in the United States (February 29) and the WHO's declaration of a global pandemic (March 11) to a global peak of 123,623 articles (March 18). Since then, coverage has decreased steadily, even as new cases reached a global peak of 78,157 (July 16). Even after the number of new cases has begun to decrease, the news coverage has continued to decrease at a faster rate. This suggests that, on a broad scale, news sources were most interested in reporting the novel events surrounding the beginning of the pandemic.

### 3.2 News Coverage by Political Bias

Of the 10.4 million articles extracted from the GKG, about 2.7 million were published by the sources evaluated for bias by MBFC or AllSides. Figure 2 shows the daily count of articles published by each bias category, each of which follow a similar trend to the total article count. This is corroborated by distance correlation tests [19] performed with respect to the daily observations of articles from all sources. These tests are summarized in Table 4 and reported correlation coefficients ≥0.980 (where 0 implies independence and 1 implies perfect correlation) for each bias category except "Scientific" and "Conspiracy-Pseudoscience," which reported coefficients of 0.852 and 0.844, respectively. In other words, all bias categories except "Scientific" and "Conspiracy-Pseudoscience" exhibited very similar temporal trends. In the case of "Scientific" sources, correlation was lowest toward the beginning of 2020, which indicates that scientific sources published more slowly during the early stages of the pandemic.

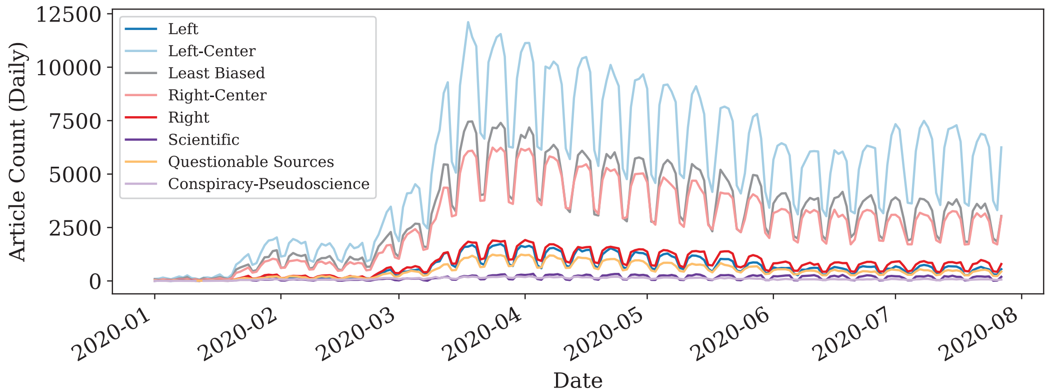


Fig. 2. The number of COVID-19 related articles extracted from the GKG and grouped by source bias, measured daily from January 1 through July 27, 2020.

Table 4. Distance Correlations for the Daily Counts of News Published by Evaluated Sources, Calculated with Respect to the Daily Counts of Articles Published by All Sources

| | Distance Correlations | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Source Bias** | **Jan** | **Feb** | **Mar** | **Apr** | **May** | **June** | **July** | **All** |
| Left | .991 | .943 | .988 | .977 | .978 | .962 | .976 | .981 |
| Left-Center | .998 | .990 | .993 | .990 | .996 | .993 | .994 | .991 |
| Least Biased | .995 | .986 | .997 | .990 | .985 | .992 | .990 | .980 |
| Right-Center | .997 | .986 | .995 | .994 | .991 | .990 | .980 | .995 |
| Right | .991 | .975 | .994 | .988 | .985 | .982 | .982 | .992 |
| Scientific | .862 | .876 | .928 | .942 | .955 | .976 | .972 | .852 |
| Questionable Sources | .986 | .936 | .985 | .977 | .970 | .945 | .982 | .980 |
| Conspiracy-Pseudoscience | .975 | .877 | .886 | .859 | .908 | .798 | .901 | .844 |

The lower correlation of the distribution of articles published by these two bias categories may be attributable to noise, at least in some degree. As Figure 3 shows, both Scientific and Conspiracy-pseudoscience represent only a small percentage of the collection of COVID-19-related articles. Further, Figure 4 reveals that the representation of articles published by Scientific sources, when measured as a percentage of total published news, is significantly lower (0.72×) for COVID-19-related news when compared to a baseline of all 2019 articles, of which Scientific sources accounted for 1.5% of the records. A similar trend is found in Least Biased sources, whose representation is 0.94× with respect to the baseline—a significant decrease given that they account for 23.7% of the 2.7 million COVID-19 articles. However, some bias categories have increased representation in COVID-19-related news: Right sources increased their representation by 1.12× and Right-center sources by 1.06×. Table 5 shows that this contrast between more biased sources (specifically Left and Right) and less-biased sources (Least Biased and Scientific) is attributable to two factors: (1) the change in amount of total news published in 2020 relative to baseline and (2) the proportion of all published articles that are about COVID-19. While Scientific sources have increased their total news output, a much smaller percentage of this news is related to COVID-19 than Right sources.

One of the main implications of this increase in representation from biased sources is that it increases the prevalence of false information. According to MBFC's evaluations, summarized in Figure 5, more politically extreme sources tend to be less factual. Consider Fox News, a Right source with a Mixed fact rating. Through July 27, 2020, we found that Fox has published 40,711 articles (1.13× baseline), of which 33.3% are related to

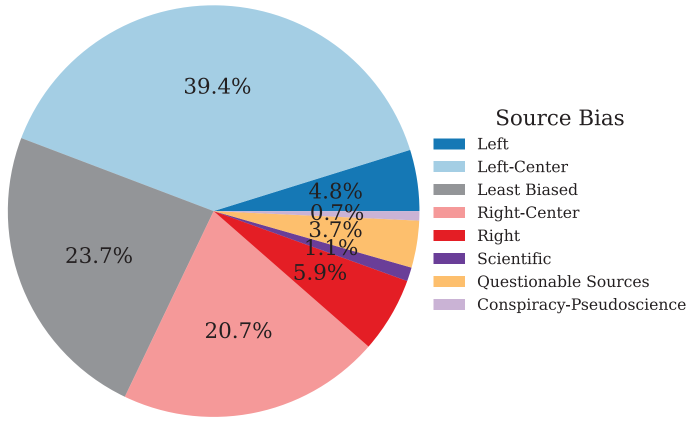


Fig. 3. The representation of each bias category in COVID-19-related news, measured as a percentage of all articles.

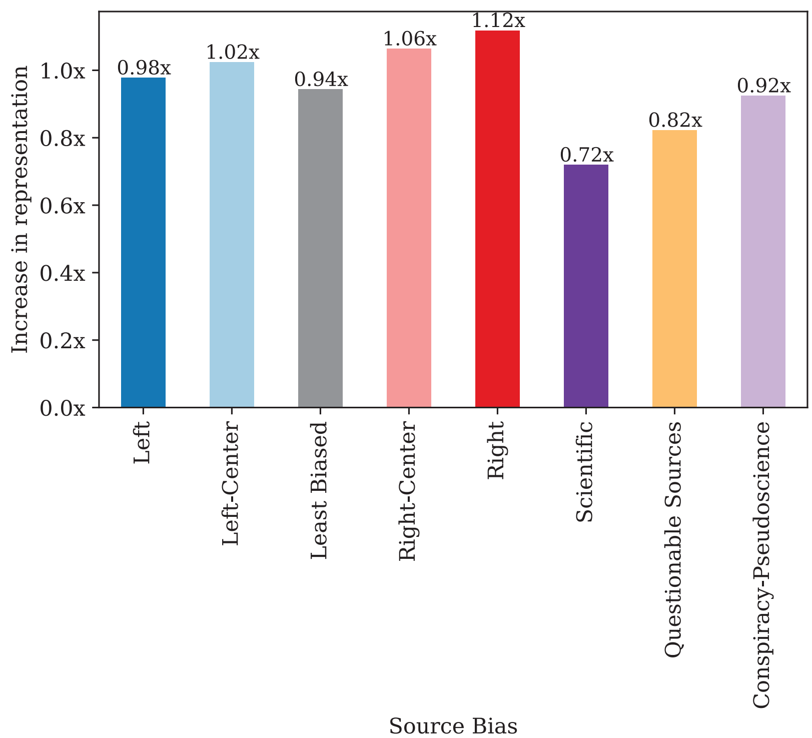


Fig. 4. The increase of each bias category's representation in COVID-19-related news, measured as a ratio of the percentage representation of articles in the COVID-19-related news against a baseline of all 2019 news.

COVID-19—accounting for almost 1% of all COVID-19 news. This is quite consequential when we consider that (a) much of Fox's coverage of the pandemic, especially in its early stages, was inconsistent with scientific reports [12], and (b) this kind of misinformation is directly correlated with problematic social patterns (such as failing to follow social distancing guidelines [10, 18] and believing that scientific and media claims about the severity of the pandemic are exaggerated [15]) and higher numbers of cases and deaths [12]. We suggest that further study

Table 5. Summary of Changes to News Patterns in 2020 (January 1 through July 27)

| **Source Bias** | $\frac{\|N_{2020}\|}{\|N_{baseline}\|}$ | **% COVID-19 related** |
|---|---|---|
| Left | 0.934 | 27.6% |
| Left-Center | 0.973 | 25.1% |
| Least Biased | 0.922 | 24.8% |
| Right-Center | 0.994 | 26.3% |
| Right | 1.051 | 27.1% |
| Scientific | 1.046 | 16.0% |
| Questionable Sources | 1.030 | 20.5% |
| Conspiracy-Pseudoscience | 0.918 | 27.5% |

$\mathbf{N_{2020}}$ is the set of all articles published thus far in 2020, and $\mathbf{N_{baseline}}$ is the baseline set of all articles published in the last seven months of 2019. **% COVID-19 related** describes the percentage of articles in $N_{2020}$ that are related to COVID-19.

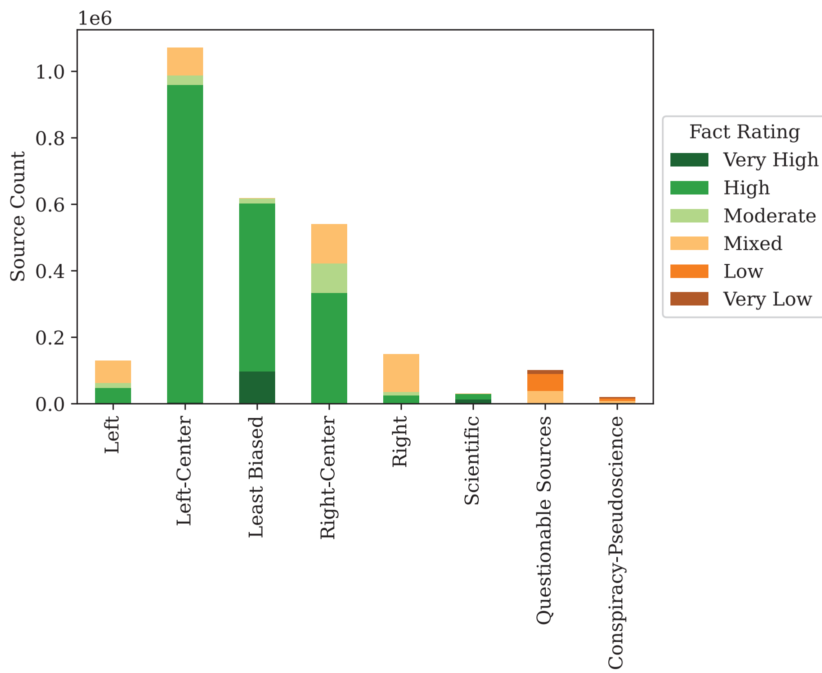


Fig. 5. The distribution of MBFC's fact ratings (grouped by the source's political bias rating) shows that more politically biased sources tend to be less factual.

of these trends in news sources and biases could provide further insight into sources of misinformation and its effects on social response to the pandemic.

## 4 CONCLUSION AND FUTURE WORK

By aggregating multimodal data from many sources that represent a variety of social signals in the United States, we have begun to explore the effects of the COVID-19 pandemic on the society as well as the reaction of the society as measured by news, search, mobility trends, and so on. Our current data includes COVID-19 case data, demographic data, longitudinal news and web search trends, media bias data, and mobility reports, but there are many other types of social signals that could be studied to better understand and model the effects of the pandemic. These could include social media trends, economic patterns, and additional healthcare data. In this article, we analyzed the quantity of news coverage and showed that the amount of COVID-19-related news peaked just after the announcement of the pandemic, after which it steadily decreased. We additionally explored media bias and demonstrated that, with respect to quantity, all groups of political biases published news in a similar pattern and that more scientific sources have significantly less representation in the COVID-19-related news when compared to their pre-pandemic baseline. We further discussed the relationships of this finding to the spread of consequences of misinformation. There are many opportunities to examine other relationships between signals, such as the influence of news on social distancing and web searches, correlations between web searches and news topics, and differences of these effects between locations and demographics. We additionally hope to extend this data and work beyond the United States to understand the effects of the COVID-19 pandemic around the world.

## REFERENCES

[1] 2015. Data Trends and Maps. Retrieved from https://www.cdc.gov/dhdsp/maps/dtm/index.html.
[2] 2015. U.S. Diabetes Surveillance System. Retrieved from https://gis.cdc.gov/grasp/diabetes/DiabetesAtlas.html.
[3] 2015. USCS Data Visualizations. Retrieved from https://gis.cdc.gov/Cancer/USCS/DataViz.html.
[4] 2019. Population Distribution by Age. Retrieved from https://www.kff.org/other/state-indicator/distribution-by-age.
[5] 2020. AllSides Media Bias Ratings. Retrieved May 11, 2020 from https://allsides.com/media-bias.
[6] 2020. Google COVID-19 Community Mobility Reports. Retrieved from https://www.google.com/covid19/mobility/.
[7] 2020. Google Trends. Retrieved from https://www.google.com/trends.
[8] 2020. Media Bias/Fact Check: The Most Comprehensive Media Bias Resource. Retrieved April 20, 2020 from https://mediabiasfactcheck.com.
[9] 2020. Unemployment Rates for States. Retrieved from https://www.bls.gov/web/laus/laumstrk.htm.
[10] Martin Andersen. 2020. Early evidence on social distancing in response to COVID-19 in the United States. *SSRN Electronic Journal* 10.2139/ssrn.3569368 (June 2020).
[11] US Census Bureau. 2020. American Community Survey (ACS). Retrieved from https://www.census.gov/programs-surveys/acs/.
[12] Leonardo Bursztyn, Aakaash Rao, Christopher Roth, and David Yanagizawa-Drott. 2020. *Misinformation During a Pandemic. University of Chicago, Becker Friedman Institute for Economics Working Paper* 2020-44 (2020).
[13] Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. 2019. COVID-19 Data Repository. Retrieved May 31, 2020 from https://github.com/CSSEGISandData/COVID-19.
[14] Felix Hamborg, Karsten Donnay, and Bela Gipp. 2019. Automated identification of media bias in news articles: An interdisciplinary literature review. *Int. J. Dig. Libr.* 20, 4 (2019), 391–415.
[15] Mark Jurkowitz and Amy Mitchell. 2020. Cable TV and COVID-19: How Americans perceive the outbreak and view media coverage differ by main news source. *Pew Research Center Journalism and Media*. Retrieved from https://www.journalism.org/2020/04/01/cable-tv-and-covid-19-how-americans-perceive-the-outbreak-and-view-media-coverage-differ-by-main-news-source/.
[16] Kalev Leetaru and Philip A. Schrodt. 2013. Gdelt: Global data on events, location, and tone, 1979–2012. In *Proceedings of the ISA Annual Convention*, Vol. 2. Citeseer, 1–49.
[17] Filipe N. Ribeiro, Lucas Henrique, Fabricio Benevenuto, Abhijnan Chakraborty, Juhi Kulshrestha, Mahmoudreza Babaei, and Krishna P. Gummadi. 2018. Media bias monitor: Quantifying biases of social media news outlets at large-scale. In *Proceedings of the 12th International AAAI Conference on Web and Social Media*.
[18] Andrey Simonov, Szymon K. Sacher, Jean-Pierre H. Dubé, and Shirsho Biswas. 2020. *The Persuasive Effect of Fox News: Non-compliance with Social Distancing During the COVID-19 Pandemic*. Technical Report. National Bureau of Economic Research.
[19] Gábor J. Székely, Maria L. Rizzo, Nail K. Bakirov, et al. 2007. Measuring and testing dependence by correlation of distances. *Ann. Stat.* 35, 6 (2007), 2769–2794.
[20] Unacast. 2020. Unacast Social Distancing Dataset. Retrieved April 17, 2020 from https://www.unacast.com/data-for-good.
[21] Unacast. 2020. US Social Distancing Scoreboard Methodology (unpublished).
[22] Seunghak Yu, Giovanni Da San Martino, and Preslav Nakov. 2019. Experiments in detecting persuasion techniques in the news. *arXiv:1911.06815*. Retrieved from https://arxiv.org/abs/1911.06815.

Received July 2020; revised October 2020; accepted October 2020