scientific reports

Check for updates

OPEN

Sentinel node approach to monitoring online COVID-19 misinformation

Matthew T. Osborne[1], Samuel S. Malloy[2], Erik C. Nisbet[3], Robert M. Bond[4] & Joseph H. Tien[5]✉

Understanding how different online communities engage with COVID-19 misinformation is critical for public health response. For example, misinformation confined to a small, isolated community of users poses a different public health risk than misinformation being consumed by a large population spanning many diverse communities. Here we take a longitudinal approach that leverages tools from network science to study COVID-19 misinformation on Twitter. Our approach provides a means to examine the breadth of misinformation engagement using modest data needs and computational resources. We identify a subset of accounts from different Twitter communities discussing COVID-19, and follow these 'sentinel nodes' longitudinally from July 2020 to January 2021. We characterize sentinel nodes in terms of a linked domain preference score, and use a standardized similarity score to examine alignment of tweets within and between communities. We find that media preference is strongly correlated with the amount of misinformation propagated by sentinel nodes. Engagement with sensationalist misinformation topics is largely confined to a cluster of sentinel nodes that includes influential conspiracy theorist accounts. By contrast, misinformation relating to COVID-19 severity generated widespread engagement across multiple communities. Our findings indicate that misinformation downplaying COVID-19 severity is of particular concern for public health response. We conclude that the sentinel node approach can be an effective way to assess breadth and depth of online misinformation penetration.

The proliferation of online misinformation has presented a challenge to public health throughout the COVID-19 pandemic[1], and is characterized by broad demographic and geographic reach[2]. Misinformation exposure reduces adherence to non-pharmaceutical interventions (NPIs)[3] and consequently has driven negative health outcomes in groups with high exposure to such content[4]. The U.S. Surgeon General issued an advisory declaring health misinformation a "serious threat to public health"[5]. The significant potential for misinformation to drive behavioral change such as vaccine hesitancy[6,7] underscores the importance of identifying and mitigating COVID-19 misinformation.

Assessing the public health risk posed by specific types of misinformation is challenging in the low signal-to-noise environment of social media platforms, which are key mechanisms for COVID-19 misinformation spread[8,9]. Since March 2020 tens of millions of tweets regarding COVID-19 have been posted daily[10,11], of which only a fraction contain misinformation. Developing techniques for researchers and public health officials to identify misinformation that is circulating online as well as to distinguish which types of misinformation should be addressed quickly is of critical importance.

Large-scale repositories of COVID-19 related tweets provide important resources for identifying COVID-19 misinformation[2,12–14]. The computational demands of working with these large datasets are, however, considerable, and determining which tweets correspond to misinformation is not straight forward. Moreover, a seemingly large volume of false or misleading COVID-19 content does not necessarily imply broad engagement. The breadth and depth of COVID-19 misinformation are important considerations: misinformation confined to a small, isolated group of users has different public health ramifications than misinformation being consumed by a large portion of the population across different communities and demographic groups.

[1]The Ohio State University, Columbus, OH, USA. [2]Glenn College of Public Affairs, The Ohio State University, Columbus, OH, USA. [3]School of Communication, Northwestern University, Evanston, IL, USA. [4]School of Communication, The Ohio State University, Columbus, OH, USA. [5]Department of Mathematics, The Ohio State University, Columbus, OH, USA. ✉email: tien.20@osu.edu

Here, we describe a longitudinal, network-based approach to characterize the breadth of online engagement with specific misinformation topics. We focus on Twitter, an important platform for the dissemination of content and influencing opinion, including COVID-19 misinformation[2,15,16]. Twitter has a natural network structure, for example through retweets, followership and co-linkage to other domains. We utilize this network structure to detect so-called 'communities', tightly-knit groupings of users and accounts, discussing COVID-19 on Twitter. Communities play a particularly important role in the propagation of online content. Content sharing within communities is a key mechanism for exposure to and amplification of viewpoints[17,18], and provides important context for breadth of misinformation penetration. Incorporating community structure and other network features into misinformation monitoring efforts can thus enhance the detection of false or misleading content that has the potential to become broadly disseminated and to drive behavior change.

From each community we selected and monitored a subset of accounts, referred to as 'sentinels', from July 2020 to January 2021. This borrows from an epidemiological concept known as sentinel monitoring[19], in which sampling sites are chosen based on their ability to both represent a key population and their production of data that is of a known and consistent quality. Objectives of epidemiological sentinel monitoring include detecting presence and patterns of disease occurrence[19]. Analogously, we define sentinel accounts here as a subset of online accounts that can be used to detect the presence and patterns of content within online communities. Sentinel nodes are thus used to identify and characterize the types of content circulating within online communities. In the context of misinformation on social media platforms, sentinel nodes provide a way to assess both the depth and breadth of misinformation penetration. Note that there are many possible ways to select sentinel nodes from a given community. Here we select a subset of influential nodes from each community to use as sentinel nodes, as a majority of Twitter content is driven by a small subset of accounts[20]. Thus, there is some overlap between the selection of sentinel accounts and opinion leadership within the constituent online communities. The concept of sentinel accounts is, however, distinct from that of opinion leaders, and other selection criteria for sentinel accounts may yield sentinels that are not themselves influential but are nonetheless useful for monitoring and diagnostic purposes.

Sentinel communities were characterized according to the link sharing behavior of their constituent accounts. Twitter community structure has been observed to be highly assortative with respect to media preferences[21], which have been shown to be correlated with attitudes regarding COVID-19[22,23]. Characterization of communities in terms of media preference allows for the establishment of baseline metrics for the propensity to engage with and propagate COVID-19 misinformation, and comparison of tweet similarity across the media preference spectrum. This is critical for assessing breadth of misinformation penetration, including identifying consequential events where specific false or misleading content propagates from a fringe group that regularly engages in conspiracy theories to a more mainstream community.

This structured 'sentinel node' approach provides a way to compare tweets across segments of the Twitter ecosystem and identify misinformation topics with broad penetration, while only requiring modest data storage and computational resources. Further, our longitudinal approach helps address issues with selection bias and temporal changes in phrase and hashtag usage that can be problematic for cross-sectional studies[24].

We find evidence that the linked-media preferences of the sentinel accounts is strongly correlated with their propensity to post COVID-19 misinformation. Accounts linking predominantly to right-leaning domains tended to post more COVID-19 misinformation than those linking predominantly to left-leaning domains. Importantly, we observe that sensational topics were largely confined to a subset of conspiracy minded communities, while misinformation on COVID-19 severity was much more widespread. This includes a large 'virality' event that appears to be catalyzed by former President Donald Trump. These results indicate that perceived COVID-19 severity is of particular concern for public health.

Results

Before describing our empirical results we briefly overview our methods, which are described in further detail below (see "Methods" section). Sentinel nodes were identified as prominent accounts within communities from the retweet network among accounts posting about COVID-19 (see Methods section "Sentinel recruitment and data collection"). We then categorized the political valence of each community by examining patterns of linking to external websites (see Methods section "Sentinel community characterization using linked domains"). Tweet content was characterized by topic and cluster by a set of human coders (see Methods section "COVID-19 misinformation by cluster"). Next, tweets containing content about vaccines or the severity of COVID-19 were identified using a set of keywords that were commonly used to discuss these topics during the period of observation (see Methods section "COVID-19 vaccines and disease severity content"). Finally, we examined inter-cluster content spread by examining the similarity in the text of tweets across communities over time (see Methods section "Flagging inter-cluster content spread").

Twitter communities and sentinel node identification. A query for tweets containing the term 'covid' yielded 168,950 tweets (74% retweets) sent on May 27, 2020 by 142,331 unique accounts (Methods, Section "Sentinel recruitment and data collection"). We constructed a weighted, directed graph with adjacency matrix A, where A_{ij} is the number of times that account j retweeted account i. The in-degree of node k is thus the number of times that k was retweeted, and the out-degree of k equals the number of times that k retweeted other accounts. Here we say that account i retweeted account j if account j was the original poster of the retweeted content. We excluded self-loops (self-retweets). The largest connected component of this 'retweet network', denoted G, contained 78,680 nodes and 87,030 total retweets.

Modularity maximization using a Louvain method[25] yielded 148 communities in G, ranging in size from 5,641 to 6 nodes (Methods, Section "Sentinel recruitment and data collection"). Let C denote the largest 28

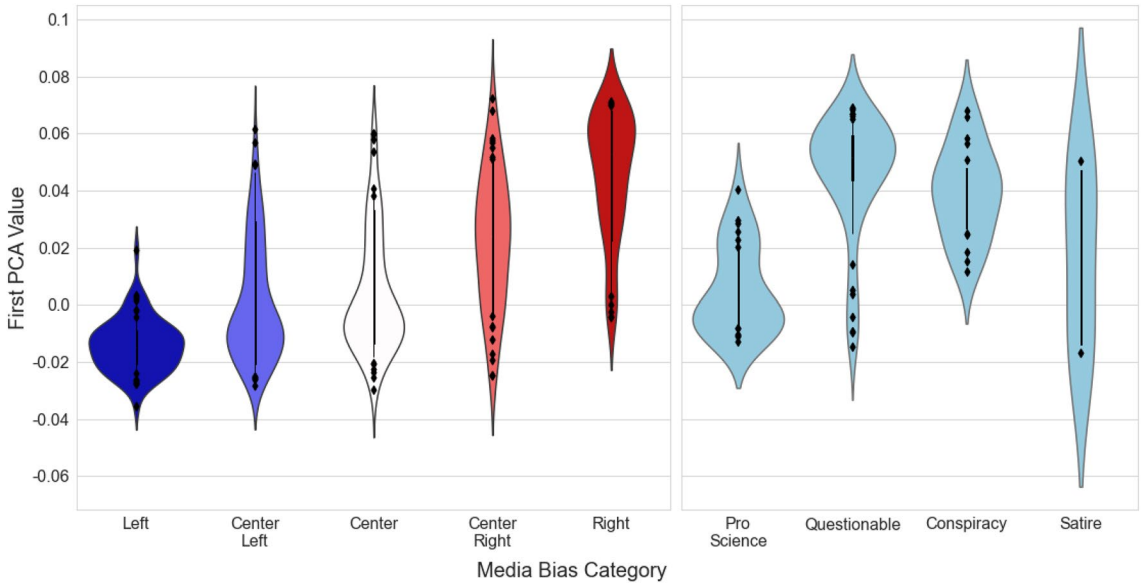

Figure 1. Distribution of linked domain first principal component entries by Media Bias Fact Chart (MBFC) categories[27]. 38% of linked domains from sentinel nodes were listed in MBFC as either Left, Center Left, Center, Center Right, or Right, and 9% were listed as either Pro Science, Questionable, Conspiracy, or Satire. Treating the political categories as an ordinal variable with 'Left' being equivalent to 1, 'Center Left' = 2, 'Center' = 3, 'Center Right' = 4, and 'Right' = 5, there is a correlation of 0.66 between the first PCA value and the political tilt of the domains represented.

communities comprised primarily of English-speaking, domestic accounts. Each community in C contained at least 429 nodes, from which we selected the 15 most-frequently retweeted accounts as *sentinel nodes* to follow longitudinally. Note that the in-degree distribution for G has a heavy tail: the 15 most highly retweeted nodes in a community account for on average 84% of the total number of times that nodes in that community were retweeted (median 86%; interquartile range 77% to 96%). We will refer to the set of sentinel nodes as S, and the sentinel nodes drawn from community c in G as S_c.

To examine robustness of the community assignments, we constructed a second retweet network $\tilde{G}$ based upon the same search term over a second, later time interval (June 8-9, 2020). While 35% (148/420) of the sentinel nodes from G were present in $\tilde{G}$, all of the 28 communities in S were represented in $\tilde{G}$ (in the sense that at least one node from each community in S was present in $\tilde{G}$). The Rand index comparing the Louvain-detected communities for nodes common to both G and $\tilde{G}$ was 0.96 (z-score 897.7)[26], indicating significant correlations in the community structure of G and $\tilde{G}$ and consistent with robustness in the community structure of the Twitter conversation regarding COVID-19 over this time period.

The most recent 3200 tweets posted by each sentinel node in S were collected at least once per week between July 1, 2020 and January 6, 2021, yielding a dataset of all 4,130,909 tweets posted by the sentinel nodes over this time period, with the possible exception of deleted tweets.

Sentinel community characterization using linked domains. We characterized sentinel communities through the domains shared in their tweets posted from 7/1/2020–10/3/2020. We used principal components analysis (PCA) to derive a linked domain score for each of our sentinel communities. Notably, we made no assumptions on the bias or reliability of the shared domains, so any correlation between the linked domain score with COVID-19 misinformation is not a contrivance of the dimension reduction process.

The 2,152,849 tweets posted from 7/1/2020 - 10/3/2020 contained 706,564 links. Of these, 147,510 linked to Twitter and 44,529 were shortened, yielding 514,525 non-shortened links to 8624 distinct non-Twitter domains. PCA on the domain frequency matrix (Methods, Section "Sentinel community characterization using linked domains") from each of the 28 communities in C resulted in the distribution of domains along the first principal component value shown in Figure 1. Note that media sources with conservative-leaning partisan scores as described in[27] are associated with positive first component values (eg. `foxnews.com`, `oann.com`), while more liberal media sources are associated with negative entries (e.g. `motherjones.com`, `washingtonpost.com`). A table of the domains associated with the 30 most positive and negative first principal component values is given in the Supplementary Materials. In the remainder we will refer to the first PCA score as the 'Left' / 'Right' linked domain score.

Left / Right linked domain scores for each sentinel node community are shown in Fig. 2. While hierarchical clustering[28] results in a dendrogram in which the highest silhouette score (0.894)[29] corresponds to two clusters, we instead chose a dendrogram cut producing three clusters (silhouette score of 0.737) as seen in Fig. 2. The communities in the 'Right' cluster were skewed to the left due to disproportionate sharing of a domain connected to a single sentinel node in each 'Right' community. Repeating the PCA and clustering process after removing

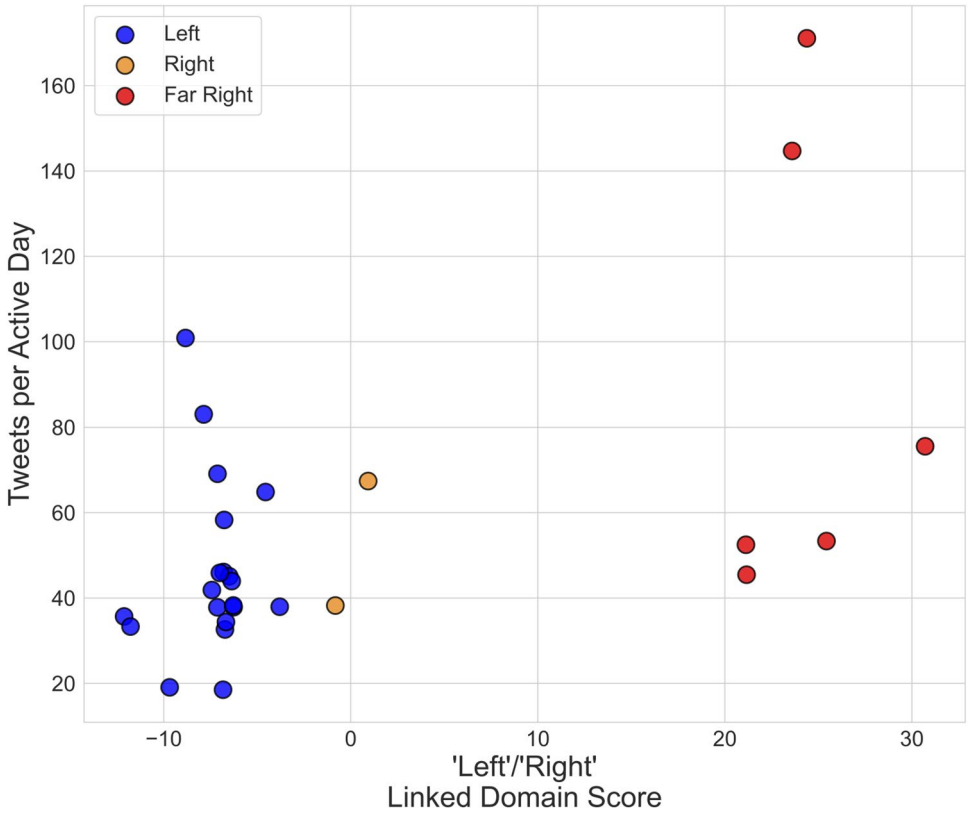

Figure 2. Sentinel communities projected onto the first component of the linked domain PCA space. Colors correspond to clusters resulting from mean linkage clustering on the first principal component score. The vertical axis corresponds to the total tweets posted by the community over the observation period, normalized by the number of active days for the community's sentinel nodes (Methods, Section "Sentinel recruitment and data collection").

these two domains shifted the 'Right' communities approximately 10 units to the right on the linked domain score, while the 'Left' and 'Far Right' clusters exhibited negligible differences in their Fig. 2 positioning (see Supplementary Materials). The dendrogram following removal of these two domains resulted in three clusters having the highest silhouette score (0.814).

The cluster in Fig. 2 with the most negative inter-cluster average linked domain score contained several Democratic politician accounts, while the cluster with the most positive inter-cluster mean score contains some Republican politician accounts. Profiles in the most positive cluster were, on average, about seven times more likely to post tweets containing strings associated with the QAnon conspiracy theory (e.g. "qanon", "wwg1wga", "new world order") than acounts in the other two clusters (see Supplementary Materials). We thus labeled the cluster with most positive average score as 'Far Right' and the cluster with most negative average score as 'Left'. The remaining cluster contained accounts associated with mainstream conservative media pundits. When combined with the additional cluster analysis described above, this led us to label the final cluster as 'Right'.

COVID-19 misinformation by cluster. To compare COVID-19 misinformation prevalence in tweets from the Left, Right and Far Right clusters, we randomly sampled 1,151 tweets stratified by cluster and topic. The four topics examined were COVID-19 mortality, masks, hydroxychloroquine, and Plandemic. Tweet identification utilized a substring search (Methods, Section "COVID-19 misinformation by cluster").

Figure 1 indicates that the 'Conspiracy' and 'Questionable' media bias categories tended to be associated with more positive (Right-leaning) linked domain scores, suggesting that tweets containing misinformation may be more prevalent in the Right and Far Right clusters. To test whether prevalence differed between clusters, we developed a coding sheet with factual background information on the most common false or misleading claims associated with the aforementioned topics (see Supplemental Material). Four human coders employed the reference sheet to individually annotate (1 =misinformation, 0 =no or unsure) whether each tweet presented misinformation. Coding results yielded a Krippendorff's alpha of 0.73, indicating adequate inter-coder reliability[30].

The summative results of the human coding are presented in Table 1. In total, 14.4% of the tweets from the Left cluster on these four topics contained misinformation, compared to 85.1% and 88.2% of tweets from the Right and Far Right clusters, respectively. This difference between the three clusters was significant, $\chi^2(2, N = 1151) = 563.3$, $p < .001$, rejecting the null hypothesis of no difference in misinformation prevalence between clusters. Additional analysis indicated the difference in misinformation prevalence between the Right and Far Right clusters was not significant, $\chi^2(1, N = 790) = 1.7$, $p = n.s.$. There was substantial variation in

Cluster	# of Tweets	Facemasks	COVID-19 Mortality	Hydroxychloroquine	Plandemic	Total
		% Misinformation	% Misinformation	% Misinformation	% Misinformation	% Misinformation
Left	361	9.5	14.0	19.0	16.1	14.4
Right	382	72.0	77.0	95.0	98.8	85.1
Far Right	408	82.4	77.5	94.1	99.0	88.2
Total	1151	54.1	56.3	69.5	79.6	64.0

Table 1. Frequency of COVID-19 misinformation within Left, Right, and Far Right clusters by topic. For each topic (facemasks, COVID-19 mortality, hydroxychloroquine, Plandemic), the listed percentage denotes the proportion of the corresponding cluster's tweets on that topic containing misinformation. A small number of tweets contained misinformation regarding multiple topics.

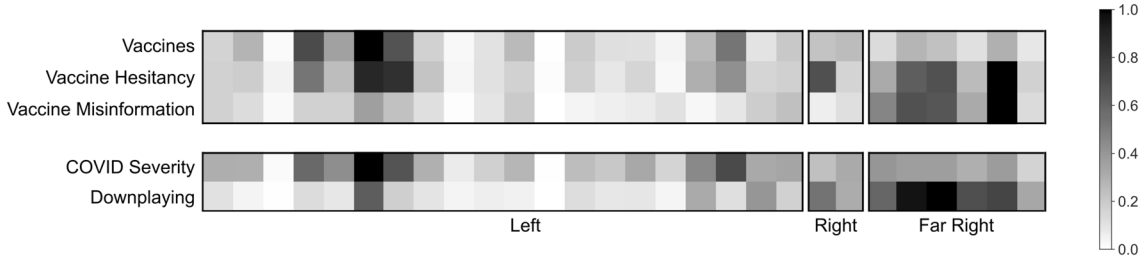

Figure 3. Distribution of topical tweets across communities. Grayscale corresponds to the percentage of the maximum within subtopic per capita tweet rate across communities. Per capita tweet rate corresponds to dividing the total number of topical tweets from each community by the number of active user days for the corresponding community. 'Vaccine Hesitancy' and 'Vaccine Misinformation' tweets are subsets of 'Vaccines' tweets; 'Downplaying' tweets are a subset of 'COVID Severity' tweets. Phrases used for topic identification are given in the Supplementary Materials.

misinformation prevalence by topic, with misinformation appearing less frequently in tweets about face masks (54.1%) and death severity (56.3%) in all three clusters as compared to those about hydroxychloroquine (69.5%) and Plandemic (79.6%).

COVID-19 vaccines and disease severity content. As described in Section "COVID-19 misinformation by cluster", there is high prevalence of misinformation regarding COVID-19 mortality specifically and COVID-19 severity more broadly in tweets posted by the Right and Far Right clusters of accounts. Abundant online vaccine misinformation has also been documented[14]. COVID-19 severity perceptions and vaccination decisions are related: perceived risks of infection versus vaccination factor into vaccination decisions[31,32], and polls indicate that perceived low risk of severe outcome is a potentially important rationale for vaccine hesitancy[32].

Tweets about vaccines and COVID-19 severity from sentinel nodes were obtained by filtering COVID-related tweets on associated keywords and phrases. Example keywords and phrases included 'death rate', 'fatality rate', and 'confirmed cases' for COVID-19 severity, and 'vaccine', 'pfizer', 'moderna', and 'johnson and johnson' for vaccinations. These two topics were further subset with additional keyword and phrase filters to identify tweets concerning vaccine hesitancy, vaccine misinformation, and tweets that downplayed COVID-19 severity. Example keywords and phrases included 'will not take' for vaccine hesitancy, 'change dna' for vaccine misinformation, and 'lower than flu' and 'mild' for downplaying severity. Complete lists are given in the Supplementary Materials.

The COVID-19 severity substring search returned 27,832 tweets from the Left, 1997 from the Right and 7377 from the Far Right. Of those, 1322 Left, 372 Right and 1831 Far Right tweets contained phrases associated with downplaying COVID-19 severity. The vaccination substring search returned 22022 tweets from the Left, 1781 from the Right and 4298 from the Far Right. Of those, 933 Left, 165 Right and 528 Far Right tweets contained vaccine hesitancy phrases, while 312 Left, 19 Right and 401 Far Right tweets contained vaccine misinformation related phrases.

Figure 3 compares tweet rates between communities for tweets concerning vaccines, vaccine hesitancy, vaccine misinformation, COVID-19 severity, and downplaying severity. For each topic and community, we first compute the tweets per active account day for that cluster by dividing the total number of topical tweets by the number of active account days (Methods, Section "Sentinel recruitment and data collection"). We then scale this quantity for community i by dividing the tweets per active account day for i by the sum of the tweets per active account day across all communities. This quantity is labeled the scaled per capita tweet rate for community i. Each row of Fig. 3 shows the scaled per capita tweet rates for a given topic.

We observed tweets containing vaccine hesitancy phrases across each cluster, though relatively few communities in the Left engaged with this topic. Further examination of vaccine hesitancy tweets from the Left indicates that the majority of these tweets were commenting on 'experimental' clinical trial progress for vaccine

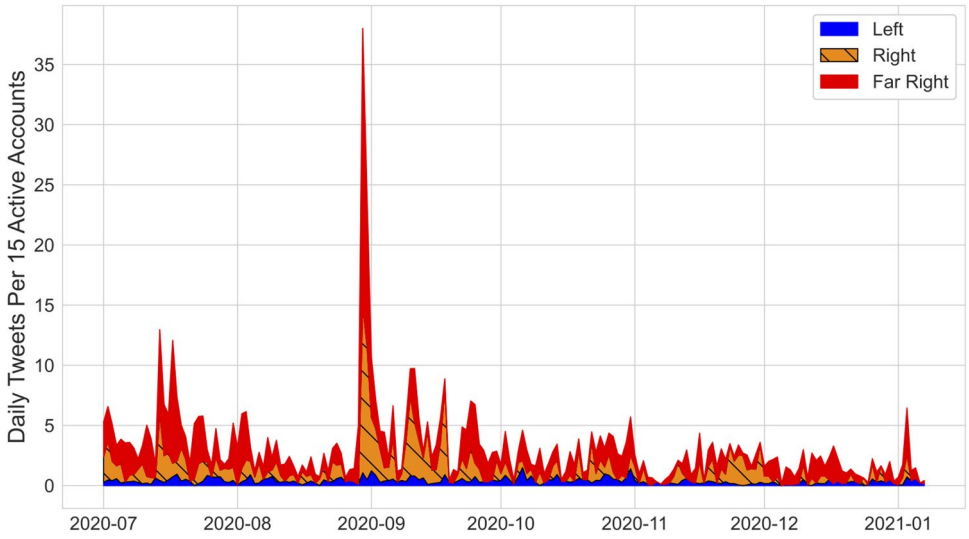

Figure 4. Daily tweets containing phrases associated with downplaying COVID-19 severity by cluster. See the Supplementary Materials for phrase lists. Vertical axis corresponds to daily tweets per 15 active accounts. Sustained high volume of tweets containing phrases downplaying COVID-19 severity are observed from the Far Right (red) and Right (orange) clusters of accounts.

development, rather than expressing hesitancy to vaccinate. Vaccine misinformation and severity downplaying tweets were largely confined to the Right and Far Right. By contrast, scaled per capita rates for all vaccine and severity tweets were highest for the Left. We thus observe selective higher engagement on vaccine hesitancy, vaccine misinformation, and downplaying severity from the Right and Far Right, despite lower engagement overall on vaccine and COVID-19 severity tweets compared to the Left.

Perceived COVID-19 severity. A time series of daily tweets containing phrases downplaying COVID-19 severity aggregated by cluster is shown in Fig. 4. We observe sustained high volume from the Far Right (red) and Right (orange) throughout the study period. We also observe occasional spikes, for example at the end of August. These 'bursts' are considered further in Sect. "Flagging inter-cluster content spread".

Widespread posting of tweets downplaying COVID-19 severity was observed across Right and Far Right communities, indicating cluster-wide engagement with the topic as opposed to engagement restricted to a handful of communities. For all but one of the communities in the Right and Far Right clusters, more than 15% of the tweets about COVID-19 severity contained phrases associated with downplaying COVID-19 risk. The proportion of COVID-19 severity tweets that downplayed COVID-19 risk was over 70% for one community in the Right.

A majority of downplaying tweets from the Right (59%) and Far Right (75%) contained phrases connected to the questioning of reported incidence and death statistics (Methods, Section "COVID-19 vaccines and disease severity content"). An inspection of these tweets reveals that both clusters tended to use words like "overcount", "inflated" and "false positives". They also highlighted various comorbidities identified in the Centers for Disease Control (CDC) death statistics as the true causes of death. Further, a large fraction of tweets downplaying COVID-19 severity from the Right (47%) and Far Right (30%) contained phrases downplaying illness severity at the individual level. Such tweets tended to emphasize the survival rate for someone that contracts COVID-19. These results suggest that the perception that COVID-19 does not have a serious impact on the morbidity and mortality of the population or those individuals that contract the disease is not confined to a small subset of communities, but is widespread among the Right and Far Right clusters.

Vaccines

Sentinel node tweets about COVID-19 vaccinations include both misinformation (e.g. "the COVID vaccine will alter your DNA") as well as vaccine hesitant sentiments (e.g. "the vaccine was rushed"). Both have the potential to undermine public vaccination campaigns. Figure 5 shows daily tweets by cluster containing vaccine misinformation phrases (Top) and vaccine hesitant phrases (Bottom) over the study period (Methods, Section "COVID-19 vaccines and disease severity content").

Tweets containing vaccine misinformation phrases were largely confined to the Far Right, which exhibited multiple days with high vaccine misinformation engagement relative to the other clusters. Three communities accounted for more than 70% of all Far Right vaccine misinformation tweets, suggesting heterogeneous engagement with such content. Inspection of these tweets indicates that vaccine misinformation content was not dominated by any one narrative. Subtopics for vaccine misinformation tweets from the Far Right are given in Table 2. Apart from those tweets related to 'Plandemic' (a conspiracy theory not solely confined to vaccines), no single topic accounts for greater than 15% of the vaccine misinformation space.

Vaccine hesitancy tweets were more widespread among the Right and Far Right clusters compared with the Left (Fig. 5, Bottom). There are several days on which both these clusters engage with hesitancy content. Notably, these days occur both before and after Pfizer's November announcement of efficacy results from their phase 3 trial. The percentage of vaccine-related tweets containing vaccine hesitancy phrases was comparable pre- and

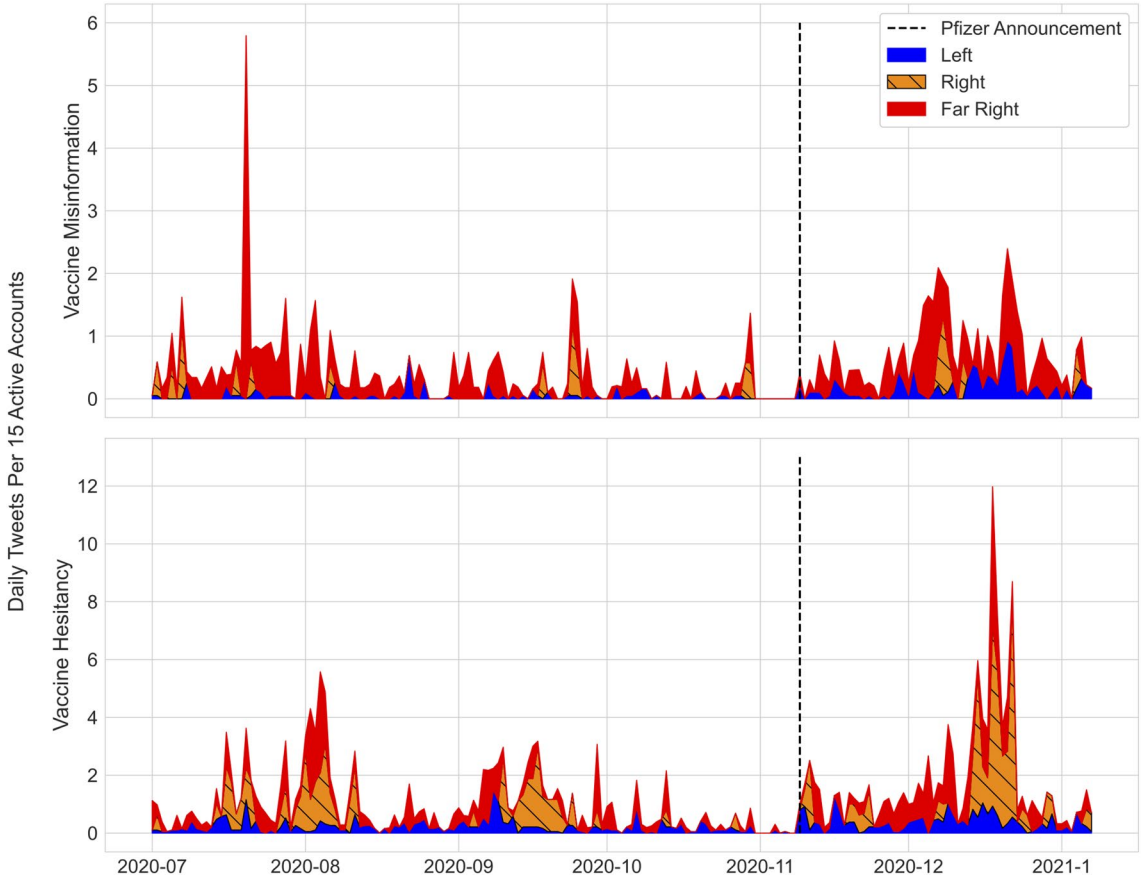

Figure 5. Daily tweets by cluster containing phrases associated with vaccine misinformation (Top) or vaccine hesitancy (Bottom). Vertical axis corresponds to daily tweets per 15 active accounts. Tweets containing vaccine misinformation phrases are primarily confined to the Far Right (red). Tweets containing vaccine hesitancy phrases occur sporadically in both the Far Right (red) and the Right (orange), at a higher volume than in the Left (blue). Vertical line corresponds to announcement of Pfizer phase three trial results on November 9, 2020.

Vaccine Misinformation Topic	Percent of Misinformation Tweets
Contains microchips	13.5%
Alters your DNA	7.3%
Will sterilize you	10.4%
Contains aborted cells	11.8%
Depopulation/genocidal weapon	14.7%
Plandemic	32.11%
Other	7.61%

Table 2. Prevalence of specific subtopics in Far Right vaccine misinformation tweets. Percentages may surpass 100% as a tweet can be flagged for more than one term or phrase.

post-Pfizer announcement for the Right (9% pre, 10% post) and Far Right (14% pre, 10% post), but decreased by more than half for the Left (7% pre, 3% post).

While the Right and Far Right both posted vaccine hesitant content, the flavor of content was slightly different between clusters. The Far Right's tweets were more likely to use phrases suggesting that the poster will not take the vaccine or urging others not to take the vaccine (38% of their total vaccine hesitant tweets compared to 6% for the Right), while the Right was more likely to mention reports of adverse reactions (42% of their hesitant tweets compared to 13% of the Far Right).

Flagging inter-cluster content spread. Figure 6 (Top) shows trigram cosine similarity between pairs of clusters over time (Methods, Section "Flagging inter-cluster content spread"). The inter-cluster similarity between the Left and the other two clusters is nearly zero for all but a handful of days. By contrast, the Right

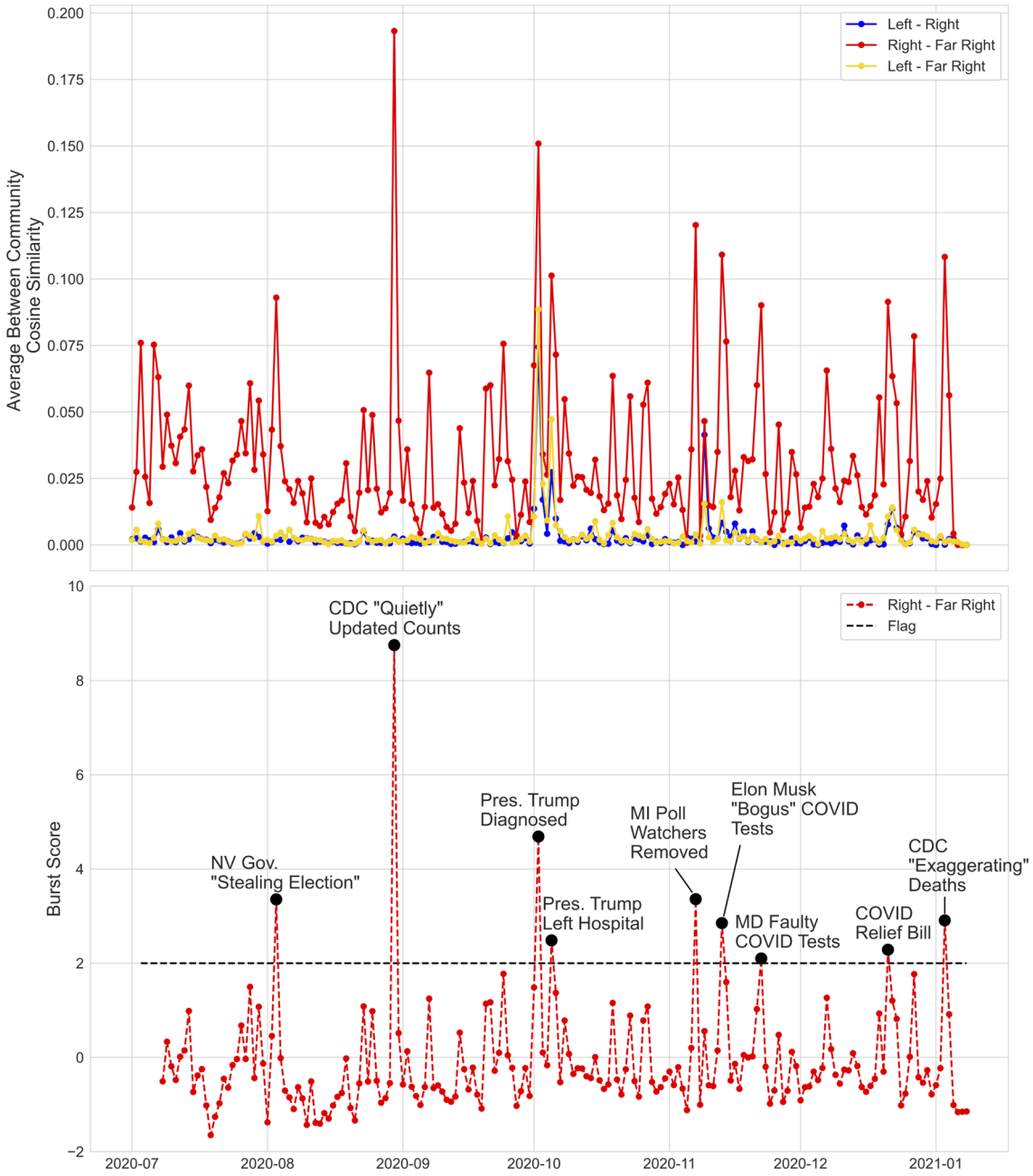

Figure 6. *Top.* Average inter-cluster similarity between the Left, Right, and Far-Right clusters for each day between July 1, 2020 and January 6, 2021. *Bottom.* Inter-cluster burst score [Equation (1)] between the Right and Far Right. Annotated days represent days in which the burst score is greater than or equal to 2. The first seven days of the observation period are omitted to establish baseline.

and Far Right clusters have a non-zero baseline similarity and exhibit multiple days of similarity exceeding 0.10. Additionally, time points are observed where similarity spikes between the Right and Far Right. These bursts may represent emerging content garnering support across different segments of the online ecosystem.

We used a modification of the burst score of Mehrotra et al[33] to identify time points for which inter-cluster similarity is unusually large. Define

$$H(A,B,t) = \frac{s_t(A,B) - \text{mean}_{\tau<t}\{s_\tau(A,B)\}}{\text{SD}_{\tau<t}\{s_\tau(A,B)\}}, \tag{1}$$

where $s_t(A,B)$ is the similarity between clusters A and B on day t and SD denotes the standard deviation. The burst score H gives a measure of inter-cluster similarity on a given day relative to historical similarity.

Figure 6 (Bottom) shows the Right-Far Right similarity burst score (1), with days where $H > 2$ (above the dashed line) flagged for further examination. *Topical tweets* for the flagged days were identified using latent semantic analysis (LSA[34]; Methods, Section "Flagging inter-cluster content spread"). Brief summaries of these

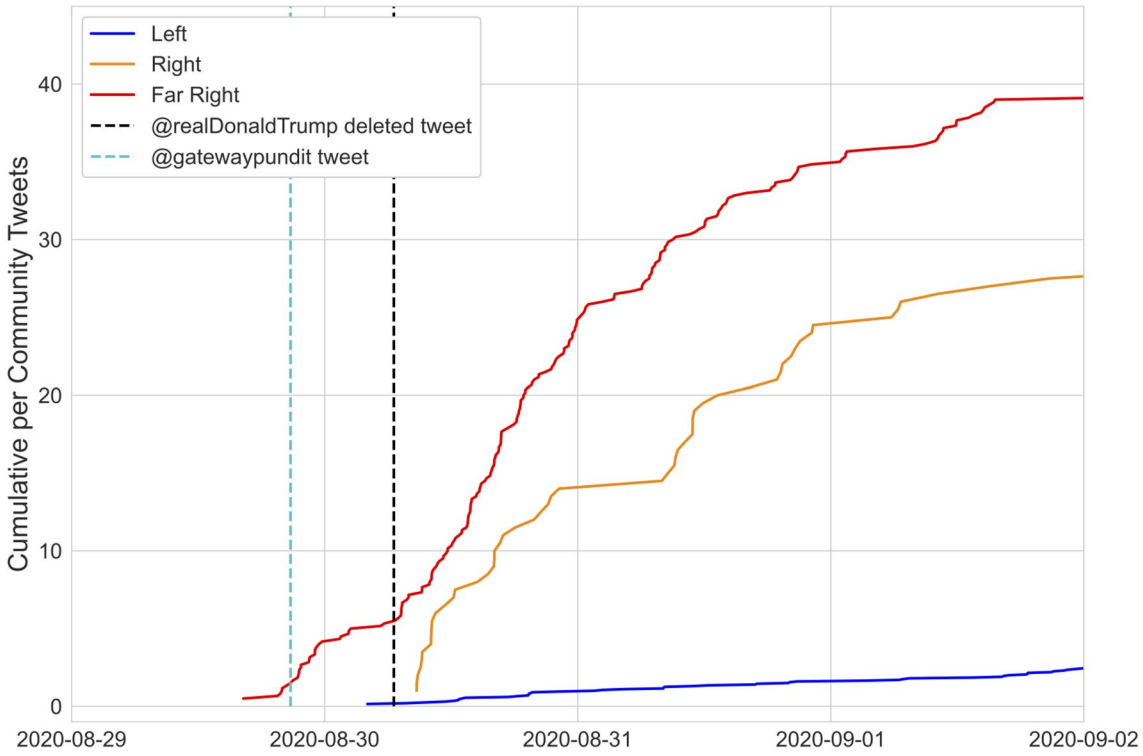

Figure 7. Cumulative per community tweet curves by cluster for tweets discussing the CDC altering its reported COVID-19 death statistics in August of 2020. The cyan dotted line denotes the time of a GatewayPundit tweet that sparked an amplification event in the Far Right and the black dotted line denotes the time of a deleted Donald Trump retweet of a QAnon account promoting the conspiracy.

'topics' are shown in Figure 6 (Bottom). Removal of topical tweets resulted in burst scores below the flagging threshold, consistent with the high burst scores on flagged days being driven by their spread. More in depth information including detailed examination of flagged days can be found in the Supplementary Materials.

Topical tweets from flagged days show three themes: downplaying COVID-19 severity (e.g. doubting CDC reporting statistics and questioning the validity of COVID tests), tying of COVID-19 to positions of the Republican party from the 2020 presidential election (i.e., opposition to mail-in ballots and deriding politicians critical of Donald Trump), and news associated with Donald Trump's COVID-19 diagnosis[35] and subsequent departure from Walter Reed Medical Center[36]. The highest burst score corresponds to a topical tweet on August 30, 2020 related to CDC reporting of COVID-19 death statistics. This is noteworthy, given the sustained posting of tweets downplaying COVID-19 severity by the Right and Far Right throughout the study period (Section "COVID-19 vaccines and disease severity content"), and motivates further examination of content associated with the August 30, 2020 flagged day in Fig. 6 (Bottom).

Example: inter-cluster spread event. The high burst score on August 30, 2020 (Fig. 6, Bottom) was driven by tweets claiming that the CDC had 'quietly updated' COVID-19 death statistics indicating that only 6% of the deaths previously categorized as being due to COVID-19 were actually caused by COVID-19, while the remaining 94% were caused by underlying conditions[37]. Figure 7 plots cumulative tweets on this topic from each cluster normalized by the number of communities in the cluster. Initial tweets occur in the Far Right, with early amplification within the Far Right corresponding to tweets linking to or retweeting the Twitter account of the Gateway Pundit (cyan dashed line). Other studies have identified the prominence of the Gateway Pundit on Twitter generally and among fake news domains specifically[38]. A subsequent amplification event occurs following a retweet by Donald Trump (black dashed line) of a known QAnon account on this topic[37]. The Right then begins to post on this topic. In this example, we thus observe a misinformation pathway that begins in the Far Right, includes accounts associated with conspiracy theorists, and subsequently spreads to more mainstream communities following amplification by an influential node.

Discussion

The structured, data-light approach we have taken to monitoring COVID-19 misinformation is, to our knowledge, distinct from other published misinformation monitoring approaches to date. Online communities within social media platforms vary widely in their posted content[39]. The presented framework for longitudinal monitoring of sentinel nodes selected from online communities can support early detection of content and narratives moving across communities, including noteworthy events such as content migration from a community with extremist tendencies into communities closer to the mainstream. This approach complements panel-based longitudinal studies that can be conducted at scale with regards to demographic information such as age, gender,

and party affiliation[38]. Here we use influential accounts within each community as sentinel nodes, as previous work has shown that COVID-19 online content generally[20] and COVID-19 misinformation specifically[16] is driven by a small set of influential accounts. We note that other criteria for sentinel node selection, such as a mixture of influential and randomly selected accounts, may also be desirable for detection and characterization of circulating content within communities. Principled selection of accounts translates into modest data requirements: we collected 4,130,909 tweets posted by 420 accounts over the six month period. By contrast Chen et. al.[12] pulled 764,613,007 COVID-19 related tweets over the same time frame. Although collecting more data may sometimes be preferable, doing so increases the costs of monitoring and makes monitoring for the spread of misinformation more complex.

The polarization of the U.S. electorate, the ramifications and connections with interactions and content on social media, and the politicization of the COVID-19 pandemic[40–42], have been extensively documented. Our finding that accounts sharing right-leaning domains more frequently posted COVID-19 misinformation compared with accounts sharing left-leaning domains is consistent with other studies[9,22,23,38,43], and demonstrates a correlation between sharing right-leaning media on Twitter and sustained posting of COVID-19 misinformation. This is of particular concern given the influence of these sentinel accounts for COVID-19 discourse on Twitter. While causal relationships between media exposure and polarization are actively debated[44,45], polarization has been shown to impact compliance with recommended NPIs[46–48], and polling data suggests that it will be detrimental to mass vaccination efforts[6,49].

Translation of misinformation exposure to public health impact depends upon many factors, including extent of misinformation penetration, specific misinformation content, and demographic characteristics and social environment of the misinformation consumer[50]. Loomba et al[6] demonstrate in an experimental setting that recent exposure to COVID-19 misinformation can produce statistically significant decreases among individuals' intent to vaccinate, although both the nature of the false or misleading content and the demographics of the information consumer can amplify or dampen this effect. The measurable impact of misinformation on vaccination rates is consequential given the continued emergence of SARS-CoV-2 variants of concern. Misinformation thus presents a critical threat in the persuasion of the vaccine hesitant, who play a key role in the direction and duration of the pandemic.

Sensationalist narratives have been the focus of much discussion and do indeed present a potential risk to public health: among misinformation narratives presented to respondents, Loomba et al. find the claim that the COVID-19 vaccine would alter host DNA to be associated with the largest decrease in vaccination intent of the misinformation types considered[6]. Hotez et al[51] include 'genetically modified humans' in their primer for healthcare providers for correcting COVID-19 vaccine misinformation. We found that the genetic alteration, microchips, and Plandemic COVID-19 conspiracy theories were largely confined to Far Right communities, and did not garner widespread cross-cluster support. However, there was widespread penetration of misinformation downplaying COVID-19 severity, including cross-cluster propagation of content claiming manipulation of CDC death counts. These latter findings are consistent with survey results[50] and other Twitter studies[52] which indicate widespread misconceptions of COVID-19 severity both online and offline.

The prevalence of tweets downplaying COVID-19 severity across both the Right and Far Right clusters has public health implications. Romer and Jamieson[53] found that belief that the CDC exaggerated COVID-19 severity was associated with decreased vaccine willingness, and surveys released slightly after our observation period ended (February 2021) suggest that perceived COVID-19 severity is a key factor in an individual's decision to vaccinate[54,55]. Signaling from many accounts that COVID-19 is not a severe disease, despite reputable evidence to the contrary, could instill this position in Twitter users that consume our sentinels' content, especially if social media is their primary news source[3,22]. This may be particularly true if perceived vaccine risks outweigh the perceived risks from infection[31].

Our sentinel approach is flexible, and can be extended to other topics and platforms (given data availability). The basis of the method consists of identifying subsets of nodes in online communities, and using these nodes to examine content similarity over time between communities or clusters of communities. We have intentionally taken a simple approach to measuring content similarity, using cosine similarity of trigrams. This technique captures similarity driven by overlap in words (and thus, for example, captures retweet-driven similarity), but would not capture similarity of tweets using different words to express related content. Using approaches such as word[56–58] or tweet embeddings[59] to address this is an area for future work.

Importantly, this analysis addresses the broader case of misinformation, which we consider to be false or misleading content regardless of intent, as opposed to the subset of misinformation known as disinformation which refers to intentionally disseminated false or misleading content within a target group to advance an agenda or to cause harm. The described framework could be used in conjunction with emerging techniques in the detection of influence operations, such as those developed by Smith et al.[60], in order to explore the extent to which such actors drive meaningful narrative shifts across the social media ecosystem.

A natural question is what specific mechanisms underly high burst score days for inter-cluster content similarity. Our data suggest that highly active accounts can serve as bridges for inter-cluster spread, but more detailed examination is needed to assess whether this is a robust feature driving high burst score days. We note that the sentinel account approach is not designed for mapping content spread in detail, as the number of accounts selected is intentionally small. Analogous to epidemiological sentinel monitoring, the online sentinel accounts allow detection of presence and patterns in the data, but more detailed investigation (analogous to contact tracing for infectious diseases) is needed to deduce mechanisms of spread. Sentinel monitoring can identify content that is garnering attention in disparate communities, prompting more detailed follow-up investigation on specific content diffusion pathways and other potential drivers of content similarity. More detailed analysis can include considerations of primary and secondary spread of content, as well as movement of content across different platforms.

A limitation of this study is its restriction to Twitter. Each social media platform has its particular biases in user base and online functions, and study of consistencies and differences between platforms with regards to health information is important for misinformation monitoring and mitigation efforts. Existing COVID-19 related work includes analysis of pro-vaccination and anti-vaccination content on Facebook[61], popularity of YouTube COVID-19 misinformation and sharing of these videos via Facebook[62], and cross-platform comparisons of Twitter and Facebook[16]. Developing similar structured approaches to misinformation monitoring within and across additional platforms is an area for future work.

In conclusion, the work presented here illustrates the promise of structured approaches to monitoring online content. Principled approaches are needed for efficiently navigating the high volume, high noise to signal online environment. Network-based approaches such as that taken here can be particularly useful, given the fundamental role of networks in online content propagation. Sentinel accounts provide a way to identify content circulating within and between online communities. Many different approaches can be taken for selecting sentinel accounts besides that which we used here. Studying different selection criteria in the context of various topics and online platforms of interest are important areas for additional research.

Methods

Sentinel recruitment and data collection. **Community structure and sentinel node identification.** Queries for tweets containing the phrase 'covid' were performed on May 27 over a twelve hour period using Twitter's API and the `tweepy` Python library. Community detection was performed on the largest connected component of the retweet network, where the weight of the arc from j to i equals the number of times node j retweeted i. Specifically, we maximize the following version of modularity for weighted, directed graphs:

$$Q = \frac{1}{w}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(A_{ij} - \frac{w_i^{in}w_j^{out}}{w}\right)\delta(C_i, C_j), \tag{2}$$

where w is the sum of all edge weights in the network, w_k^{in} and w_k^{out} the weighted in-degree and out-degree, respectively, of node k, C_k the community assignment of node k, and δ corresponds to the Kronecker delta. Modularity maximization was performed using a GenLouvain method[25,63], implemented with symmetrization of the modularity matrix $A_{ij} - \frac{w_i^{in}w_j^{out}}{w}$.

Sentinel nodes were selected by first considering the fifty largest communities in the retweet network, and then selecting the 15 most highly retweeted nodes from each of the communities that consisted of predominantly English-speaking, domestic accounts. Whether a community was predominantly English-speaking was determined by taking a random sample of 100 tweets from that community and applying Google's language detection algorithm. Communities whose sample was at least 80% English were classified as English-speaking. A final inspection was done on the filtered sentinel communities and any community whose sentinel accounts were clearly not based in the United States was removed.

Stability of community structure. To assess stability of community structure in Twitter content regarding COVID-19, we assembled a second retweet network using the same phrase search ('covid') over a 24 hour period beginning June 8 and ending June 9, 2020. Modularity maximization was performed on this second retweet network, and then the community structures of the May 27 and June 8 networks were compared using the Rand score as described in[26]. Statistical significance of community similarity was assessed using the z-Rand score ([26], equation 2.1).

Longitudinal data collection and observation period

We collected tweets from sentinel nodes from July 1, 2020 through January 6, 2021 using Twitter's public-facing API. This observation period was split into two sets using 10/4/2020 at 12:00 AM (ET) as the demarcation point. The first portion of the data (from 7/1/2020 - 10/3/2020) was used to characterize sentinel communities according to their linked domains (Section "Sentinel community characterization using linked domains") and establish baseline similarity between clusters (Eq. (1) and Section "Flagging inter-cluster content spread").

Node attrition, active accounts, and active user days. Node attrition over the observation period may occur due to several possible reasons, including a user deleting their account or suspension by Twitter. Each cluster retained at least 80% of their initial sentinel nodes through mid-December. Comparison of attrition over time between clusters is given in the Supplementary Materials. In order to account for sentinel attrition we define the concept of an *active account* as well as the corresponding notion of *active account days*. We consider a sentinel account to be active on a given day if we observe a tweet from that account on or after that day. As an extension we define active account days to be the number of days a particular account is deemed active.

Sentinel community characterization using linked domains. We examined the links that each community shared in their tweets posted from 7/1/2020 - 10/3/2020. Specifically, we used principal components analysis (PCA)[64] to produce a scalar measure of linked domain preference, and then clustered sentinel communities according to this preference.

A community-level domain frequency matrix was formed by examining the unique domains posted more than 10 times by any given community. The i, j entry of this matrix was the fraction of links from community i to that linked to domain j, excluding links to `twitter.com`, domains from URL shortening services (e.g. `bit.ly`) and any domain shared less than 10 times across all communities. PCA was performed for this matrix and the output was cross-referenced with the media bias fact chart provided in[27].

Sentinel communities were projected onto the resulting first principal component, giving each community a linked domain score. Hierarchical clustering based upon the distance between cluster centroids[28] was performed

on the linked domain scores. Cut points in the dendrogram were compared using the corresponding silhouette scores[29].

COVID-19 misinformation by cluster. COVID-19 misinformation by topic and cluster was determined through human coding of a random sample of tweets that were posted 7/1/2020-10/3/2020. The considered topics were COVID-19 mortality, hydroxychloroquine, facemasks and Plandemic.

Relevant tweets were selected with substring searches. First, a subset of tweets pertaining to COVID-19 was selected by finding any tweet that contained at least one of "covid", "coronavirus", "sars-cov" or "pandemic". These COVID-19 tweets were further subset for each of the four topics below by searching for the strings: "plandemic" and "scamdemic" for Plandemic; "hcq", "hydrox" and "chloroq" for hydroxychloroquine; "mask" for facemasks; and "fatality rate", "death rate", "survival rate", "death numbers", "covid-19 death", "covid death", "covid19 death", "died from covid" and "died of covid" for covid mortality.

A random sample of 100 tweets from each cluster for each topic was selected for coding. If a cluster did not have 100 tweets on a particular topic, we took all tweets from that cluster on that topic. When possible, each community within a cluster was equally represented in that cluster's random sample, so that a single community did not disproportionately impact the coding results.

The coders consisted of four undergraduates from a Midwestern private university. Coders were given a reference sheet (see Supplementary Materials) and a collection of tweet texts with the time the tweet was posted. The time was provided so that coders could cross-reference with the scientific consensus about the topic at the time the tweet was posted. All tweets were presented through an untimed Qualtrics survey.

COVID-19 vaccines and disease severity content. Tweets pertaining to COVID-19 severity and vaccination were identified through a substring search of COVID-19-related tweets (those found to contain any of the the strings "covid", "coronavirus", "sars-cov" or "pandemic"). Search strings for COVID-19 severity were seeded by reading the COVID-19 tweets and searching for phrases related to the morbidity and mortality of COVID-19, for example those that mention disease incidence or prevalence (e.g. 'case spike' or 'confirmed cases') as well as death count statistics (e.g. 'death count'). For vaccination we used 'vaccine', the stem 'vaccinat' as well as the names of the vaccine manufacturers whose vaccines received approval from the U.S. Food and Drug Administration. Exact substring lists can be found in the Supplementary Materials.

To identify which topical tweets may be related to misinformation surrounding severity or vaccinations we constructed additional substring lists related to downplaying COVID-19 severity, vaccine hesitancy and vaccine misinformation which were used to filter severity and vaccination tweets respectively. Complete lists can be found in the Supplementary Materials.

Flagging inter-cluster content spread. Inter-cluster content similarity was measured using a cosine similarity score[65] applied to the trigrams generated by each community's tweets. Specifically, the 'documents' used to generate the similarity score were all COVID-19 tweets (Section "COVID-19 misinformation by cluster") sent by a particular community on a given day. These tweets were cleaned to remove stopwords, urls and mentions of Twitter users, and then trigram frequency vectors were generated using the NLTK package in Python[66]. We define the daily similarity between clusters A and B as the arithmetic mean of the similarity scores between distinct pairs of communities in A and B.

A day was flagged as containing viral content if the similarity between two clusters was anomalously large according to the 'burst-score' given in Eq. (1). This metric is an adaptation of the burst-score introduced by[33] for inter-cluster similarities. Specifically, (1) measures between similarity on day t in terms of standard deviations from the historic mean. Implicit in this measure is the assumption that the similarity between clusters A and B is stationary in time and does not exhibit a trend. Validity of this stationarity assumption was assessed with an augmented Dickey-Fuller test[67], performed with the `adfuller` model in the statsmodels[68] Python package using a 'constant only' model with 0 lag (see Supplementary Materials).

Latent semantic analysis (LSA)[34] was used to identify topical tweets that drove high similarity on days in which at least one pair of clusters was flagged. LSA was separately applied to the tweets posted by both flagged clusters, where documents corresponded to individual tweets and terms corresponded to the tweet trigrams. The singular document vectors associated with the five largest singular values were used to identify topical tweets from each cluster on the flagged day. Specifically, a sharp drop in the magnitude of the document vector components was identified and those tweets with component magnitudes above the drop were selected as topical tweets. Topical tweets common to both clusters were removed and the between cluster similarity was recalculated. Removed tweets were considered to be drivers of the burst on that day if the recalculated burst score was lower than the flagging criterion.

Data availability

Tweet IDs and user IDs corresponding to sentinel nodes that were verified accounts as of July 22, 2020 are available at: https://github.com/joetien/sentinel-node-misinfo.

Received: 1 December 2021; Accepted: 20 April 2022
Published online: 14 June 2022

References

1. Brennen, J. S., Simon, F., Howard, P. N. & Nielsen, R. K. Types, sources, and claims of COVID-19 misinformation. *Reuters Institute* **7**, 3–1 (2020).

2. Gallotti, R., Valle, F., Castaldo, N., Sacco, P. & De Domenico, M. Assessing the risks of 'infodemics' in response to COVID-19 epidemics. *Nat. Hum. Behav.* **4**(12), 1285–1293 (2020).
3. Bridgman, A. *et al.* The causes and consequences of COVID-19 misperceptions: Understanding the role of news and social media. *Harvard Kennedy School Misinf. Rev.* https://doi.org/10.37016/mr-2020-028 (2020).
4. Ash, E., Galletta, S., Hangartner, D., Margalit, Y. & Pinna, M. The effect of Fox News on health behavior during COVID-19." Available at SSRN: https://ssrn.com/abstract=3636762, (2020).
5. Murthy, V. H. Confronting health misinformation: the U.S. Surgeon General's advisory on building a healthy information environment, tech. rep., United States Department of Health and Human Services, (2021). https://www.hhs.gov/sites/default/files/surgeon-general-misinformation-advisory.pdf.
6. Loomba, S., de Figueiredo, A., Piatek, S. J., de Graaf, K. & Larson, H. J. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nat. Hum. Behav.* **5**(3), 337–348 (2021).
7. Bubar, K. M. *et al.* Model-informed COVID-19 vaccine prioritization strategies by age and serostatus. *Science* **371**(6532), 916–921 (2021).
8. Allcott, H. & Gentzkow, M. Social media and fake news in the 2016 election. *J. Econ. Perspect.* **31**(2), 211–36 (2017).
9. Evanega, S., Lynas, M., Adams, J., Smolenyak, K. & Cision Global Insights, Coronavirus misinformation: quantifying sources and themes in the COVID-19 'infodemic', *JMIR Preprints*, (2020).
10. Lamsal, R. Design and analysis of a large-scale COVID-19 tweets dataset. *Appl. Intell.* **51**, 2790–2804 (2021).
11. Qazi, U., Imran, M. & Ofli, F. Geocov19: a dataset of hundreds of millions of multilingual COVID-19 tweets with location information. *SIGSPATIAL Special* **12**(1), 6–15 (2020).
12. Chen, E., Lerman, K. & Ferrara, E. Tracking social media discourse about the COVID-19 pandemic: development of a public coronavirus Twitter data set. *JMIR Public Health Surveill.* **6**(2), e19273 (2020).
13. DeVerna, M. R., Pierri, F., Truong, B., Bollenbacher, J., Axelrod, D., Loynes, N., Torres-Lugo, C., Yang, K., Menczer, F. & Bryden, J. CoVaxxy: A collection of English Twitter posts about COVID-19 vaccines., *Proceedings of the 15th International Conference on Web and Social Media*, (2021).
14. Muric, G., Wu, Y. & Ferrara, E. COVID-19 vaccine hesitancy on social media: building a public Twitter dataset of anti-vaccine content, vaccine misinformation, and conspiracies. *JMIR Public Health and Surveill.* **7**(11), e30642 (2021).
15. Cinelli, M. *et al.* The COVID-19 social media infodemic. *Sci. Rep.* **10**(1), 16598 (2020).
16. Yang, K.-C. *et al.* The COVID-19 infodemic: Twitter versus Facebook. *Big Data Soc.* **8**(1), 20539517211013860 (2021).
17. Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W. & Starnini, M. The echo chamber effect on social media, *Proceedings of the National Academy of Sciences*, vol. 118, no. 9, (2021).
18. Flaxman, S., Goel, S. & Rao, J. M. Filter bubbles, echo chambers, and online news consumption. *Public Opin. Q.* **80**, 298–320 (2016).
19. Declich, S. & Carter, A. O. Public health surveillance: historical origins, methods and evaluation. *Bull. World Health Organ.* **72**(2), 285 (1994).
20. Gallagher, R. J., Doroshenko, L., Shugars, S., Lazer, D. & Welles, B. F. Sustained online amplification of COVID-19 elites in the United States. *Soc. Media Soc.* **7**(2), 20563051211024956 (2021).
21. Tien, J. H., Eisenberg, M. C., Cherng, S. T. & Porter, M. A. Online reactions to the 2017 'Unite the Right' rally in Charlottesville: Measuring polarization in Twitter networks using media followership. *Appl. Network Sci.* **5**(1), 1–27 (2020).
22. Jamieson, K. H. & Albarracin, D. The relation between media consumption and misinformation at the outset of the SARS-CoV-2 pandemic in the US, *The Harvard Kennedy School Misinformation Review*, vol. 1, no. 2, (2020).
23. Motta, M., Stecula, D. & Farhart, C. How right-leaning media coverage of COVID-19 facilitated the spread of misinformation in the early stages of the pandemic in the US. *Can. J. Political Sci.* **53**(2), 335–342 (2020).
24. Tufekci, Z. Big questions for social media big data: representativeness, validity and other methodological pitfalls, in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, (2014).
25. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech: Theory Exp.* **2008**(10), P10008 (2008).
26. Traud, A. L., Kelsic, E. D., Mucha, P. J. & Porter, M. A. Comparing community structure to characteristics in online collegiate social networks. *SIAM Rev.* **53**(3), 526–543 (2011).
27. "Media Bias / Fact Check." https://mediabiasfactcheck.com, (2021).
28. Müllner, D. fastcluster: Fast hierarchical, agglomerative clustering routines for R and Python. *J. Stat. Softw.* **53**(9), 1–18 (2013).
29. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).
30. Krippendorff, K. *Content Analysis: An Introduction to Its Methodology* (Sage, Thousand Oaks, 2013).
31. Bauch, C. T. & Earn, D. J. Vaccination and the theory of games. *Proc. Natl. Acad. Sci.* **101**(36), 13391–13394 (2004).
32. Kirzinger, A., Sparks, G., Hamel, L., Lopes, L., Kearney, A., Stokes, M. & Brodie, M. KFF COVID-19 Vaccine Monitor: July 2021, tech. rep., Kaiser Family Foundation. https://www.kff.org/coronavirus-covid-19/poll-finding/kff-covid-19-vaccine-monitor-july-2021/?utm_campaign=KFF-2021-polling-surveys, (2021).
33. Mehrotra, R., Sanner, S., Buntine, W. & Xie, L. Improving LDA topic models for microblogs via tweet pooling and automatic labeling, in *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 889–892, (2013).
34. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. & Harshman, R. Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci.* **41**(6), 391–407 (1990).
35. Lemire, J. & Knickmeyer, E. What we know, and what we don't, about Trump's diagnosis. AP News. https://apnews.com/article/virus-outbreak-donald-trump-amy-coney-barrett-hope-hicks-archive-1d48efc6d80a91430d060106749aca07, (2020).
36. Miller, Z., Colvin, J. & Madhani, A. Trump, still infectious, back at White House — without mask. AP News. https://apnews.com/article/virus-outbreak-donald-trump-archive-d39bd670e8a280b6283abcdfc91d4794, (2020).
37. Dupuy, B. CDC has not reduced the death count related to COVID-19. AP News. https://apnews.com/article/virus-outbreak-us-news-health-technology-142575f3ba64158dd0b24a8e2fb23579, (2020).
38. Lazer, D., Ruck, D. J., Quintana, A., Shugars, S., Joseph, K., Grinberg, N., Gallagher, R. J., Horgan, L., Gitomer, A., Bajak, A., Baum, M. A., Ognyanova, K., Qu, H., Hobbs, W. R., McCabe, S. & Green, J. The COVID States Project #18: fake news on Twitter, *OSF Preprints*, (2020), https://doi.org/10.31219/osf.io/vzb9t
39. Singh, L. *et al.* Understanding high- and low-quality url sharing on COVID-19 Twitter streams. *J. Comput. Soc. Sci.* **3**(2), 343–366 (2020).
40. Allcott, H. *et al.* Polarization and public health: Partisan differences in social distancing during the coronavirus pandemic. *J. Public Econ.* **191**, 104254 (2020).
41. Gollwitzer, A. *et al.* Partisan differences in physical distancing are linked to health outcomes during the COVID-19 pandemic. *Nat. Hum. Behav.* **4**(11), 1186–1197 (2020).
42. Green, J., Edgerton, J., Naftel, D., Shoub, K. & Cranmer, S. J. Elusive consensus: Polarization in elite communication on the COVID-19 pandemic. *Sci. Adv.* **6**(28), eabc2717 (2020).
43. Garrett, R. K. & Bond, R. M. Conservatives susceptibility to political misperceptions. *Sci. Adv.* **7**(23), eabf1234 (2021).

44. Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, Chen, H., Fallin Hunzaker, M. B., Lee, J., Mann, M., Merhout, F. & Volfovsky, A. Exposure to opposing views on social media can increase political polarization, *Proceedings of the National Academy of Sciences*, vol. https://doi.org/10.1073/pnas.1804840115(2018).
45. Wojcieszak, M., de Leeuw, S., Menchen-Trevino, E., Lee, S., Huang-Isherwood, K. M. & Weeks, B. No polarization from partisan news: Over-time evidence from trace data, *Int. J. Press/Politics*, pp. 1–26, (2021).
46. Engle, S., Stromme, J., & Zhou, A. Staying at home: Mobility effects of COVID-19. Available at SSRN. https://ssrn.com/abstract=3565703, (2020).
47. Grossman, G., Kim, S., Rexer, J. M. & Thirumurthy, H. Political partisanship influences behavioral responses to governors' recommendations for COVID-19 prevention in the United States. *Proc. Natl. Acad. Sci.* **117**(39), 24144–24153 (2020).
48. Painter, M. & Qiu, T. Political beliefs affect compliance with government mandates. *J. Econ. Behav. Organiz.* **185**, 688–701 (2021).
49. SteelFisher, G. K., Blendon, R. J. & Caporello, H. An uncertain public-encouraging acceptance of COVID-19 vaccines. *N. Engl. J. Med.* **384**, 1483–1487 (2021).
50. Druckman, J. N. *et al.* The role of race, religion, and partisanship in misperceptions about COVID-19. *Group Process. Intergroup Relations* **24**(4), 638–657 (2021).
51. Hotez, P. *et al.* Correcting COVID-19 vaccine misinformation. *EClinicalMed.* **33**, 100780 (2021).
52. Jamison, A. M., Broniatowski, D. A., Dredze, M., Sangraula, A., Smith, M. C. & Quinn, S. C. Not just conspiracy theories: Vaccine opponents and proponents add to the COVID-19 'infodemic' on Twitter, *Harvard Kennedy School Misinformation Review*, vol. 1, no. 3, (2020).
53. Romer, D. & Jamieson, K. H. Conspiracy theories as barriers to controlling the spread of COVID-19 in the U.S. *Soc. Sci. Med.* **263**, 113356 (2020).
54. Nguyen, K. H. *et al.* COVID-19 vaccination intent, perceptions, and reasons for not vaccinating among groups prioritized for early vaccination – United States, September and December 2020. *Am. J. Transplant.* **21**(4), 1650–1656 (2021).
55. Ruiz, J. B. & Bell, R. A. Predictors of intention to vaccinate against COVID-19: Results of a nationwide survey. *Vaccine* **39**(7), 1080–1086 (2021).
56. Pennington, J., Socher, R. & Manning, C. D. Glove: Global vectors for word representation, In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, (2014).
57. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient Estimation of Word Representations in Vector Space, in *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings* (Y. Bengio and Y. LeCun, eds.), (2013).
58. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding, in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, (2019).
59. Vosoughi, S., Vijayaraghavan, P. & Roy, D. Tweet2vec: Learning tweet embeddings using character-level CNN-LSTM encoder-decoder, in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pp. 1041–1044, (2016).
60. Smith, S. T., Kao, E. K., Mackin, E. D., Shah, D. C., Simek, O., & Rubin, D. B. Automatic detection of influential actors in disinformation networks, *Proceedings of the National Academy of Sciences*, vol. 118, no. 4, (2021).
61. Sear, R. F. *et al.* Quantifying COVID-19 content in the online health opinion war using machine learning. *IEEE Access* **8**, 91886–91893 (2020).
62. Knuutila, A., Herasimenka, A., Au, H., Bright J., Nielsen, R. & Howard, P. N. COVID-related misinformation on YouTube, tech. rep., Oxford Internet Institute, (2020).
63. Jeub, L. G. S., Bazzi, M., Jutla, I. S. & Mucha, P. J. A generalized Louvain method for community detection implemented in MATLAB. http://netwiki.amath.unc.edu/GenLouvain, 2011–2016. Version 2.0.
64. Tipping, M. E. & Bishop, C. M. Mixtures of probabilistic principal component analyzers. *Neural Comput.* **11**(2), 443–482 (1999).
65. Manning, C. & Schutze, H. *Foundations of Statistical Natural Language Processing* (MIT Press, 1999).
66. Bird, S., Klein, E. & Loper, E. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit* (Springer, 2009).
67. Dickey, D. A. & Fuller, W. A. Distribution of the estimators for autoregressive time series with a unit root. *J. Am. Stat. Assoc.* **74**(366a), 427–431 (1979).
68. Seabold, S. & Perktold, J. "statsmodels: Econometric and statistical modeling with Python," in *9th Python in Science Conference*, (2010).

Acknowledgements

The authors would like to thank Rod Abhari and David King for helpful discussions. This work was supported by the Erdős Institute and the Office of Research at the Ohio State University.

Author contributions

M.O. and J.T. collected the data. M.O. conducted network analysis and analysis of tweet content. E.N. led coding of tweets for misinformation and associated statistical analysis. R.B., E.N., S.M., M.O., and J.T. designed the study. M.O. created the figures. All authors wrote and reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-12450-8.

Correspondence and requests for materials should be addressed to J.H.T.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.