Hindawi
Human Behavior and Emerging Technologies
Volume 2023, Article ID 9094933, 14 pages
https://doi.org/10.1155/2023/9094933

Research Article

The Evolution of Polarization in Online Conversation: Twitter Users' Opinions about the COVID-19 Pandemic Become More Politicized over Time

Weize Zhao, Lukasz Walasek, and Gordon D. A. Brown

Department of Psychology, University of Warwick, Coventry CV4 7AL, UK

Correspondence should be addressed to Weize Zhao; w.zhao.6@warwick.ac.uk

Received 24 October 2022; Revised 26 April 2023; Accepted 9 June 2023; Published 5 July 2023

Academic Editor: Zheng Yan

Political polarization on social media has been extensively studied. However, most research has examined polarization about topics that have preexisting associations with ideology, while few studies have tracked the onset of polarization about novel topics or the evolution of polarization over a prolonged period. The occurrence of COVID-19 provides a unique opportunity to study whether social media discourse about a novel event becomes increasingly polarized along ideological lines over time. This paper analyzes trends in Twitter polarization in relation to COVID-19 and other geopolitical events of 2020. The first two studies use topic analysis to examine the evolving difference over time in discussions of COVID-19 and other topics by liberals and conservatives on social media. COVID-19-related polarization is initially absent but gradually increases over time, in contrast to polarization related to other events. A third study examines structural polarization in retweet networks and finds that the frequency of counterpartisan retweets reduces over time. Across all three studies, we find evidence that Twitter discussion of COVID-19 has become more polarized over time.

1. Introduction

There is growing concern about the extent of political polarization on online media such as Twitter, as selective attention to social media sources may lead to reinforcement of preexisting beliefs (as with echo chamber effects, e.g., [1, 2]). People's opinions may become polarized along partisan lines, which may in turn lead to the expression of more extreme opinions [3] and to other social problems such as the spread of misinformation [4] and distrust in society [5]. Although the extent and consequences of such polarization on social media remain debated [6, 7], it is of both theoretical and practical importance to understand the development of polarization over time.

Natural language processing (NLP) approaches have been widely used to study political polarization: Németh [8] identifies over 150 studies published since 2010. Previous studies have quantified polarization on social media in a number of ways. Several researchers have trained machine learning models to classify messages into different groups (liberal or conservative, positive or negative in sentiment, etc.) based on words and hashtags [9–12]. Polarization can then be quantified with "identification scores" (as we do in the present paper). For example, higher polarization exists when it is increasingly possible to identify a person's political partisanship from the language he/she uses in tweeting about a particular topic. Sentiment analysis has also been used to quantify polarization [13–15]. Another strand of research focuses on structural isolation and polarization in communication networks (e.g., [16–18]), analyzing interactions between users including their following (accounts one chooses to pay close attention to), retweeting (spreading of messages sent by someone else), and mentioning (using @+username in messages).

Most previous research into the development of polarization on social media, although extensive, is limited in important ways. First, such research has typically examined transient (e.g., [17, 19–21]) or short-term (e.g., [10, 22–24]) trends in polarization rather than tracking long-term changes. Second, studies have generally focused on polarization in

discussions of long-discussed and highly politicized topics such as gun control [25] and presidential elections [26]. Many new events related to these topics (such as political demonstrations and presidential elections) lead to extensive discussion on Twitter and elsewhere. However, such events typically relate to well-formed and preexisting individual belief systems and social network structures and hence may be associated with polarized discussion from the outset. Analyses of Twitter discussions following such events therefore cannot inform understanding about the development, over time, of polarization in discussions of novel topics that may not tap directly into preformed belief systems.

To understand the development of polarization on social media, it is therefore essential to examine (a) the extent to which polarization is present at the time a new topic enters the social media discourse and (b) how polarization subsequently develops. This paper provides such an examination.

Understanding the development of polarization on social media, over lengthy periods of time and for novel and initially unpoliticized topics, is important for theoretical as well as practical reasons. Many researchers have used agent-based models to simulate the dynamic process of opinion formation and expression and have shown how significant polarization can emerge within an initially unpolarized system (e.g., [27, 28]). For example, social judgment theory [29, 30] and a family of bounded confidence models [27, 28, 31–34] assume that individuals are only influenced by other opinions if those opinions lie within a threshold distance of their own [35]. Some models additionally assume that exposure to opposing voices will solidify rather than weaken prior opinions [36–38]. The spiral of silence theory [39, 40] suggests that people are more likely to keep silent when their opinions are in the minority and explains why diversity of information disappears in an isolated group or "echo chamber." There are also model-based simulations of polarization of multidimensional opinions [41], complex interaction networks [42], and agents with psychologically plausible characteristics [3].

Agent-based models of polarization typically assume that the environment is initially unpolarized: Individuals' opinions (represented by a continuous value) are uniformly distributed at the outset, but this distribution eventually comes to follow a bimodal or multimodal pattern [27]. Field data from social media are also widely used by researchers to test their models (e.g., [34, 43]). However, as noted above, in many real-world contexts, the relevant social environment is already polarized. Most field evidence from social media studies the *outcomes* of polarization processes—situations in which the opinion dynamics may have already become stable. Although many studies have measured the intensity of polarization on social media, relatively few have tracked changes in intensity over time (e.g., [23]), especially over a long period. Thus, the assumption (made by many models) that significant polarization can emerge from an initially unpolarized environment remains underinvestigated.

The outbreak of COVID-19 in 2020 provides an opportunity to test the growth of polarization concerning a novel topic for which there may be relatively weak preexisting associations between political attitudes and the new topic, as COVID-19 was unknown by almost everybody at the beginning of our data collection period (start of 2020). Here, we examine how polarization levels respond to different types of geopolitical events. In Study 1 and Study 2, we compare the development of polarization in discussions of COVID-19-related topics to the development of polarization regarding two other significant geopolitical events in the same year (BLM activity and the 2020 U.S. Presidential election). Although discussion of all three topics reflects ideology to some degree, a unique feature of COVID-19 discussion is its novelty. BLM activity started in 2013, in contrast, and racism has been a widely discussed social problem in the U.S. for many decades. Similarly, the presidential election in the U.S. occurs every four years, and arguments between Democratic and Republican supporters are long-standing. The aim of this paper is to answer the following research questions:

R1: Are social media users' attitudes toward both the novel topic of COVID-19 and long-existing topics of BLM activity/presidential election significantly polarized?

R2: Does the degree of polarization for each of the three topics increase over time?

R3: Are increases in polarization greater for a novel topic (COVID-19) that is not initially associated with political ideology?

Before moving on to our methodologies for measuring polarization, we clarify how we use the term "polarization" because it is used in several ways in the research literature [44]. Polarization as we use the term in this paper describes the difference in people's behaviors as a function of their membership of one of two partisan groups (Democratic and Republican supporters). According to this definition, if polarization is insignificant, we will not be able to identify a person as Democratic or Republican through his/her behavior on social media. Across our three studies, we quantify this behavioral difference in two ways. The first measures the tendency for Democratic and Republican supporters to use different language when discussing the same topics in approximately 30 M tweets from one dataset (referred to as dataset A) sent between February 2020 and January 2021. Topics are identified using latent Dirichlet allocation topic modeling [45]. We examined both long-term (throughout 2020 in Study 1) and short-term (100 days in Study 2) changes in polarization in discussions about COVID-19 and other topics. Our second method, in Study 3, examines structural network polarization, defined as the tendency for Democratic and Republican supporters to receive information from, and propagate it to, different sources using a different dataset (B).

2. Study 1: Long-Term Polarization

Our first two studies focus on polarization in tweets' content. The aim of these two studies is to answer three specific questions. The first is simply whether or not evidence for polarization is available in our dataset. The second question, assuming that polarization does indeed occur, is whether the average levels of polarization differ over the three

important political topics in 2020 (COVID-19, BLM activity, and Presidential election). The third question, central to the present investigation, is whether the level of polarization changes over time for each topic.

2.1. Data. Our data are publicly available tweets posted by accounts categorized as conservative (i.e., Republican) and liberal (i.e., Democrat) from 1 Feb 2020 to 17 Jan 2021. We used the Twitter API to collect tweets sent by the followers (mainly from the U.S.) of a set of seed accounts. Seed accounts are well-known Twitter accounts owned by sources with clearly identifiable partisanship, including media (such as @FoxNews), politicians (@JoeBiden), and organizations (@GOP). Table S1 in the supplemental material lists the seed accounts that were used in the present study. We selected 35,000 accounts that primarily follow Democrats and 35,000 accounts that primarily follow Republicans using the following criteria: First, accounts were included only if they follow at least three seed accounts associated with one party and fewer than three accounts associated with the other party. We used the tweetsbotornot2 package in R to identify bots and removed all accounts with an estimated probability (of being bots) > 0.5. We also removed accounts created later than 1 Feb 2020, non-English language accounts, users who did not send any tweets from 1 Feb 2020 to 17 Jan 2021, protected or private accounts, and highly active users (those who sent more than 3200 tweets from 1 Feb 2020 to 17 Jan 2021). The Twitter API only allows collection of the most recent 3200 tweets sent by each Twitter user. Therefore, if we include those highly active users who sent many thousands of tweets in a short period, we can only collect their tweets sent during the last few months of our dataset, and our dataset will not reflect Twitter users' real activity level at different periods as the activity level during the last few months will be overestimated. This is also the reason why we removed Twitter accounts created later than 1 Feb 2020. Any resulting bias is manageable as those highly active users comprise only approximately 1% of accounts in our dataset.

Figure 1 shows the number of seed accounts followed by users in our dataset. A clear partisan division is evident: 63.5% of Democratic users and 70.9% of Republican users follow no accounts in the opposite group, and 22.7% of Democratic users and 35.3% of Republican users follow over six accounts in their own group and no accounts in the opposite group. We then collected up to 3200 most recent tweets from each user but only included tweets sent between 1 Feb 2020 and 17 Jan 2021. The result is a dataset with 30 M tweets, including both original tweets and retweets.

2.2. Methodology

2.2.1. Associating Tweets with Topics. We used the Mallet latent Dirichlet allocation (LDA) [46] topic modeling tool to identify a topic for each tweet. To train the LDA model and reduce memory load, we preprocessed and tokenized tweets as follows (tokenization is the process of dividing a large quantity of text into smaller pieces called tokens): All text was lowercased, and all URLs, hashtags, punctuations, emoticons, and numbers were removed. We removed short messages because a message's polarization level is calculated using an average score for each word used in the message (see below), and so if the number of words is small, the measurement will be unreliable. All tweets with fewer than five words were removed. Stopwords (such as "the" and "a," selected from the SMART (System for the Mechanical Analysis and Retrieval of Text) information retrieval system) were also removed, and the remaining words (except nouns) were lemmatized (i.e., inflected forms of words like "said" were changed to the relevant base form, like "say"). Common bigrams and trigrams (those that appeared over 500 times in our corpus) were turned into single words (e.g., "Donald Trump" became "donaldtrump" and "Black lives matter" became "blacklivesmatter"). This step is needed because two words can represent different concepts when they appear in pairs—e.g., the words "white" and "house" are typically uncorrelated with political issues when viewed in isolation but not when taken in combination. Rare words (defined as those appearing fewer than 20 times in our corpus of 30 M tweets) were removed, and finally, we removed tweets that contained only one word after the preceding steps were completed.

After the preprocessing, 29,501,806 tweets remained in the corpus. These provided the data for the topic modeling, which we now describe. LDA assumes that the probability of a word appearing in a document, $p(\text{word}|\text{document})$, is equal to $\sum_{\text{topic}=1}^{T} p(\text{word}|\text{topic})p(\text{topic}|\text{document})$, where T is the number of topics. Here, the documents are tweets. This function can be represented in matrix form as a $D \times W$ document-word matrix which is equal to the product of a $D \times T$ document-topic matrix and a $T \times W$ topic-word matrix, where D and W are the numbers of documents and words in the corpus, respectively. To choose the optimal number of topics, we used 10% of the tweets in our dataset to train LDA models with T ranging from 10 to 70. Figure S1 shows the coherence score (a measurement of classification performance based on the semantic similarity between representative words in each topic) for each number of topics. We selected 35 topics (where the coherence score converged to 0.55). Although 35 is not the number of topics that gives the highest coherence score, choosing a larger number only produces a small increase in coherence but reduces the number of tweets available for each topic in the next step of the analysis. The result of running Mallet LDA on the whole dataset was a $29{,}501{,}806 \times 35$ document-topic matrix with each element, $p(\text{topic}|\text{document})$, indicating the estimated probability of each document, i.e., tweet, belonging to the corresponding topic. Figure S2 in the supplementary material visualizes the topic modeling outcomes for 2 of 35 topics.

The next step removed tweets that could not clearly be assigned to particular topics (to reduce noise caused by multi-topic tweets). We searched for the highest (mean 0.730, median 0.743) and second-highest scores (mean 0.142, median 0.079) values in the document-topic matrix for each tweet. Next, we calculated the ratio of these two scores and filtered out the 20% tweets with the lowest ratios. This procedure had the effect of preserving only tweets with clear topic assignments. The final corpus contains 23,602,737 tweets.

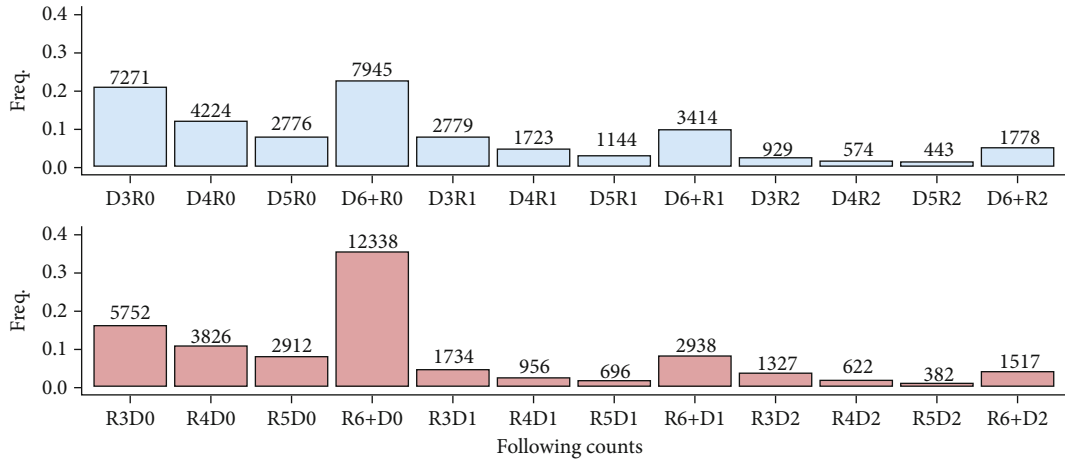

Figure 1: Number of Democrat (blue/lighter bars) and Republican (red/darker bars) users in our dataset who follow different numbers of seed accounts in Table S1. The subscript AnBm indicates users who follow n seed accounts from group A and m seed accounts from group B. For example, the top left bar indicates that there were 7271 Democrats who followed three Democrats and no Republicans.

Next, we used the topic-word matrix to find the most representative tokens for each topic. Each element in the matrix, i.e., each conditional probability $p(\text{word}|\text{topic})$, can be interpreted as the representativeness of the token for the corresponding topic. Table S2 shows the five most representative tokens for each topic. 35 topics include issues related to politics, such as the 2020 U.S. election, diplomacy, and national security, and nonpolitical issues, like sports, religion, showbiz, and cooking. Detailed topic modeling results are presented in the supplemental material (available here).

2.2.2. Measuring Polarization. We measured polarization by determining the probability with which a Twitter user could be correctly classified as a Republican or Democrat by the content of his/her tweets. Higher classification probabilities correspond to increased polarization.

Specifically, we used the leave-out (LO) estimator introduced by Gentzkow et al. [47]. This algorithm defines the partisanship of a user as the posterior probability of classifying that user as a member of the corresponding group on the basis of his/her tweets. Tokens from the users' tweets are treated as samples for updating the posterior probabilities. The polarization level associated with tweets about each topic is then defined as the mean of the posterior probabilities for all users, measured by the following function:

$$\begin{aligned}
\Pi_{T,t} &= \frac{1}{2}\frac{1}{\left|R_{T,t}\right|}\sum_{i\in R_{T,t}}\widehat{\boldsymbol{q}}_{i,T,t}\widehat{\boldsymbol{\rho}}_{-i,T,t} + \frac{1}{2}\frac{1}{\left|D_{T,t}\right|}\sum_{i\in D_{T,t}}\widehat{\boldsymbol{q}}_{i,T,t}\left(1-\widehat{\boldsymbol{\rho}}_{-i,T,t}\right),\\
\widehat{\boldsymbol{q}}_{i,T,t} &= \frac{\boldsymbol{c}_{i,T,t}}{m_{i,T,t}},\\
\widehat{\boldsymbol{\rho}}_{-i,T,t} &= \frac{\widehat{\boldsymbol{q}}^{R}_{-i,T,t}}{\widehat{\boldsymbol{q}}^{R}_{-i,T,t}+\widehat{\boldsymbol{q}}^{D}_{-i,T,t}},
\end{aligned} \tag{1}$$

where $\Pi_{T,t}$ is the polarization level for topic T at week t; $D_{T,t}$ and $R_{T,t}$ are the sets of Democratic and Republican users, respectively; $\boldsymbol{c}_{i,T,t}$ is a vector for token counts for user i; and $m_{i,T,t}$ is the total number of tokens used by user i for topic T at week t, so $\widehat{\boldsymbol{q}}_{i,T,t}$ in the second equation represents a vector for token frequency for user i. We removed short text in the preprocessing steps to ensure that each user in our dataset sent a certain number of tokens. $\widehat{\boldsymbol{q}}^{P}_{-i,T,t}$ is the frequency vector for all tokens given by all members of a group $P(P\in[R,D])$ except the user i. Therefore, $\widehat{\boldsymbol{\rho}}_{-i,T,t}$ in the third equation is a vector in which each element measures the probability (calculated from a corpus without text from user i) that the token (used by at least two users) is sent by a Republican user. $\widehat{\boldsymbol{q}}_{i,T,t}\widehat{\boldsymbol{\rho}}_{-i,T,t}$ in the first equation measures the probability of assigning user i to the Republican group based on tokens used by that individual, and the first/second component in the first equation represents the average probability of correctly attaching each Republican (Democratic) user to the Republican (Democratic) group (i.e., identifiability). Overall, $\Pi_{T,t}$ is the average identifiability of the two groups. If $\Pi_{T,t}$ is close to 0.5, messages sent by the two groups are similar, and the classifier performs no better than flipping a coin, while if $\Pi_{T,t}$ is significantly higher than 0.5, messages sent by Democratic and Republican users are significantly different, which means that the topic is polarized. Note that the key feature of the LO estimator algorithm is it measures a user's identifiability using a corpus that excludes messages from the user him/herself. This step removes the influence of tokens used only by a single user, which is important for an unbiased estimation. Imagine that there is a user from the Republican group whose message contains a unique token, like *aaa*; then if we use the full corpus including user i's message to estimate his/her identifiability, we will overestimate the value as the token *aaa* will be misdiagnosed as a very representative token for Republicans (because it only appears in the Republican corpus).

2.3. Results

2.3.1. Robustness of Polarization Measure. First, for robustness, we confirm that the LO estimator methodology we

used can quantify the intensity of polarization in our dataset. Figure S3 in the supplementary material shows that if we assign users to random rather than partisan groups, then the LO estimator is close to 0.5. In addition, the LO estimators for most nonpolitical topics are lower than for political topics, as shown in Figure S4 and Figure S5. Therefore, a high LO estimator value reflects a significant difference in language usage and hence the polarization intensity, between supporters from both parties. We conclude that the LO estimator is appropriate for measuring polarization in our dataset.

2.3.2. Extent of Polarization. Our first research question was whether polarization exists and for which topics. Figure 2 shows polarization (the LO estimator values) over 50 weeks for tweets for the 6 of the 35 topics that were most clearly related to COVID-19 (Figures 2(a)–2(c)), BLM (Figures 2(d) and 2(e)), or the 2020 election (Figure 2(f)). The supplemental material includes analytical results for the remaining 29 topics. The bar charts in Figure 2 show the number of tweets users sent at the corresponding time period.

We find high polarization in discussion of all six topics: Leave-out estimators for both tweet sets (with and without retweets) were significantly higher than 0.5 for most weeks. The two values in the parentheses following representative tokens for each panel in Figure 2 (e.g., 0.573 and 0.524 for Figure 2(a)) give the average LO estimator over 50 weeks for each topic. Figures 2(a)–2(c) show the time course of polarization in discussions of three topics related to the COVID-19 pandemic. The volume of discussions reached a peak for each topic in the middle of March; COVID-19 was a relatively new concept for most Twitter users at that time.

To identify typical tokens used by Republican and Democrat accounts, we calculated the odds ratios of particular tokens (in each week) being used by Democrats (vs. Republicans). The log odds ratio $O_{i,t}$ for a token i in week t is as follows:

$$O_{i,t} = \ln\left\{\frac{\text{Count}_{D,i,t}/(T\text{Count}_{D,t} - \text{Count}_{D,i,t})}{\text{Count}_{R,i,t}/(T\text{Count}_{R,t} - \text{Count}_{R,i,t})}\right\}, \quad (2)$$

where $\text{Count}_{D,i,t}$ is the number of times the token i appears in the Democrat (or Republican for $\text{Count}_{R,i,t}$) corpus. $T\text{Count}_{D,i,t} = \sum_i \text{Count}_{D,i,t}$ is the total number of tokens in the Democrat corpus, and $T\text{Count}_{R,i,t}$ is the equivalent number for the Republican corpus. A negative $O_{i,t}$ means that Republicans are more likely to use the token i and vice versa for Democrats. Typical tokens used by Republican accounts in their comments about COVID-19 include "lockdown" (mean log odds ratio over 50 weeks = −0.371), "flu" (-0.898), and "China" (-1.024), and tokens with strong Democratic identity include "pandemic" (0.489), "dead" (0.600), and "Trump" (1.016).

2.3.3. Changing Polarization over Time for COVID-19-Related Topics. Our second research question was whether the discussion of each topic shows a growing level of polarization over time. We first report changes in polarization regarding COVID-19-related topics. We used a linear function ($y = \beta_0 + \beta_1 x$) to describe the change in polarization level over time. The dependent variable y represents the estimated level of polarization, i.e., the LO estimator value, and the independent variable x represents the time period by week. Ordinary least squares (OLS) regression outcomes are given in Table S3. The polarization of the two topics, test|covid|virus|death|case ($\beta_1 = 6.787e-04$, $p < 0.001$with retweets/ $\beta_1 = 3.076e-04$, $p < 0.001$ without retweets) and life|feel|wearmask|people|mental ($\beta_1 = 2.227e-04$, $p < 0.001$ with retweets/ $\beta_1 = 1.032e-04$, $p < 0.001$ without retweets), shows highly significant linear growth. The vaccine|doctor|hospital|patient|medical topic, also related to COVID-19, just failed to show significant growth ($\beta_1 = 2.965e-04$, $p = 0.065$ with retweets/ $\beta_1 = 1.103e-04$, $p = 0.039$ without retweets), possibly because this topic emphasizes the medical system, which is also a long-discussed political issue in the U.S. Overall, there was a clear tendency for the discussion of COVID-19-related topics to become more polarized over time.

2.3.4. Changing Polarization over Time for Non-COVID-19 Topics. Next, we examined the time course of polarization for topics unrelated to COVID-19 (i.e., topics with preexisting associations with ideology). Representative tokens for the topic shown in Figure 2(d) are related to protest activities. The distribution of this topic's popularity over time is similar to that seen in Figure 2(e), which shows activity related to racism-related topics, with a sharp increase from an initially low level of activity. Although this topic does not show any long-term linear tendency, it appears that the first polarization peak of Figure 2(d) (in June 2020) reflects the lively discussion about the BLM protest in 2020. Consistent with this interpretation, the video of George Floyd's death was in widespread circulation at the end of May. We tested a bilinear model of polarization in this topic using the first linear function to describe data before week 17 (24 May 2020-30 May 2020) and a second linear function to describe data after that time point, shown by two dashed lines in Figure 2(d). Then, we compared the Bayesian information criterion score (BIC) for this two-component model with the score for the linear model. We found strong evidence (ΔBIC > 10) that a bilinear function (BIC = −235.0) fits better than a linear function (BIC = −214.5). We observed a sharp increase in polarization immediately after the event, coinciding with the sudden increase in the volume of tweets, but this polarization reduced as the discussion volume reduced: the estimator ($\beta_1 = -1.187e-03$, $p < 0.001$ with retweets/ $\beta_1 = -1.162e-03$, $p < 0.001$ without retweets) in the linear regression of the second part of the bilinear model is highly significant and negative.

Unlike Figure 2(d), however, Figure 2(e) does not show a sharp increase in polarization after the event (presumably because the topic is generic and less exclusively related to the specific relevant event). Figure 2(f) illustrates polarization around the topic of the 2020 election, where a large increase in tweet volume occurred around 1 Nov 2020. Twitter users were highly polarized throughout the year (the LO estimators are significantly higher than 0.5 for all weeks) for the election

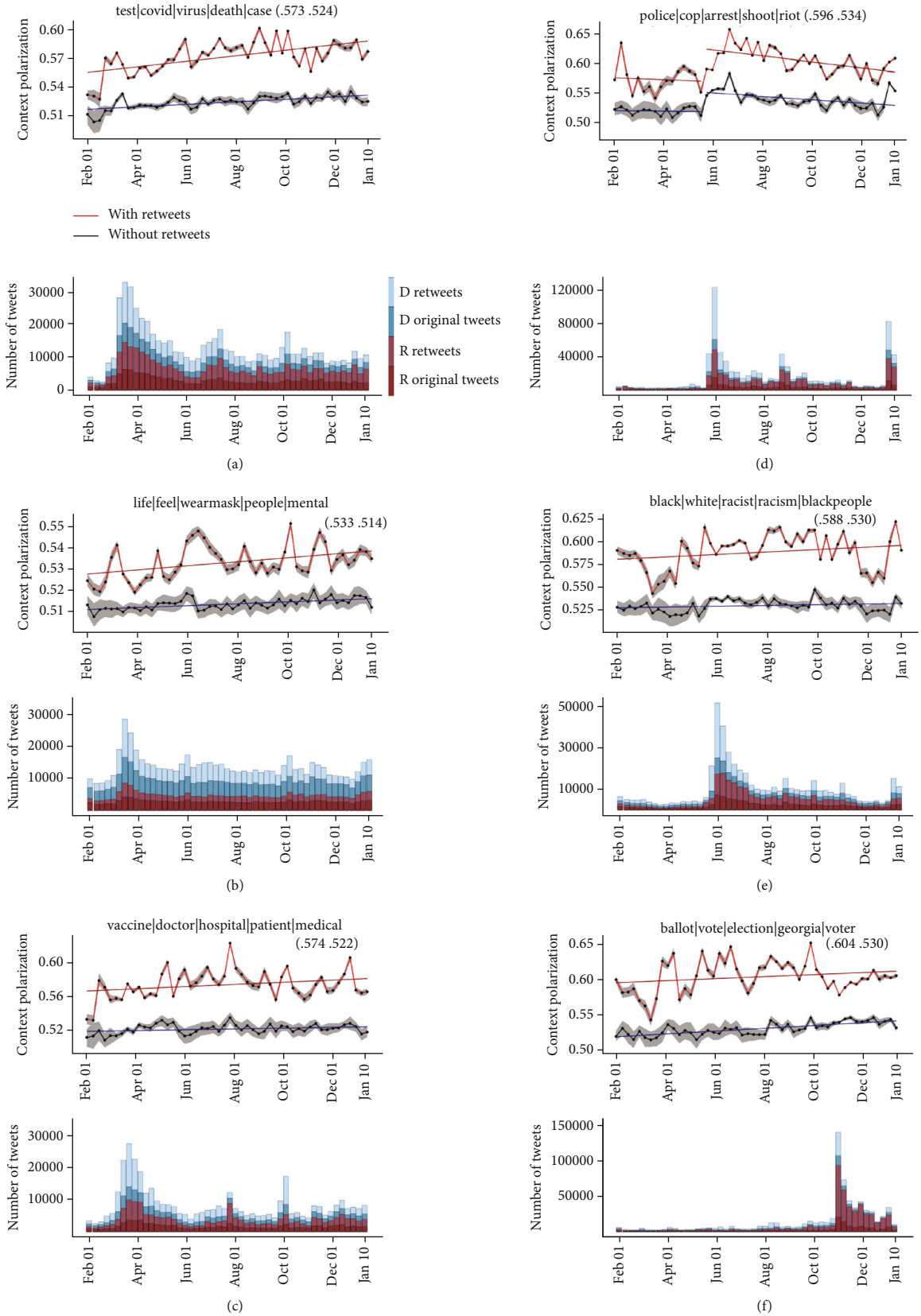

Figure 2: Week-based changes for polarizations and popularities of six topics. Titles for each subplot give the five most representative tokens from the LDA outcome, and the two values in each set of parentheses show the average leave-out estimator for whole/original tweet sets of each topic.

topic. Only the original tweet set without retweets for this topic shows a significant increase ($\beta_1 = 4.615e-04$, $p < 0.001$), but the overwhelming majority of messages were retweets (as shown in bar charts of Figure 2(f)). This phenomenon is consistent with the hypothesis that public attitudes will not become more polarized if those events relate to long-discussed affairs about racism and elections. An unexpected finding is the sharp and temporary increase in polarization level after the death of George Floyd (see Figure 2(d)).

Overall, the analysis of tweets over 50 weeks for the three topics related to COVID-19 reveals significant polarization. Even though Twitter users' discourse concerning COVID-19 was not as polarized as their discussion about topics related to the election and racism, the public's attitude toward COVID-19 became increasingly polarized over time. For other topics, in contrast, there was no systematic long-term increase in polarization over the same time period. These results underline the need to study online discourse about novel topics if the development of polarization is to be understood.

3. Study 2: Short-Term Polarization

According to models that assume polarization levels will increase gradually from an initially unpolarized environment (e.g., [27]), increasing polarization should be more likely during the early stage of discussion when people's opinions are still unstable. In other words, if our hypothesis that the trends in discussions about novel topics like COVID-19 will be different from trends for preexisting topics is correct, we will find growing polarization during the early stages of discussion when people are still unfamiliar with COVID-19, and this tendency will not exist for discussion about BLM activity and the 2020 election. Study 1 finds that polarization for topics related to COVID-19 increased over the year. However, this growth may appear some time after COVID-19 first comes into public view. The bar charts in Figure 2 show that the volume of tweets is unevenly distributed for each topic. Most tweets were sent during particular periods. For topics about COVID-19, extensive discussions start from March. In addition, the large numbers of tweets sent within a short time period enable us to compute a more precise measurement—of daily polarization intensity, rather than the weekly intensity examined in Study 1. Therefore, Study 2 supplements Study 1 in two ways. First, it examines polarization during and shortly after the period when each topic attracted substantial public attention, and second, it tracks changes in daily polarization intensity.

3.1. Data. To examine changes in the Twitter environment over a shorter period during and after trigger incidents, we selected tweets sent within 100 days (78 days for the election topic, because our dataset ends on 17 Jan 2021) after the critical event for each topic from the whole tweet sets. We ran a linear regression ($y = \beta_0 + \beta_1 x$) for day-based leave-out estimators. Independent variables were the number of days after 29 Feb 2020 (the first COVID-19 death case in the U.S.) for three COVID-19 topics, after 25 May 2020 for two racism-related topics, and after 1 Nov 2020 for the election topic. The proportions of retweets for political and event-sensitive topics were high, as shown in the bar charts of Figure 2. We used the whole tweet (i.e., including retweets) set rather than the original tweet set because the latter was too small to produce reliable results for day-based regressions.

3.2. Results. Changes in polarization (the leave-out estimator) for each day, and regression results, are shown in Figure 3 and Table S4. The line in each panel in Figure 3 gives the change of polarization intensities (LO estimator values) over 100 days and also shows the day-based linear regression models for each topic; the bar chart gives the number of tweets users sent each day. The test|covid|virus|death|case topic in Figure 3(a) again shows significant linear growth ($\beta_1 = 3.422e-04$, $p < 0.001$) in polarization as measured by the day-based leave-out estimator. Another topic related to COVID-19 in Figure 3(c), with the five most representative tokens, vaccine|doctor|hospital|patient|medical which emphasizes medical systems during the pandemic, also shows significant linear growth ($\beta_1 = 3.171e-04$, $p < 0.001$) in polarization. The life|feel|wearmask|people|mental topic in Figure 3(b), however, shows no significant growth ($\beta_1 = 8.017e-05$, $p = 0.064$).

The police|cop|arrest|shoot|riot topic in Figure 3(d) shows a trend in polarization similar to the one seen in the week-based regression. Attitudes became increasingly polarized until reaching a peak around the end of June. To identify whether polarization increases after the critical event and reduces afterwards, we ran a quadratic regression ($y = \beta_0 + \beta_1 x + \beta_2 x^2$) on polarization in the discussion of this topic, which gives a significant and positive β_2 estimator ($\beta_2 = 1.707e-05$, $p < 0.001$), as represented by the dashed line in Figure 3(d). Estimators for the election topic and the racism topic (Figures 3(e) and 3(f)) again showed no significant change in polarization after critical events.

In summary, two topics about COVID-19 showed a significant upward trend in day-based estimated polarization levels during the 100 days after 29 Feb 2020. The test|covid|virus|death|case topic shows significant linear growth in both long-term (Study 1) and short-term (Study 2) datasets. In contrast, the discussions about racism and election showed no significant change in their polarization level after important events. These results provide further evidence that political polarization for a novel topic increases over time.

4. Study 3: Polarization of Network Structure

In Study 3, we study the change in polarization by studying the *structure* of partisan retweet networks over time.

Our first two studies found that American Twitter users' discussion about COVID-19 becomes polarized over time as reflected in language usage. To check the robustness of our conclusion, Study 3 examines the development of polarization in the discussion of a novel topic (i.e., COVID-19) using a very different methodology—analysis of the structure of the social network of Twitter users.

One widely used measurement of political polarization is the frequency of cross-partisan interaction (e.g., [48–50]), that is, how often individuals choose to communicate with others from the opposite group. The higher the proportion

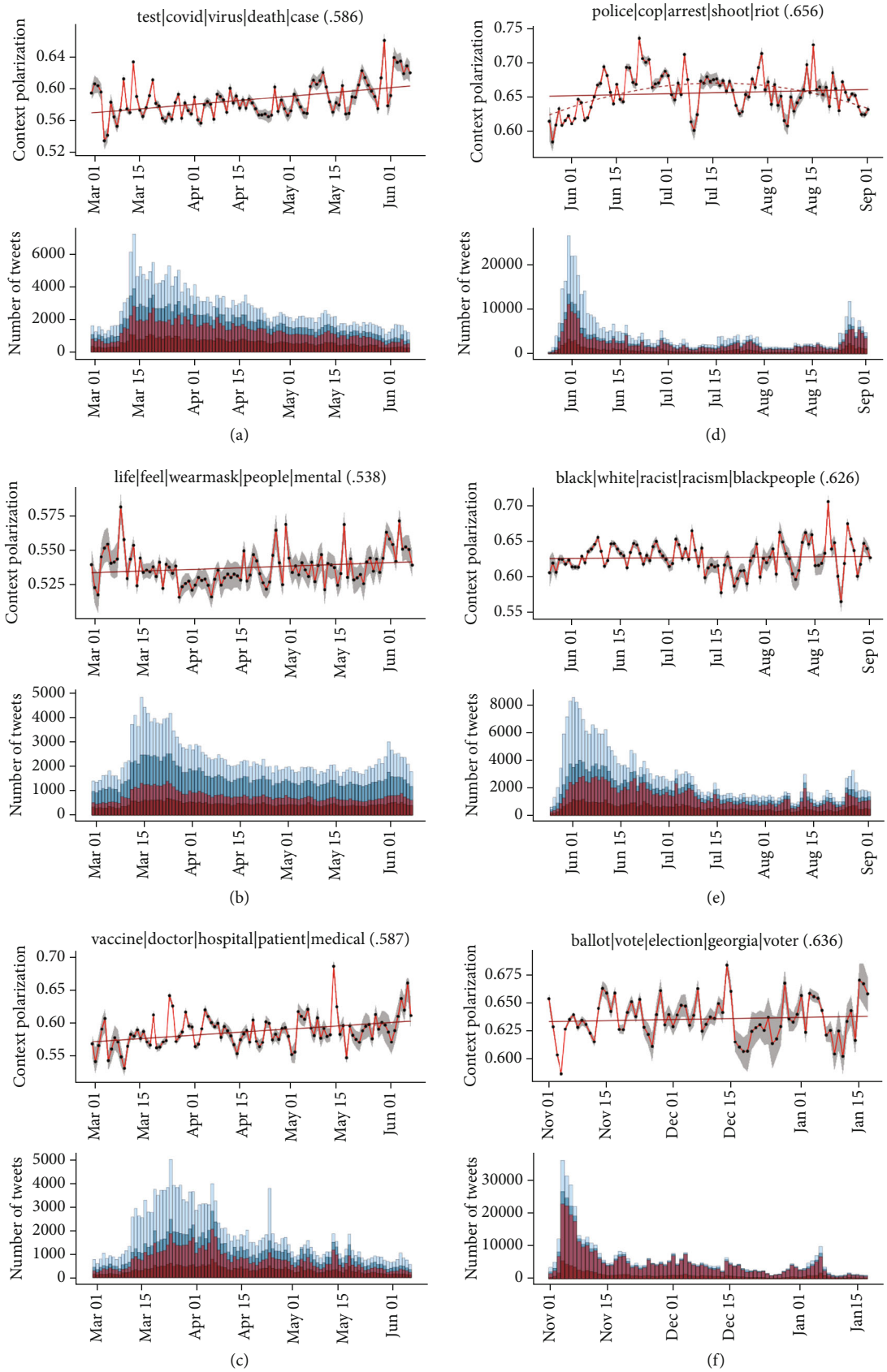

FIGURE 3: Day-based changes in polarization and popularity of six topics. Titles for each subplot give the five most representative tokens from LDA outcome, and the values in parentheses show the average leave-out estimator for each topic.

of cross-group interactions, the lower the polarization. We used the retweet network to represent these interactions. Nodes in our network represent Twitter users, and each edge (i.e., the connection between any two nodes) indicates at least one retweet relationship between a pair of users. A retweet relationship reflects the retweeter's willingness to spread the retweetee's message and often (although not always) indicates agreement with the message. A "cluster" in a social network indicates a group of nodes with internally dense connections and externally sparse connections. If the discussion environment is polarized, Democratic and Republican users will form two segregated clusters, and there will be many within-cluster edges relative to the number of between-cluster edges.

4.1. Data. We downloaded the dataset collected by Qazi et al. [51], which includes approximately 100 M English language tweets about COVID-19 from 1 Feb 2020 to 30 Apr 2020 (the period during which COVID-19 was one of the most discussed topics on Twitter). The tweets were collected by using a set of keywords (e.g., "*coronavirus*") and hashtags (e.g., "*#coronavirus*"). Full information for this dataset can be found at https://crisisnlp.qcri.org/covid19.

4.2. Methodology

4.2.1. Selection of Nodes and Tweets. In order to enable examination of time trends in network structure, we first divided the tweet set over three months into eighteen subsets, each covering a 5-day interval. We then identified and visualized the largest component of the retweet network for each subset. To clarify the visualization, we first removed self-loops (users who retweet themselves) and pure followers (users who only retweet others but are never retweeted). We then selected popular influencers (the 30% of accounts that received the highest number of retweets) to be nodes in the visualization. To estimate the political leaning of each of these nodes, we used the same lists of seed accounts as in Studies 1 and 2 (Table S1) and constructed separate sets of Democrat supporters (66,758,942 accounts who follow more Democrat accounts in Table S1) and Republican supporters (34,386,997 accounts). Nearly 50% of tweets were sent by accounts in one of our two supporter sets. Next, we estimated each influencer (node)'s partisanship as follows:

$$A_{i,t} = \frac{N_{i,t}^{D}}{N_{i,t}^{R} + N_{i,t}^{D}}, \tag{3}$$

where $A_{i,t}$ measures node i's partisanship at time interval t. $N_{i,t}^{P}(P \in [R, D])$ is the number of times node i was retweeted by Republicans or Democrats. Because the Democratic supporter set is larger than the Republican supporter set and Democratic supporters tend to be more active than Republican supporters, $N_{i,t}^{D}$ is larger than $N_{i,t}^{R}$ for most influencer nodes. Therefore, we cannot assume that a node i with $A_{i,t}$ equal to 0.5 (i.e., a node that received the same number of Democratic and Republican retweets) is neutral. Rather, this node i is more likely to be a Republican because the total number of retweets from Democrats is larger than that from Republicans. We therefore estimated the "neutral" rate of $A_{i,t}$ as follows:

$$\mathrm{Neu}_t = \frac{\sum_{j \in D}^{T_D} n_{j,t}}{\sum_{j \in D}^{T_D} n_{j,t} + \sum_{j \in R}^{T_R} n_{j,t}}. \tag{4}$$

T_R and T_D are the sizes of Republican and Democrat supporter sets. $n_{j,t}$ is the number of times user j retweets a message from any node. The average Neu_t over eighteen intervals is approximately 0.75. Any node i with $A_{i,t}$ smaller (larger) than Neu_t is therefore classified as a Republican (Democrat) influencer.

4.2.2. Polarization Measure. To quantify structural polarization in the network at each time step, we used the random walk controversy (RWC) measure defined by Garimella et al. [49]. The intuition behind this algorithm is that a random walk from one node to another is more likely to end up in the same cluster to the extent that the network is structurally polarized.

The algorithm was implemented as follows: First, nodes with the highest 25% (for Democrats) and lowest 25% (for Republicans) $A_{i,t}$ values were defined as "extreme" accounts. We selected the $0.5\% \times N$ nodes with the highest degree (number of connected edges) from each extreme set, where N is the number of nodes in the network. Second, we initialized a random walk from a randomly selected node. Imagine that there is a walker who can move along edges from a node to one of the node's neighbors but will stop when it reaches a high-degree node from either extreme set. We defined the conditional probability P_{AB} as the probability that a walker starts from a node in partisan set A and stops after reaching a high-degree node in extreme set B:

$$P_{\mathrm{AB}} = \mathrm{Prob}[\text{start from partisan set A} | \text{end in extreme set B}], \tag{5}$$

where $A, B \in [D, R]$, and each node with $A_{i,t}$ higher (lower) than the neutral rate Neu_t was classified as a member of partisan set D for Democrats (partisan set R for Republicans). Finally, we calculated the "controversy" of the network as follows:

$$\mathrm{RWC} = P_{\mathrm{DD}} P_{\mathrm{RR}} - P_{\mathrm{RD}} P_{\mathrm{DR}}. \tag{6}$$

A high RWC level, therefore, indicates that the two partisan sets are isolated from each other because the random walk is more likely to start from and stop in the same cluster than to enter the opposite cluster. We calculated the average RWC level of 100 rounds of random walks for each five-day interval. For each round, we randomly selected 1000 nodes from each partisan set to initialize the random walk.

4.3. Results. Figure 4 shows the network structures among American influencers (who account for nearly 37.3% of all influencers) for four out of eighteen time intervals (end of February, middle of March, end of March, and middle of April; Figure S6 in the supplemental material gives graphs

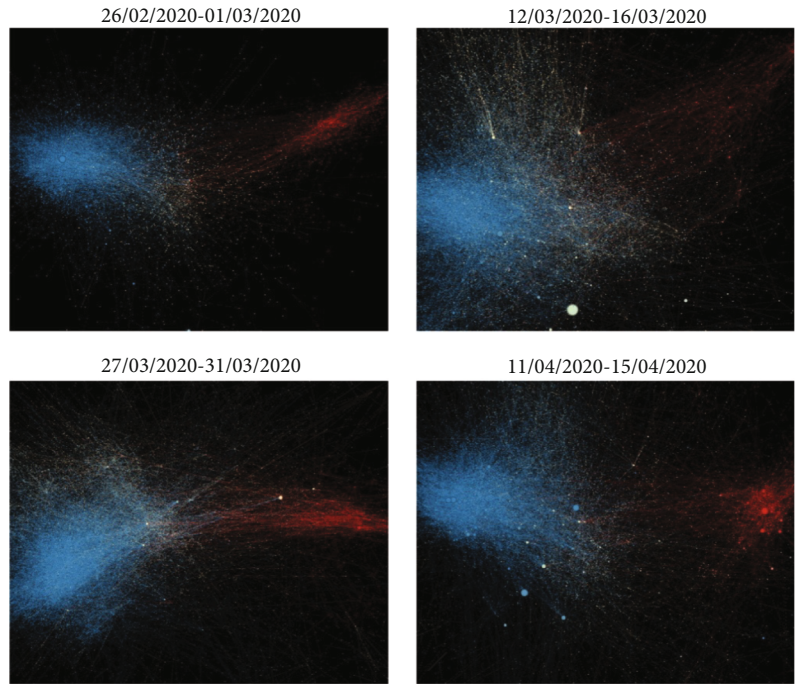

Figure 4: Structure of retweet networks. The title for each subplot gives the time interval.

for all intervals). The colors of the nodes in Figure 4 represent the estimated strength of partisanship of each influencer, i.e., $A_{i,t}$ values, from 0 (dark red; Republicans) to Neu_t (white) to 1 (dark blue; Democrats). Thus, dark color nodes exhibit clear partisanship as their $A_{i,t}$s are close to 1 or 0, and light color nodes are relatively neutral as their $A_{i,t}$s are close to Neu_t. The sizes of nodes represent the number of times they were retweeted during the corresponding time interval. Networks were visualized on Gephi with the ForceAlter2 layout algorithm [52].

Figure 4 shows that Democrats provided the majority of COVID-19 discussion on Twitter (i.e., the blue cluster is larger). Republican supporters were less active on this topic before the end of March, as shown in Figure 4(b) where the red cluster is sparse compared with the blue one. More Republicans joined the discussion from April, and Figures 4(c) and 4(d) show more significant red clusters and segregation. In addition, there are many large nodes in the blue clusters for each graph, representing influential Democratic political accounts (such as @JoeBiden and @BarackObama) that received many retweets. However, large nodes in the red cluster can be clearly observed only in Figure 4(d), representing Republican political accounts (such as @WhiteHouse45 and @GOPleader).

Figure 5 shows the changes of the estimated level of structural polarization from 21 Feb 2020 to 30 Apr 2020. (The number of American accounts before 21 Feb 2020 is too small to be used. There are fewer than 1000 nodes in either or both partisan sets.) The first panel of Figure 5 shows the strength of polarization estimated by RWC values, which shows sharp growth at the end of February, then decreases before the middle of March, and finally increases steadily afterwards.

To test the robustness of the results obtained using the RWC estimate of polarization, we used another network modularity measurement algorithm introduced by Blondel et al. [53]. This algorithm classifies nodes into multiple clusters based on the network topology and estimates the level of segregation between clusters with a scale value between -1 (low segregation) and 1 (high segregation). If the network shows significant structural polarization, i.e., high segregation between clusters, nodes will be densely connected within clusters and loosely connected between clusters. Results are shown in the second panel of Figure 5, which show a similar trend to the RWC measurement of polarization. We also find that the modularity algorithm assigns around 15% of nodes into the second largest cluster (of Republican influencers) in the second time interval from 26 Feb 2020 to 1 Mar 2020. For the third to fifth time intervals, i.e., from 2 Mar 2020 to 16 Mar 2020, in contrast, only around 5% of nodes become assigned to the second largest cluster (i.e., Republicans; over 70% nodes were assigned to the largest, Democratic, cluster), while this proportion increased to over 20% after the sixth interval. This explains why the intensity of structural polarization measured by RWC value in Figure 5 reduces sharply from the second to the fifth interval: Republican users sent many fewer tweets about COVID-19 than their Democratic peers did during this time period. Therefore, a random walk starting from a Republican node has a higher chance of ending in the Democratic cluster. The levels of structural polarization in time period 3-5 are therefore low because there were no significant Republican clusters. Other researchers have also found that Republican Twitter users were less concerned about COVID-19 than Democratic users during the early stage of the pandemic [54, 55].

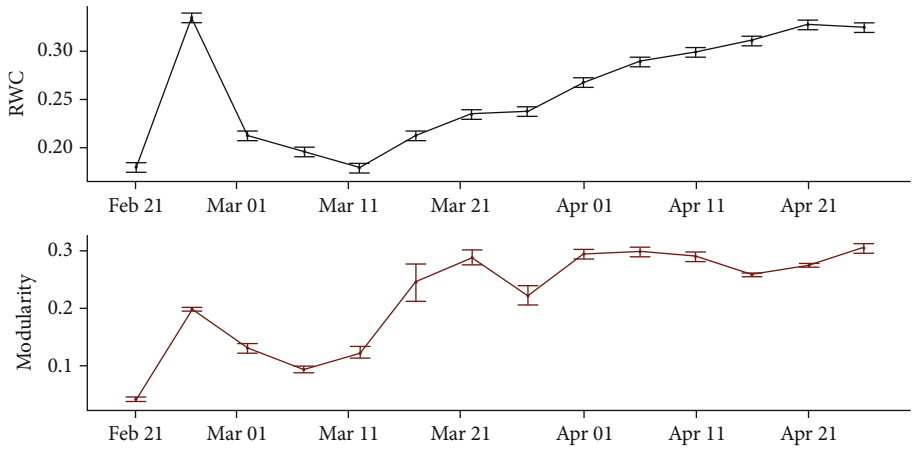

Figure 5: Change of network polarization level for American accounts from 21 Feb to 30 Apr, measured by both the random walk controversy (RWC) methodology and by modularity. The error bars represent standard deviations from 100 rounds of experiments.

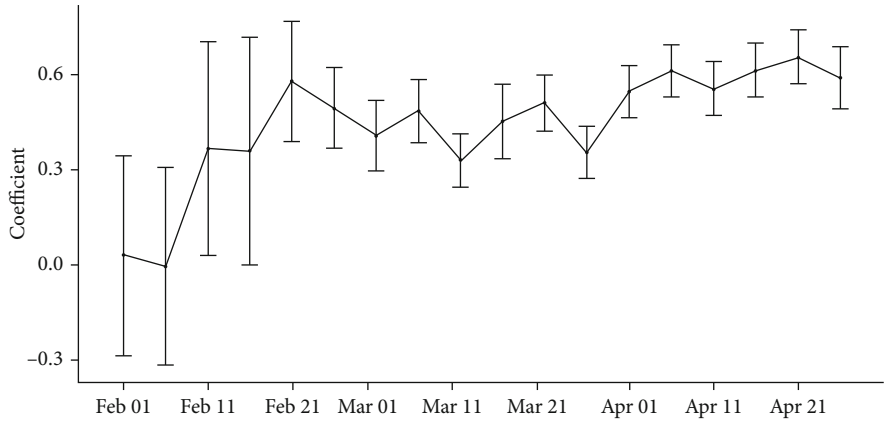

Figure 6: Change of estimated β_1 for the regression: #retweets = $\beta_0 + \beta_1 \times$ extremeness for American accounts.

Because we only chose influencers (accounts that were retweeted many times) as nodes, the network structure and RWC level only represent the interaction between influential accounts such as those owned by famous politicians, media, and commentators. We therefore also used linear regressions to test whether the broader public also became polarized in choosing retweet targets. The independent variable was each influencer's extremeness (distance from neutrality), and the dependent variable was the number of retweets the influencer received (#retweets = $\beta_0 + \beta_1$extremeness). The extremeness of an influencer i in period t was measured by the following function:

$$\text{Extremeness}_{i,t} = \begin{cases} \dfrac{A_{i,t} - \text{Neu}_t}{1 - \text{Neu}_t} \text{ if } A_{i,t} > \text{Neu}_t, \\ \dfrac{\text{Neu}_t - A_{i,t}}{\text{Neu}_t} \text{ if } \text{Neu}_t \geq A_{i,t}. \end{cases} \quad (7)$$

Thus, the extremeness level represents the gap between estimated partisanship for a Democratic or Republican (with $A_{i,t} > \text{Neu}_t$ or $\text{Neu}_t \geq A_{i,t}$) influencer and the neutral partisanship rate in the corresponding time period. The change of estimated β_1 is shown in Figure 6. The magnitude of the estimated coefficient shows an overall upward trend and is significantly larger than 0 after the fourth interval (21 Feb 2020–25 Feb 2020). Accounts with clear partisanship received more retweets than neutral accounts from the end of February. If Twitter users were apolitical in their discussions about COVID-19, neutral accounts would receive more retweets than their partisan peers because neutral accounts receive retweets from both rather than single parties. However, our analysis shows that the opposite pattern is seen in practice.

To sum up, in Study 3, we find Twitter users became polarized as measured by their choices of information sources to retweet from, and the magnitude of this polarization increased from Feb 2020 to Apr 2020. Influential Twitter accounts were more likely to retweet from, and be retweeted by, others with the same partisanship. Twitter users in the general public became more and more likely to retweet information from partisan rather than neutral accounts.

5. Conclusion

The primary aim of the present study was to examine the time course of polarization on social media for a novel topic

(COVID-19) and to compare it to polarization for existing, and already politicized, topics. Our results show that the average levels of polarization in Twitter users' discussions about BLM activity and the 2020 election are higher than for discussion of COVID-19. On the other hand, we also find long-term growth in polarization intensity reflected in the different language usage for the discussion about COVID-19, while such growth is insignificant for the other two non-COVID-19 topics. In addition, we also find Twitter users from opposite partisanships become increasingly isolated over time in their retweet network for the COVID-19 topic.

Public opinion toward COVID-19 is a focus of recent political research, and much evidence has shown that people's attitudes toward some topics have become polarized over ideological lines, especially in the U.S. (e.g., [56]). Our research shows that such attitude polarization is not preexisting at the time of a specific event and does not emerge suddenly but instead grows steadily over time. Our results also raise a question for future research about polarization and echo chamber effects in the online environment. This question concerns how claims about polarization relate to preexisting belief systems about the relevant topic. For example, many researchers have used messages posted online by voters during elections to demonstrate the existence of attitude polarization. However, our study shows that when their attitudes are compared with those before the 2020 U.S. election, Twitter users did not become more polarized during and after the election. Appendix A in the supplementary material shows the dynamic change of leave-out estimators of polarization for other political topics in Study 1, such as national security, economy, and abortion. Besides topics related to COVID-19, only one topic with five representative tokens, biden|debate|joebiden|joe|bernie, shows significant ($p < 0.01$) growth in estimated polarization intensity (see Table S5 in the supplementary material; Table S6 also gives the estimates for the nonpolitical topics). This outcome shows that even though Twitter users are polarized for many political topics, the strength of polarization is relatively stable. Therefore, future research may need to shift the focus away from those long-discussed political topics to novel debates which are initially unrelated to ideology, because the latter are more likely to cause the aggregation of partisan conflict.

As noted in Introduction, many models about opinion dynamics offer accounts of how polarization may emerge in initially unpolarized environments. However, as we also noted, it has been difficult to evaluate the general idea that polarization will increase over time, because most existing studies of polarization on social media have examined discussions about long-standing, and often politicized, topics which may already be polarized at the time of investigations. By taking advantage of the emergence of a novel topic (COVID-19) and by examining time trends over a long period, we have been able to provide empirical test of the assumption that polarization of a novel topic may increase over time. Of course, our data do not show that any novel topic will become polarized during discussion on social media, and future research will be needed to investigate the boundary conditions of this phenomenon.

Our study is subject to both strengths and limitations. A relatively underutilized advantage of online datasets is continuity. In the dataset we used for Study 1, the average number of tweets posted or retweeted by each user over the year is 24 (median 8) for discussions related to COVID-19, which is insufficient to track the opinion change for an individual based on messages he/she posted at different time points. Our research illustrates polarization trends in the online environment by using between-group comparison to analyze the differences in behaviors (language usage and retweeting) between the two parties, but it fails to track opinion dynamics at the individual level. In other words, a limitation of our work is it cannot say whether the environment becomes polarized because people become more opposed as individuals or because more extreme people join in the discussion. Different methodologies are required for analyzing information carried by individual rather than group corpora.

In summary, our research emphasizes the importance of analyzing the dynamic trend of polarization on social media, which is a research field of practical and theoretical importance. We hope future work can extend this tracking of dynamic opinions for different topics (especially novel topics such as COVID-19), social media platforms, and cultural backgrounds.

Data Availability

Our data collection and analyses have been given full approval by the Psychology Department Research Ethics Committee at the University of Warwick. However, we are not permitted to share the dataset used in Study 1 and 2, as it contains personal information of Twitter users. The dataset for Study 3 is a public dataset, currently available at https://crisisnlp.qcri.org/covid19.

Conflicts of Interest

The authors report there are no competing interests to declare.

Acknowledgments

Weize Zhao is supported by a studentship from the Chinese Scholarship Council (CSC ID: 201908060159), and this research has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant Agreement No. 788826).

Supplementary Materials

The Supplementary material includes the (1) seed Twitter accounts we used for the classification of partisan users (Table S1); (2) Mallet topic modeling results for representative tokens for each topic (Table S2), change of coherence scores with different numbers of topics (Figure S1), and visualization for two estimated topics about protest activities and COVID-19 (Figure S2); (3) validity test based on the comparison between LO estimators from partisan and random groups

(Figure S3); (4) week-based changes for LO estimators of all 35 topics in Study 1, including political topics (FigureS4) and nonpolitical topics (Figure S5); (5) linear regression results of week-based polarization intensity growth for 6 topics in Study 1 (Table S3), other political topics (Table S5), and nonpolitical topics (Table S6); (6) linear regression results of day-based polarization intensity growth for 6 topics in Study 2 (Table S4); and (7) visualizations of retweet networks about COVID-19 discussion for each of 5-day durations from 1 Feb 2020 to 30 Apr 2020 (Figure S6). *(Supplementary Materials)*

References

[1] M. Cinelli, G. D. F. Morales, A. Galeazzi, W. Quattrociocchi, and M. Starnini, "The echo chamber effect on social media," *Proceedings of the National Academy of Sciences*, vol. 118, no. 9, 2021.

[2] M. Van Alstyne and E. Brynjolfsson, "Global village or cyberbalkans? Modeling and measuring the integration of electronic communities," *Management Science*, vol. 51, no. 6, pp. 851–868, 2005.

[3] G. D. A. Brown, S. Lewandowsky, and Z. Huang, "Social sampling and expressed attitudes: authenticity preference and social extremeness aversion lead to social norm effects and polarization," *Psychological Review*, vol. 129, no. 1, pp. 18–48, 2022.

[4] P. Törnberg, "Echo chambers and viral misinformation: modeling fake news as complex contagion," *PLoS One*, vol. 13, no. 9, article e0203958, 2018.

[5] C. Rapp, "Moral opinion polarization and the erosion of trust," *Social Science Research*, vol. 58, pp. 34–45, 2016.

[6] E. Dubois and G. Blank, "The echo chamber is overstated: the moderating effect of political interest and diverse media," *Information, Communication & Society*, vol. 21, no. 5, pp. 729–745, 2018.

[7] J. Haidt and C. Bail, *Social Media and Political Dysfunction: A Collaborative Review*, New York University, 2022.

[8] R. Németh, "A scoping review on the use of natural language processing in research on political polarization: trends and research prospects," *Journal of Computational Social Science*, vol. 6, no. 1, pp. 289–313, 2023.

[9] M. Anjaria and R. M. R. Guddeti, "Influence factor based opinion mining of Twitter data using supervised learning," in *Sixth International Conference on Communication Systems and Networks*, pp. 1–8, Bangalore, India, 2014.

[10] J. Green, J. Edgerton, D. Naftel, K. Shoub, and S. J. Cranmer, "Elusive consensus: polarization in elite communication on the COVID-19 pandemic," *Science Advances*, vol. 6, no. 28, article eabc2717, 2020.

[11] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," *LREc*, vol. 10, no. 2010, pp. 1320–1326, 2010.

[12] G. A. Ruz, P. A. Henríquez, and A. Mascareño, "Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers," *Future Generation Computer Systems*, vol. 106, pp. 92–104, 2020.

[13] J. A. Frimer, M. J. Brandt, Z. Melton, and M. Motyl, "Extremists on the left and right use angry, negative language," *Personality and Social Psychology Bulletin*, vol. 45, no. 8, pp. 1216–1231, 2019.

[14] H. Saif, Y. He, and H. Alani, "Semantic sentiment analysis of Twitter," in *International Semantic Web Conference*, pp. 508–524, Springer, Berlin, Heidelberg, 2012.

[15] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe, "Predicting elections with Twitter: what 140 characters reveal about political sentiment," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 4, no. 1, pp. 178–185, 2010.

[16] K. Garimella, G. D. F. Morales, A. Gionis, and M. Mathioudakis, "Quantifying controversy on social media," *ACM Transactions on Social Computing*, vol. 1, no. 1, pp. 1–27, 2018.

[17] L. Guo, J. A. Rohde, and H. D. Wu, "Who is responsible for Twitter's echo chamber problem? Evidence from 2016 US election networks," *Information Communication & Society*, vol. 23, no. 2, pp. 234–251, 2020.

[18] M. Mosleh, C. Martel, D. Ecklec, and D. Rand, "Shared partisanship dramatically increases social tie formation in a Twitter field experiment," *Proceedings of the National Academy of Sciences*, vol. 118, no. 7, 2021.

[19] L. A. Adamic and N. Glance, "The political blogosphere and the 2004 US election: divided they blog," in *Proceedings of the 3rd international workshop on Link discovery*, pp. 36–43, Chicago Illinois, 2005.

[20] E. Bakshy, S. Messing, and L. A. Adamic, "Exposure to ideologically diverse news and opinion on Facebook," *Science*, vol. 348, no. 6239, pp. 1130–1132, 2015.

[21] S. Hong and S. H. Kim, "Political polarization on Twitter: implications for the use of social media in digital governments," *Government Information Quarterly*, vol. 33, no. 4, pp. 777–782, 2016.

[22] J. Jiang, E. Chen, S. Yan, K. Lerman, and E. Ferrara, "Political polarization drives online conversations about COVID-19 in the United States," *Human Behavior and Emerging Technologies*, vol. 2, no. 3, pp. 200–211, 2020.

[23] P. Wicke and M. M. Bolognesi, "COVID-19 discourse on Twitter: how the topics, sentiments, subjectivity, and figurative frames changed over time," *Frontiers in Communication*, vol. 6, p. 45, 2021.

[24] S. Yardi and D. Boyd, "Dynamic debates: an analysis of group polarization over time on Twitter," *Bulletin of Science, Technology & Society*, vol. 30, no. 5, pp. 316–327, 2010.

[25] D. Demszky, N. Garg, R. Voigt et al., "Analyzing polarization in social media: method and application to tweets on 21 mass shootings," 2019, http://arxiv.org/abs/1904.01596.

[26] P. Grover, A. K. Kar, Y. K. Dwivedi, and M. Janssen, "Polarization and acculturation in US Election 2016 outcomes - can Twitter analytics predict changes in voting preferences," *Technological Forecasting and Social Change*, vol. 145, pp. 438–460, 2019.

[27] R. Hegselmann and U. Krause, "Opinion dynamics and bounded confidence: models, analysis and simulation," *Journal of Artificial Societies and Social Simulation*, vol. 5, no. 3, 2002.

[28] J. Lorenz, "Continuous opinion dynamics under bounded confidence: a survey," *International Journal of Modern Physics C*, vol. 18, no. 12, pp. 1819–1838, 2007.

[29] A. E. Allahverdyan and A. Galstyan, "Opinion dynamics with confirmation bias," *PLoS One*, vol. 9, no. 7, article e99557, 2014.

[30] M. Sherif and C. I. Hovland, *Social Judgment: Assimilation and Contrast Effects in Communication and Attitude Change*, Yale University Press, New Haven, CT, 1961.

[31] V. D. Blondel, J. M. Hendrickx, and J. N. Tsitsiklis, "On Krause's multi-agent consensus model with state-dependent

connectivity," *IEEE Transactions on Automatic Control*, vol. 54, no. 11, pp. 2586–2597, 2009.

[32] G. Deffuant, S. Huet, and F. Amblard, "An individual-based model of innovation diffusion mixing social value and individual benefit," *American Journal of Sociology*, vol. 110, no. 4, pp. 1041–1069, 2005.

[33] G. Deffuant, D. Neau, F. Amblard, and G. Weisbuch, "Mixing beliefs among interacting agents," *Advances in Complex Systems*, vol. 3, p. 11, 2001.

[34] X. Wang, A. D. Sirianni, S. Tang, Z. Zheng, and F. Fu, "Public discourse and social network echo chambers driven by sociocognitive biases," *Physical Review X*, vol. 10, no. 4, article 041042, 2020.

[35] R. S. Nickerson, "Confirmation bias: a ubiquitous phenomenon in many guises," *Review of General Psychology*, vol. 2, no. 2, pp. 175–220, 1998.

[36] W. Jager and F. Amblard, "Uniformity, bipolarization and pluriformity captured as generic stylized behavior with an agent-based simulation model of attitude change," *Computational and Mathematical Organization Theory*, vol. 10, no. 4, pp. 295–303, 2005.

[37] D. Baldassarri and P. Bearman, "Dynamics of political polarization," *American Sociological Review*, vol. 72, no. 5, pp. 784–811, 2007.

[38] L. Salzarulo, "A continuous opinion dynamics model based on the principle of meta-contrast," *Journal of Artificial Societies and Social Simulation*, vol. 9, no. 1, 2006.

[39] E. Noelle-Neumann, "The spiral of silence a theory of public opinion," *Journal of Communication*, vol. 24, no. 2, pp. 43–51, 1974.

[40] B. Ross, L. Pilz, B. Cabrera, F. Brachten, G. Neubaum, and S. Stieglitz, "Are social bots a real threat? An agent-based model of the spiral of silence to analyse the impact of manipulative actors in social networks," *European Journal of Information Systems*, vol. 28, no. 4, pp. 394–412, 2019.

[41] S. E. Parsegov, A. V. Proskurnikov, R. Tempo, and N. E. Friedkin, "Novel multidimensional models of opinion dynamics in social networks," *IEEE Transactions on Automatic Control*, vol. 62, no. 5, pp. 2270–2285, 2017.

[42] A. Flache and M. W. Macy, "Small worlds and cultural polarization," *The Journal of Mathematical Sociology*, vol. 35, no. 1-3, pp. 146–176, 2011.

[43] F. Baumann, P. Lorenz-Spreen, I. M. Sokolov, and M. Starnini, "Modeling echo chambers and polarization dynamics in social networks," *Physical Review Letters*, vol. 124, no. 4, article 048301, 2020.

[44] A. Bramson, P. Grim, D. J. Singer et al., "Understanding polarization: meanings, measures, and model evaluation," *Philosophy of Science*, vol. 84, no. 1, pp. 115–159, 2017.

[45] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[46] A. K. McCallum, "Mallet: a machine learning for language toolkit," 2002, http://mallet.cs.umass.edu.

[47] M. Gentzkow, J. Shapiro, and M. Taddy, "Measuring group differences in high-dimensional choices: method and application to congressional speech," *Econometrica*, vol. 87, no. 4, pp. 1307–1340, 2019.

[48] M. Conover, J. Ratkiewicz, M. Francisco, B. Goncalves, F. Menczer, and A. Flammini, "Political polarization on Twitter," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 5, no. 1, pp. 89–96, 2011.

[49] K. Garimella, G. D. F. Morales, A. Gionis, and M. Mathioudakis, "Political discourse on social media: echo chambers, gatekeepers, and the price of bipartisanship," in *International World Wide Web Conferences Steering Committee*, pp. 913–922, Lyon France, 2018a.

[50] A. J. Morales, J. Borondo, J. C. Losada, and R. M. Benito, "Measuring political polarization: Twitter shows the two sides of Venezuela," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 25, no. 3, article 033114, 2015.

[51] U. Qazi, M. Imran, and F. Ofli, "GeoCoV19," *SIGSPATIAL Special*, vol. 12, no. 1, pp. 6–15, 2020.

[52] M. Jacomy, T. Venturini, S. Heymann, and M. Bastian, "ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software," *PLoS One*, vol. 9, no. 6, article e98679, 2014.

[53] V. D. Blondel, J. L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, article P10008, 2008.

[54] E. Jing and Y. Ahn, "Characterizing Partisan Political Narratives about COVID-19 on Twitter," *EPJ Data Science*, vol. 10, no. 1, 2021.

[55] A. Panda, D. Siddarth, and J. Pal, "COVID, BLM, and the polarization of US politicians on Twitter," 2020, http://arxiv.org/abs/2008.03263.

[56] J. Kerr, C. Panagopoulos, and S. van der Linden, "Political polarization on COVID-19 pandemic response in the United States," *Personality and Individual Differences*, vol. 179, article 110892, 2021.