UNIVERSITY OF COPENHAGEN

M.Sc. in Social Data Science

Ideological and affective polarization in social networks

A case study of the Covid-19 Twitter debate in the US

Marilena Hohmann

Supervised by Hjalmar Alexander Bang Carlsen & August Lohse

May 31, 2022

Marilena Hohmann

Ideological and affective polarization in social networks

M.Sc. in Social Data Science, May 2022

Supervisors: Hjalmar Alexander Bang Carlsen & August Lohse

Character Count: 95,759 characters in total

University of Copenhagen

Faculty of Social Sciences

Master's Degree in Social Data Science

Øster Farimagsgade 5

1353 Copenhagen K

Acknowledgements

I would like to thank my thesis supervisors Hjalmar Alexander Bang Carlsen and August Lohse for their invaluable guidance on the project. I am extremely grateful for your help, encouragement, and feedback throughout the thesis writing process. Tusind tak!

A seriously big thank you to Michele Coscia for (informally) supervising this and previous projects. Thank you so much for almost a year of weekly discussions about networks and polarization, for all your time, and for making sense of my confusion.

Many thanks, et kæmpestort tak, und ein riesengroßes Dankeschön to Alysha Chamadia, Aurelia Duisberg, Ellen Linnert, Louis Hohmann, and Nicklas Müller for the emotional and content-related support. I really appreciate that you took the time to discuss my project with me.

Abstract

In recent years, the algorithmically-driven exposure to information on digital platforms has fuelled concerns about polarized digital communities. As a consequence, the analysis and quantification of political polarization has attracted immense interest across scientific disciplines. However, the polarization measures used in prior research only partly capture the concepts under investigation. I argue that a stronger focus on social connections that foster homophily is needed and that theoretically justified measures of political polarization should take the network of interpersonal interactions into account. To address this task, the thesis first discusses and identifies appropriate measures for quantifying polarization in social media networks and subsequently applies them to a large-scale Twitter data set. As part of the research design, I collect approximately 140 million tweets discussing Covid-19 which were posted by US-based Twitter users in between February 2020 and July 2020. By leveraging computational methods including transfer learning and network science tools, I conduct an exploratory analysis of ideological and affective polarization in the Covid-19 debate on Twitter. I find that both types of polarization were low in early February 2020 and then increased to moderately high levels in the following months before reaching very high levels in July 2020.

Contents

Part I

SCIENTIFIC ARTICLE

1 Introduction

Researchers, politicians, and journalists have widely discussed the issue of political polarization in the recent past. For the conceptualization of polarization, both the increasing extremity of ideological views as well as the structure of the interactions between individuals matter (Baldassarri & Page, 2021). When people only associate with others whose views they share, homogeneous and polarized communities emerge (Baldassarri & Page, 2021). Depending on what drives these in-group versus out-group dynamics, the literature distinguishes between two types of polarization (Wilson et al., 2020). Ideological polarization refers to opposing ideological views among partisans; and affective polarization is concerned with the affective attitude towards others (Wilson et al., 2020).

In this paper, I use the conceptualization outlined above to identify appropriate measures of ideological and affective polarization in social media networks. Moreover, I show that these measures outperform other popular approaches to quantifying polarization. In a case study of the debate about Covid-19 restrictions, I apply these measures to a large-scale data set comprising tweets posted by Twitter users over the course of half a year in 2020.

The exploratory analysis of this data set shows that both ideological and affective polarization were low in early February 2020 when Covid-19 was not yet a broadly discussed topic. In the following weeks, the Twitter debate became both ideologically and affectively polarized and I find moderately high levels of both types of polarization. I observe a peak in ideological polarization in mid-April 2020, at a time when policies regarding face masks were changing. The highest levels of affective polarization in the Twitter data set are reported in July 2020 which coincides with an intense public debate about whether to reopen schools after the summer break.

2 Literature Review

Political polarization is an umbrella term that covers a range of different concepts which are often understood along two axes: ideology and affect (Kubin & von Sikorski, 2021).[1] As the name suggests, ideological polarization refers to increasing differences in the ideological leaning among individuals and political elites. The literature conceptualizes ideological polarization as follows: on the one hand, it is understood as the placement of individuals on an ideological, e.g., liberal–conservative scale, or a partisan scale involving parties on the political spectrum (Abramowitz & Saunders, 2008; Fiorina & Abrams, 2008). When partisans move towards the opposing ends of those scales, ideological polarization increases (Abramowitz & Saunders, 2008). On the other hand, ideological polarization can refer to increasing differences in policy stances (Fiorina & Abrams, 2008). The opinions that people hold regarding political and social issues, such as immigration or labor market policies, indicate to what extent society is ideologically polarized. Both partisan identity and policy stances are related, but it is important to point out that the term ideological polarization can refer to either, or both, of these processes (Kubin & von Sikorski, 2021).

Affective polarization describes the affective attitude towards like-minded and disagreeing others and it is thus concerned with in-group versus out-group dynamics (Druckman & Levendusky,

[1] Note that a further distinction is sometimes made between the polarization of political elites (elite polarization), or society at large (mass polarization) (Carothers & O'Donohue, 2019). Since the measures proposed here can be equally applied to elite and mass discourses and the case study focuses on the micro-blogging platform Twitter where private users and political elites discuss in a digital public space, I do not distinguish between elite and mass polarization here.

2019). The concept refers to either in-group favoritism, out-group hostility, or both affective reactions taking place simultaneously (Iyengar et al., 2012). While there is widespread consensus that affective polarization has increased over the last decades (Iyengar et al., 2019; Mason, 2016), it is disputed whether in-group favoritism or out-group hostility is the main driver of this development. For instance, Iyengar et al. (2012) argue that in-party affection has only slightly changed in the United States since the 1980s while out-party dislike has increased steeply. Conversely, Park et al. (2021) conclude that affective polarization is mainly driven by in-group favoritism rather than out-group hostility.

Recently, the relation between ideological and affective polarization has attracted scholarly attention and several competing explanations have emerged. First, partisan identity is said to reinforce affective polarization (Dias & Lelkes, 2022). When the identification with a political party becomes fundamental to the perception of one self and others, this can fuel hostility directed towards political adversaries (Dias & Lelkes, 2022). Second, stark ideological disagreements might strengthen animosity towards other parties and their voters (Lelkes, 2021; Orr & Huber, 2020; Webster & Abramowitz, 2017). According to these explanations, ideological polarization reinforces affective polarization. However, other studies argue that it is in fact affective polarization that drives ideological polarization. Druckman et al. (2021a, 2021b) for instance show that pre-pandemic out-party hostility shapes the subsequent pandemic-related beliefs and practices that individuals develop. To conclude, the direction of the relation between ideological polarization and affective polarization is disputed and several explanations including mutual reinforcement between the two types of polarization are possible (Baldassarri & Page, 2021).

2.1 Traditional Measures of Political Polarization

Traditionally, surveys such as the American National Election Survey are used to measure the two types of polarization (Iyengar et al., 2019). To quantify ideological polarization, survey respondents are asked to indicate how liberal or conservative they are, which party they usually vote for, or what opinion they hold regarding different political issues (Kubin & von Sikorski, 2021). When those ideological leanings diverge strongly, ideological polarization is high.

A survey-based measure of affective polarization not only includes the ideological leanings of the survey respondents, but also how they feel towards others holding similar or opposing views (Iyengar et al., 2012). For instance, participants are asked to rate how they feel towards the out-group on a 'feeling thermometer' that ranges from 0 (very cold) to 100 (very warm) (Iyengar et al., 2012). Other approaches include 'social distance' measures that ask respondents to indicate whether they would want to have a colleague, neighbor, be friends with, or marry someone with an opposing ideological leaning (Druckman & Levendusky, 2019). Furthermore, respondents might indicate how much they trust their in-group versus out-group and which stereotypes or traits, such as intelligence, independence, selfishness, or ignorance, they associate with them (Druckman & Levendusky, 2019). Affective polarization is considered to be high when there is a clear difference between the in-group versus out-group feelings reported by the participants.

Although surveys have been predominantly used to measure political polarization, it is important to note that these approaches are subject to several shortcomings (Iyengar et al., 2019). For example, the external validity of surveys is low and the survey responses might differ from the respondents' real-world behavior when interacting with their in-group versus out-group (Iyengar et al., 2019; Lelkes, 2021). Moreover, survey instruments are reactive and the results might thus be influenced by the framing of the questionnaire items (Iyengar et al., 2019). As recent research shows, this is especially problematic for the quantification of affective polarization (Druckman & Levendusky, 2019). Survey participants consistently overestimate social identities and cleavages along party lines both with regards to their in-group (Vandeweerdt, 2021) and their out-group (Ahler & Sood, 2018). For example, Ahler and Sood (2018) show that respondents estimate that 31.7% of Democrats belong to the LGBTQ+ community although the actual share is 6.3%. In another example, respondents estimate that 38.2% of Republicans earned over $250,000 per year, whereas the real fraction is 2.2.%.

2.2 Political Polarization on Social Media

The findings outlined above emphasize that survey-based measures need to be interpreted cautiously and that more information on actual behavior among and towards partisans is needed. As a consequence, recent research has turned to behavioral trace data collected on social media

platforms. A small number of studies have attempted quantifying affective polarization in social media data by investigating the sentiment expressed towards other users (Marchal, 2022; Mentzer et al., 2020; Tyagi et al., 2021; Yarchi et al., 2021). The lexica underlying sentiment analysis tools specify, for instance, that 'happy' has a positive valence while 'sad' has a negative valence. The tool calculates the overall sentiment of a sentence by taking rules such as negation into account (Marchal, 2022; Tyagi et al., 2021). Affective polarization is then measured as the difference between the sentiment directed towards the in-group versus the out-group.

A second strand of research investigates ideological polarization on social media (e.g. Barberá, 2015; Cinelli et al., 2021). The most recent studies have focused on ideological polarization in the Covid-19 debate on digital platforms, in particular on Twitter. All studies find that the Twitter debate regarding Covid-19 was ideologically polarized: J. Jiang et al. (2020) show that liberal and conservative Twitter users mostly engage in separated communities and that partisanship is an indicator of support towards the administration at the time. Similarly, Simchon et al. (2022) report that Twitter users communicate partisan opinions in the Covid-19 debate. In relation to face masks, Lang et al. (2021) conclude that the majority of users in their sample support face masks and that the debate between pro-mask and anti-mask users was emotionally charged. The remaining studies focus on the vaccination discourse and find that this topic led to especially high levels of ideological polarization (X. Jiang et al., 2021; Reiter-Haas et al., 2022). Conservative Twitter users express themselves more skeptically towards vaccines, while liberal users tend to show higher trust in vaccination and the medical field in general (X. Jiang et al., 2021; Reiter-Haas et al., 2022). These findings are in line with survey-based research examining polarization of the Covid-19 debate which confirms that the beliefs held by individuals differ along ideological lines (Bernacer et al., 2021; Bruine de Bruin et al., 2020; Druckman et al., 2021a, 2021b; Green et al., 2020; Kerr et al., 2021; Pennycook et al., 2021).

3 Research Objectives

As outlined above, the conceptualization of ideological and affective polarization is based on the idea that the connections between individuals and the affective attitudes they hold play a central

role. Although social ties are at the center of the conceptual understanding of polarization, the measures proposed in prior studies do not always explicitly account for them. For example, studies that solely focus on the distribution of ideological views or hostile attitudes fail to capture any information on the social ties between the individuals which are considered in the analysis.

It follows that a theoretically justified quantification of ideological and affective polarization needs to take the social connections between individuals into account. I argue that methods of social network analysis are well suited for this task. In a network representation, individuals can be modelled as nodes and the social relations between them can be represented as edges. Since a network perspective on political polarization can suitably capture the ideas that underlie the understanding of the concepts at hand, the first research question asks:

RQ1: How can ideological and affective polarization be quantified in a social network?

Turning from polarization measures to their application, I note that there is a gap in the literature on polarization in the Covid-19 debate on social media. The studies discussed above exclusively focus on ideological polarization but they do not analyze affective polarization. Since only investigating one type of polarization provides an incomplete picture of the online polarization dynamics in the Covid-19 debate, the second research question addressed here is:

RQ 2: How do ideological and affective polarization develop in the US-based Twitter debate over time?

4 Operationalization and Methods

Since RQ1 focuses on quantifying ideological and affective polarization in social networks, I start this section by introducing basic network-related terms (see Figure 4.1). In a social network, each individual is modelled as a node and their social ties are represented as edges. Figure 4.1(a) shows a simple graph consisting of two nodes and one edge connecting them, while (b) is an example of a directed network in which the arrow indicates the direction of the connection. In order to model the ideological leaning of each individual in the network, node attributes are used

as shown in (c) where blue reflects liberal individuals and red indicates conservative individuals. Moreover, edge attributes are assigned to model how favorable (dashed line) or hostile (solid line) the ties between the individuals are.

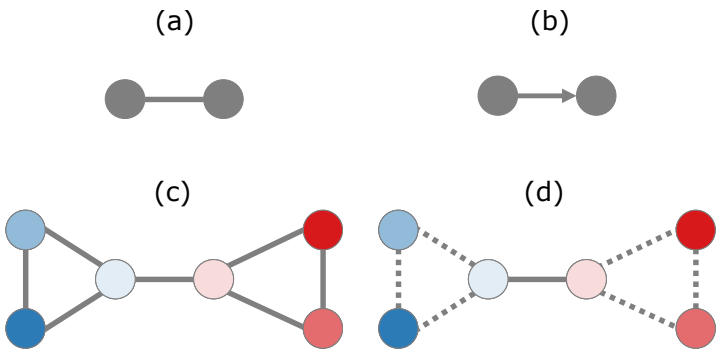


Figure 4.1: Basic network representations. Individuals are nodes (circles) and their relations are edges (lines). (a) undirected network, (b) directed network, (c) undirected network in which the node color reflects ideological leaning (blue for liberals, red for conservatives), (d) undirected network in which the edge pattern reflects favorability (dashed line) or hostility (solid line).

4.1 Ideological Polarization

Ideological polarization refers to an increasing gap in the ideological positions held by individuals, as well as the clustering of those individuals in communities of like-minded others (Baldassarri & Page, 2021). In previous work, we have identified relevant dimensions which a measure of ideological polarization should take into account (Hohmann et al., 2022). Below are the two dimensions which I focus on here:

1. The ideological dimension: Figure 4.2(a) illustrates a non-polarized setting with many moderate and only a few strongly partisan leanings (in the left plot); and a polarized environment with many and opposing partisan views (in the right plot).

2. The structural dimension: Figure 4.2(b) shows an example of a non-polarized network in which all individuals are randomly connected and there are no communities (on the left) and a network which fosters polarization as it is split into several separated clusters (on the right).

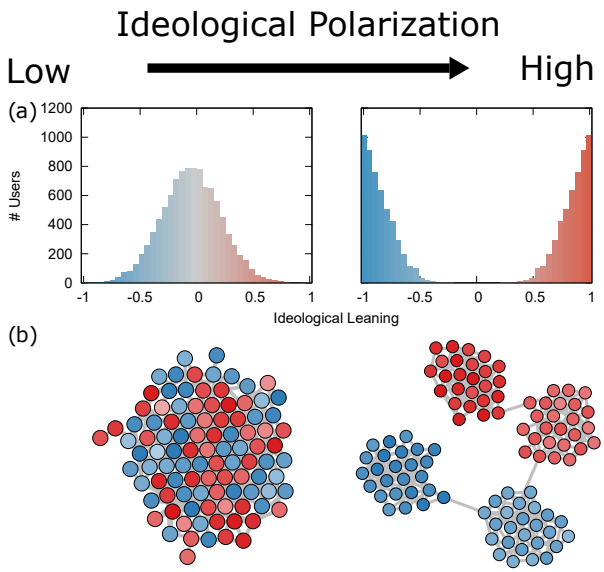


Figure 4.2: Two dimensions of ideological polarization (adapted from Hohmann et al., 2022). (a) Ideological dimension: the plots show the distribution of ideological leanings on a scale from liberal (-1) to conservative ($+1$). (b) Structural dimension: users are represented as nodes and connected by an edge if they interact on social media. Node color reflects the users' ideological leaning (blue for liberals, red for conservatives).

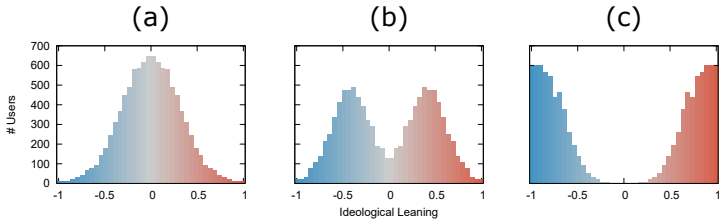
The goal of the Twitter analysis is to compare the ideological polarization over the course of six months and I therefore need a measure that can be easily compared across many networks. Among the measures that condense the ideological polarization of a network into a numerical score, I choose the Generalized Euclidean Polarization Index to measure ideological polarization (Hohmann et al., 2022). This measure captures the distance between all agreeing and disagreeing nodes in a network. The formal definition of the measure and details on its properties are provided in the Companion Paper Section 8.

4.2 Ideological Polarization Measure: Experiments

To demonstrate how the Generalized Euclidean Polarization Index can be interpreted and how it outperforms other comparable measure, I repeat variations of some of the experiments presented in earlier work (Hohmann et al., 2022). In particular, I focus on the Assortativity Coefficient which quantifies to what extent the users across a network are directly linked to like-minded others (Mønsted & Lehmann, 2022). Moreover, I take the Random Walk Controversy measure

into account which quantifies how well connected two communities are by simulating random walks between them (Garimella et al., 2018).

Ideological Dimension. The results of the first experiment are shown in Figure 4.3. Here, I generate a random network with no community structure. In a random network, all nodes can randomly connect to one another and, based on the network structure alone, I expect the polarization to be low.



	(a)	(b)	(c)
Generalized Euclidean Polarization Index	**1.56**	**3.43**	**6.02**
Assortativity Coefficient	-0.017	-0.018	-0.017
Random Walk Controversy	0.055	0.064	0.052

Figure 4.3: The ideological dimension. The first row shows the distribution from which user leanings were drawn. The other rows specify the values of the different measures compared here. All values are calculated based on a random graph with $n = 100$ nodes and approximately $m = 340$ edges as well as the respective distribution of leanings as shown in the first row. The values reported are averages over 100 iterations of the experiment.

Figure 4.3(a) then starts with a normal distribution of ideological leanings, which gets more polarized from 4.3(b) to 4.3(c). The Generalized Euclidean Polarization Index captures this change appropriately since it grows over (a)-(c). In contrast, the Assortativity Coefficient and the Random Walk Controversy measure cannot account for the change in the distribution of ideological leanings.

Structural dimension. For the second experiment, I generate three networks with different topologies but the same number of nodes and edges as before. In Figure 4.4(a), the network structure is random and, as in the previous experiment, there are no clearly distinguishable communities. In Figure 4.4(b) communities emerge and they get increasingly separated in 4.4(c). I therefore expect the ideological polarization score to grow from (a) to (c). The distribution of node leanings is similar to the right-most histogram in the previous figure and it is fixed for all three networks shown below. Based on the results, I conclude that all three measures can

capture the structural polarization dimension as expected. In summary, these experiments show that the Generalized Euclidean Polarization Index is the only measure that suitably captures and distinguishes between both dimensions of ideological polarization.

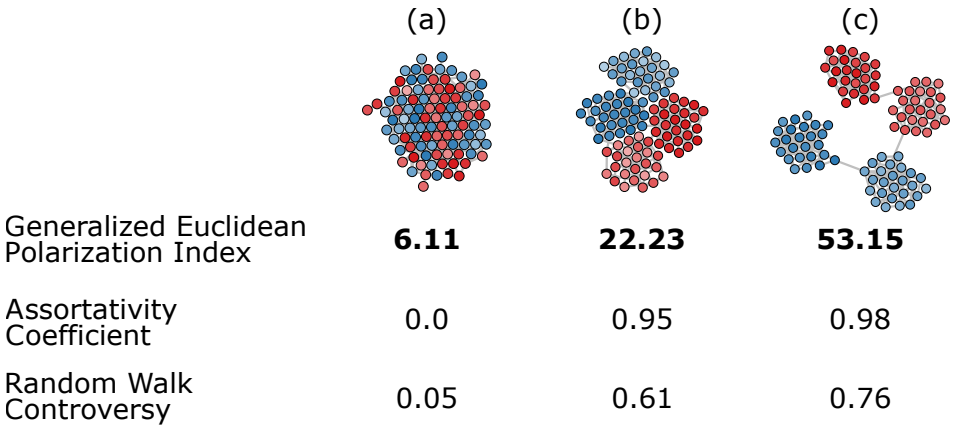


Figure 4.4: The structural dimension. The first row shows three stochastic block models with $n = 100$ nodes and approximately $m = 340$ edges each. In (a)-(c), the communities in the network become increasingly separated. The bottom rows show the values of the three measures compared here. The values reported are averages over 100 iterations of the experiment.

4.3 Affective Polarization

Similar to the previous section, I define two dimensions of affective polarization that a network-based measure should be able to account for. As noted in the literature review on affective polarization, the concept can either refer to favoritism among agreeing individuals, hostility among disagreeing individuals, or both. Since hostility-based measures are predominantly used in other studies (see e.g. Bougher, 2017; Druckman et al., 2021a), I choose to focus on hostility here. However, the measure proposed below could be easily adjusted to include favorability or combine both.

1. The interplay between ideological difference and hostility: Figure 4.5 exemplifies the relationship between (dis)agreement and hostility in the user interactions. If there is no relation between the ideological difference and the hostility directed at another user, then the affective polarization is low. If agreement correlates with low levels of hostility and disagreement correlates with high levels of hostility, then affective polarization is high.

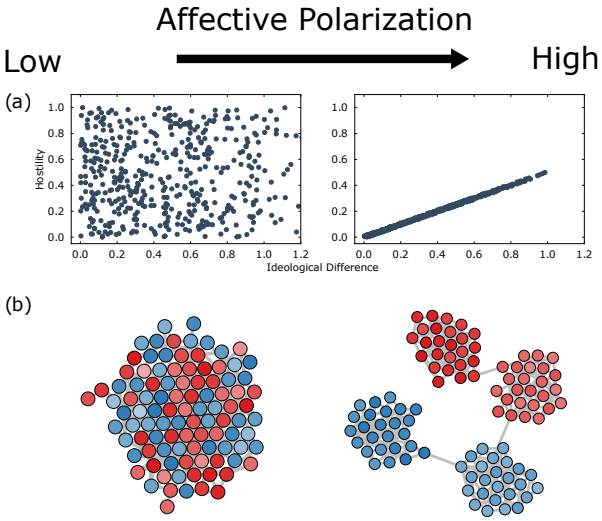


Figure 4.5: Dimensions of affective polarization. (a) Interplay between ideological difference and hostility: the scatter plots show how these variables can be correlated. (b) Structural dimension: users are randomly connected in the left graph, whereas there are distinguishable communities in the right graph.

2. The structural dimension: In the left network, there are no communities and users therefore have the chance to observe many other (dis)agreements and their related hostility levels around them. I expect affective polarization to be comparatively low in this case. If, however, users are embedded in separated communities and thus only surrounded by agreeing others whose interactions are not hostile, this might reinforce their feelings towards the in-group versus out-group, and affective polarization is therefore high.

Many survey-based studies of affective polarization rely on a Pearson correlation coefficient to measure how hostile respondents feels towards others with opposing ideological leanings. However, a Pearson correlation coefficient cannot capture any topological information about the network. I therefore use the recently introduced Pearson correlation for complex networks (Coscia, 2021) to define a new measure of affective polarization, the Affective Polarization Coefficient. This measure calculates the correlation between the ideological difference and hostility, given the correlation of the two variables in the vicinity of each edge. The influence of neighboring relations decays exponentially over the network; i.e, with each step that is further away from a node pair, the values of other node pairs have exponentially less influence. The experiments below show how this measure can be interpreted and a formal definition is provided in the Companion Paper Section 8.

4.4 Affective Polarization Measure: Experiments

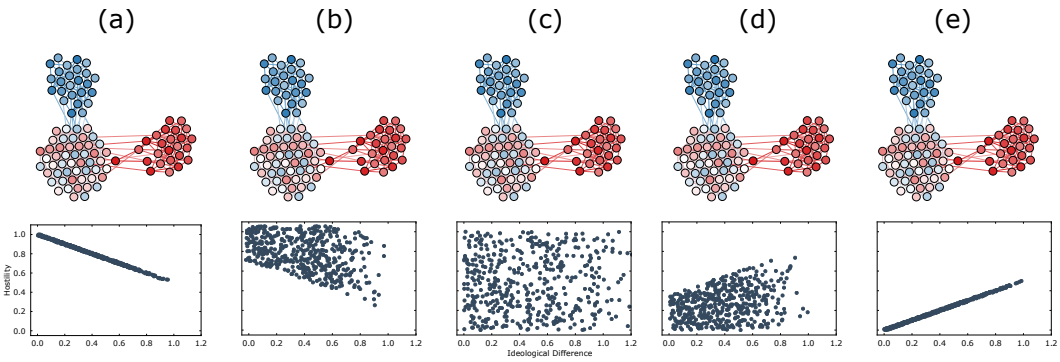
I test whether the Affective Polarization Coefficient captures both dimensions of affective polarization. Moreover, I compare it to the Pearson Correlation Coefficient as one of the established, yet not network-based, measures of affective polarization. Lastly, I investigate the only other network-based affective polarization measure that is reported in the literature: Tyagi et al. (2021) suggest an approach that relies on the Earth Mover's Distance (EMD), a metric to calculate the distance between two distributions, to quantify affective polarization. Their approach relies on the idea that the individuals in the network can be categorized into two groups; in the case of Tyagi et al. (2021) these groups consist of climate change believers and disbelievers. Affective polarization is then calculated for each of the two groups separately as the difference between in-group versus out-group hostility. In the experiments presented below, I split the nodes into a liberal (blue) and conservative (red) group to calculate the EMD measure.

Interplay between ideological difference and hostility. In this experiment, I test how the three measures account for the relation between ideological difference and hostility. I generate a network with three communities; a small community of liberal nodes (blue), a small community of conservative nodes (red) and a large community of mixed, moderate nodes as shown in Figure 4.6. The network structure and the ideological leaning assigned to each node stays fixed for all experiments (a)-(e). I only change the hostility value assigned to each edge between two nodes as shown in the scatter plots.[1]

In Figure 4.6(a) and 4.6(b), the hostility decreases as the ideological difference increases. This relation gets weaker from (a) to (b); and in (c) the two variables are entirely uncorrelated. In (d) and (e), the hostility and ideological difference are aligned and I therefore expect the highest affective polarization in these cases.

As expected, both the Pearson Correlation and the Affective Polarization Coefficient increase over (a)-(e). Both measures can thus account for the interplay between ideological difference and hostility. On the contrary, the Earth Mover's Distance does not behave as expected. According to

[1] I rely on scatter plots to show how node pairs differ in their ideological leaning and hostility values since it is not possible to meaningfully visualize this in the networks in the first row of Figure 4.6.



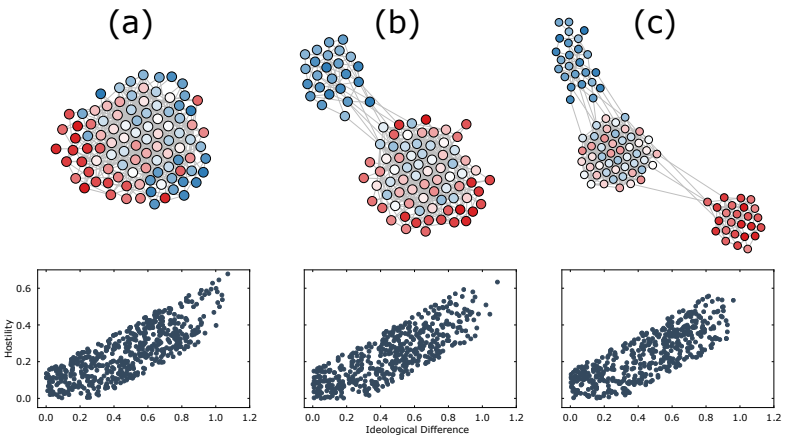
	(a)	(b)	(c)	(d)	(e)
Affective Polarization Coefficient	**-1.0**	**-0.78**	**0.0**	**0.78**	**1.0**
Pearson Correlation Coefficient	-1.0	-0.37	0.0	0.37	1.0
EMD (Blue)	-0.02	-0.06	0.0	0.06	0.02
EMD (Red)	-0.03	-0.08	0.0	0.07	0.04

Figure 4.6: The first row shows the network topology used in this experiment where the node color reflects opinion (blue as liberals, and red as conservatives). The second row shows how ideological difference and hostility are related across all node pairs in the experiment. The rows below report the results for the measures compared here. Since the EMD measure proposed by Tyagi et al. (2021) is calculated per group, I specify a value for the group of blue nodes and red nodes separately. All values reported are averages over 100 iterations of the experiment.

this measure, the example in (b) is less polarized than (a), and (e) is less polarized than (d) which should not be the case based on the argumentation above.

Structural dimension. Next, I investigate how the measures capture the network topology. I generate three different networks in which the number of nodes and edges as well as the ideological difference and the hostility between the node pairs stays constant. It is thus only the network topology that changes. In Figure 4.7(a), the nodes are randomly connected and there is no community structure. In (b), there is a community of liberal nodes (blue) which is clearly separated from the remaining network. In (c), there is both a liberal (blue) and a conservative (red) community as well as a mixed community of moderate nodes. Since the networks evolve from random (a) to modular (c), I expect the affective polarization to grow.

The Affective Polarization Coefficient increases from (a) to (c) as expected. The EMD values are smallest for (a) and larger for the other two networks. Moreover, the EMD measure seems to account for the fact that there is a clearly distinguishable community of blue nodes in (b) as



	(a)	(b)	(c)
Affective Polarization Coefficient	**0.86**	**0.93**	**0.96**
Pearson Correlation Coefficient	0.78	0.78	0.78
EMD (Blue)	0.03	0.12	0.06
EMD (Red)	0.04	0.03	0.07

Figure 4.7: The first row shows the network topology used in this experiment where the node color reflects ideological leaning (blue for liberals and red as conservatives). The second row shows how ideological difference and hostility are related across all node pairs in the experiment. The rows below report the results for the measures compared here. The values reported are averages over 100 iterations of the experiment.

the values for the blue group (0.12) are notably higher than the values for the red group (0.03). Lastly, the Pearson correlation stays constant for all three network topologies since it cannot account for any topological features. I conclude that among the three measures tested here, only the Affective Polarization Coefficient can appropriately capture both theoretical dimensions of affective polarization.

To summarize, in this section I have identified dimensions that are relevant to the quantification of ideological and affective polarization. Moreover, I have outlined two measures, the Generalized Euclidean Polarization Index and the Affective Polarization Coefficient, which suitably capture the respective dimensions and thus quantify the two types of polarization in social networks (RQ1). In the following section, I use these measures to analyze the development of ideological and affective polarization in the debate about Covid-19 restrictions on Twitter (RQ2).

5 Case Study: Twitter Covid-19 debate

5.1 Data Collection and Preprocessing

The data collection is based on TBCOV, a large longitudinal Covid-19 data set spanning tweets from 218 countries and a time frame of 1 February 2020 – 31 March 2021 (Imran et al., 2022). This data set has been used in several studies on Covid-19 before (Chen et al., 2021; Goetz et al., 2022; Jia et al., 2021; Trad & Spiliopoulou, 2021; Zhunis et al., 2022),[1] and it contains tweet IDs and location information related to each tweet. Especially the geo-location information is valuable since the Covid-19 restrictions were confined by national borders and the tweets in the data set should therefore be clearly allocated to a country.

I limit the data collection to tweets that are located in the United States and published from February 2020 to July 2020. Based on this subset of tweet IDs, I collect the full tweet object including the content of the tweet, user information, and other meta data provided by the Twitter Application Programming Interface (API) for each tweet ID specified. It is important to note that the final Twitter data set used here differs from the original TBCOV data set: some users might have deleted some of their tweets or removed their Twitter accounts entirely and this data is therefore not available any longer. Moreover, since the Twitter API only returns publicly available tweets, all messages from users who changed their account settings to private do not appear in the data set.

Next, I identify all English-language tweets using a pre-trained language detection model. I further filter the data set so that the remaining tweets contain at least one keyword related to Covid-19 restrictions. The initial keywords in this list are manually curated and supplemented by semantically similar words which I found by training a word2vec model (for a detailed account of the preprocessing and filtering steps, see the Companion Paper Section 12). The final data set obtained contains approximately 47 million tweets by 4.1 million users.

[1] Note that the studies cited here refer to *GeoCov19* (Qazi et al., 2020), a previous (smaller) version of the TBCOV data set curated by the same authors.

5.2 Ideological User Leaning

To estimate the ideological leaning of the users in the data set, I analyze how often users retweeted posts by political and media accounts. This approach is based on the idea that liberal users are more likely to interact and share posts by liberal political and media accounts, while conservative users prefer conservative sources (Barberá, 2015). The list of political and media accounts, which I will refer to as *baseline accounts*, comprises the Twitter accounts of the US politicians in the 118th US Congress (2019-2021). It is important to note that I only consider accounts that were still available at the time of collecting the data. Since the Twitter account of the former President Donald Trump has been banned, all retweets of Trump's posts are not available any longer. I use the DW-NOMINATE scores (Lewis et al., 2022), an established political science approach to estimating the ideological leaning of US Congress members based on their roll call votes (Poole & Rosenthal, 1997), to assign each politician's account an ideological leaning score between -1 (liberal) to $+1$ (conservative).

Moreover, the list of baseline accounts contains news outlets which I retrieved from mediabias-factcheck.com. This website scores the leaning of news outlets and this data has been used for similar purposes in previous studies (Cinelli et al., 2021). A detailed account of how the news outlets were curated is available in the Companion Paper Section 12. The distribution of baseline scores is shown below (Figure 5.1).

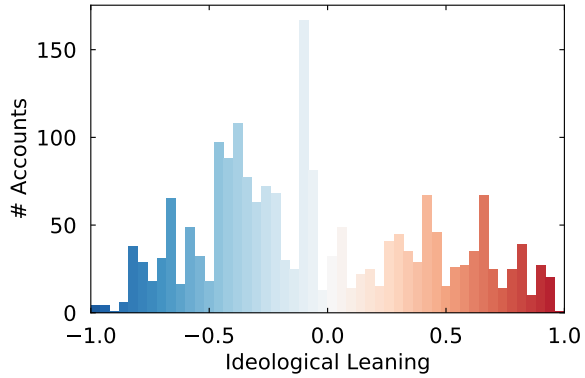


Figure 5.1: Distribution of the ideological leaning of the baseline scores.

To estimate the leaning of the users in the data set, I count how often each user retweets posts by the baseline accounts and I calculate the average of the scores associated with each baseline

account. Moreover, I only consider users who have retweeted at least 5 posts by the baseline accounts in order to ensure that the user scores reflect an actual preference for liberal or conservative accounts and are not just the product of retweeting viral posts. Figure 5.2 shows the distribution of the users that are considered in the remaining analysis. These results are robust regardless of whether the threshold is set at 2 or more retweets (see Companion Paper Section 13), but it is important to note that this approach only allows me to classify approximately 104,000 users (2.54% of the entire user base in the sample).

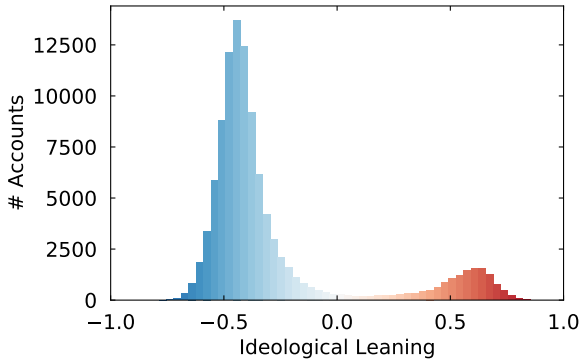


Figure 5.2: Distribution of the user scores for users who retweeted at least 5 different baseline accounts.

As the histogram shows, the distribution of ideological leanings is skewed in this sample as there are considerably more liberals than conservatives. This finding is consistent with other analyses on the ideological divide in the US-based pandemic debate on Twitter (Lazer et al., 2020), but the results presented below should nevertheless be treated with caution and interpreted given the fact that the distribution is skewed.

5.3 Hostility Classification

As outlined in the literature review, all prior studies investigating affective polarization on social media have used sentiment as a measure of favorability or hostility. I diverge from this approach since the sentiment of social media posts is heavily influenced by the topics that users discuss which is particularly problematic in the Covid-19 debate. The sentiment of pandemic-related posts is overall very negative because users discuss topics such as death, disease, isolation, and mental health issues (this is e.g. shown in Imran et al., 2022). The negative sentiment is given by

the somber topic of conversation, but it should not be interpreted as a sign of out-group hostility or affective polarization.

Instead, I choose to use offensive language as an indicator for hostility in user interactions. For this classification task, I use a RoBERTa-base model that was re-trained on approximately 58 million tweets and fine-tuned for the detection of offensive language by Barbieri et al. (2020). The details of the classification approach are outlined in the Companion Paper Section 14. Figure 5.3 shows the distribution of offensive language tweets from February 2020 to July 2020. In the first few weeks in February, the fraction of offensive tweets is very small and it increases to approximately 15% in the following weeks.

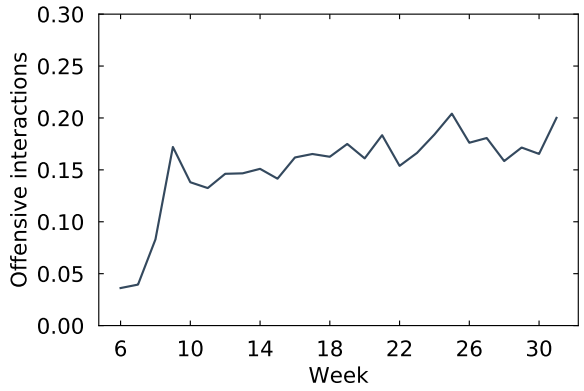


Figure 5.3: Fraction of offensive tweets from February 2020 to July 2020.

5.4 Networks

To observe the development of ideological and affective polarization over time, I split the data set into subsets. Each subset spans a week from Monday to Sunday in order to avoid any weekday or weekend effects. From these subsets, I extract all replies and mentions between users for which there are ideological leaning scores and I build undirected networks based on that. Since there can be more than one reply or mention between two users, I calculate the average offensiveness of all their interactions and use those as edge attributes. As outlined in the Companion Paper Section 14, the results are robust to the way that the offensiveness scores of multiple edges between one node pair are summarized. I discard the directionality of the edges because the measures used here quantify the ideological and affective polarization as aggregates over the

entire network. In other words, I am interested in how frequently and how hostile liberal and conservative users interact at large, but in this macro-level view it does not make a difference whether it is a liberal user addressing a conservative user or vice versa. Once I have created the network for each time frame, I extract the largest connected component because the ideological polarization metric requires a connected graph.

5.5 Results: Ideological and Affective Polarization on Twitter

The results are summarized in Figure 5.4 below which shows the development of ideological and affective polarization in the Covid-19 debate on Twitter over the course of half a year. Below some exemplary visualizations of the network structures encountered in the Twitter sample, the second row visualizes how the Generalized Euclidean Polarization Index evolves over time. As argued above, this ideological polarization measure (y-axis) shows the extent to which liberals and conservatives reply to and mention each other. If there are many interactions between users of opposing leanings and users are thus exposed to a range of views, the score is low (the minimum is 0). If, however, there are only few interactions between disagreeing users and they mainly interact with like-minded others in homogeneous communities, the ideological polarization measure is high (the maximum is an arbitrary positive value).

As expected, the ideological polarization is low throughout February 2020 since the Covid-19 regulations were not a widely discussed topic yet. The ideological polarization increases from late February to mid-April 2020. At this time, all states reported widespread cases of Covid-19 and the US became the country with the highest number of Covid-19-related deaths in the world (Centers for Disease Control and Prevention, 2022). This peak in ideological polarization also coincides with a change in policy regarding face masks. In April 2020, the Centers for Disease Control and Prevention started recommending the use of face masks as a preventive measure although officials had discouraged wearing face masks until that point (Centers for Disease Control and Prevention, 2022). To investigate whether this topic is mirrored in the Twitter discussion, I analyse how often the users referred to the Covid-19 restriction keywords that

I used to filter the data set (see Companion Paper Section 12 for a list of all keywords). This analysis shows that the most frequently occurring terms in weeks 15–17 are *mask* and *test* and I conclude that the Twitter users in the sample indeed discussed the face mask policy in spring 2020. After the first peak in April 2020, the ideological polarization score decreases slightly, and then increases again to a high level in July 2020.

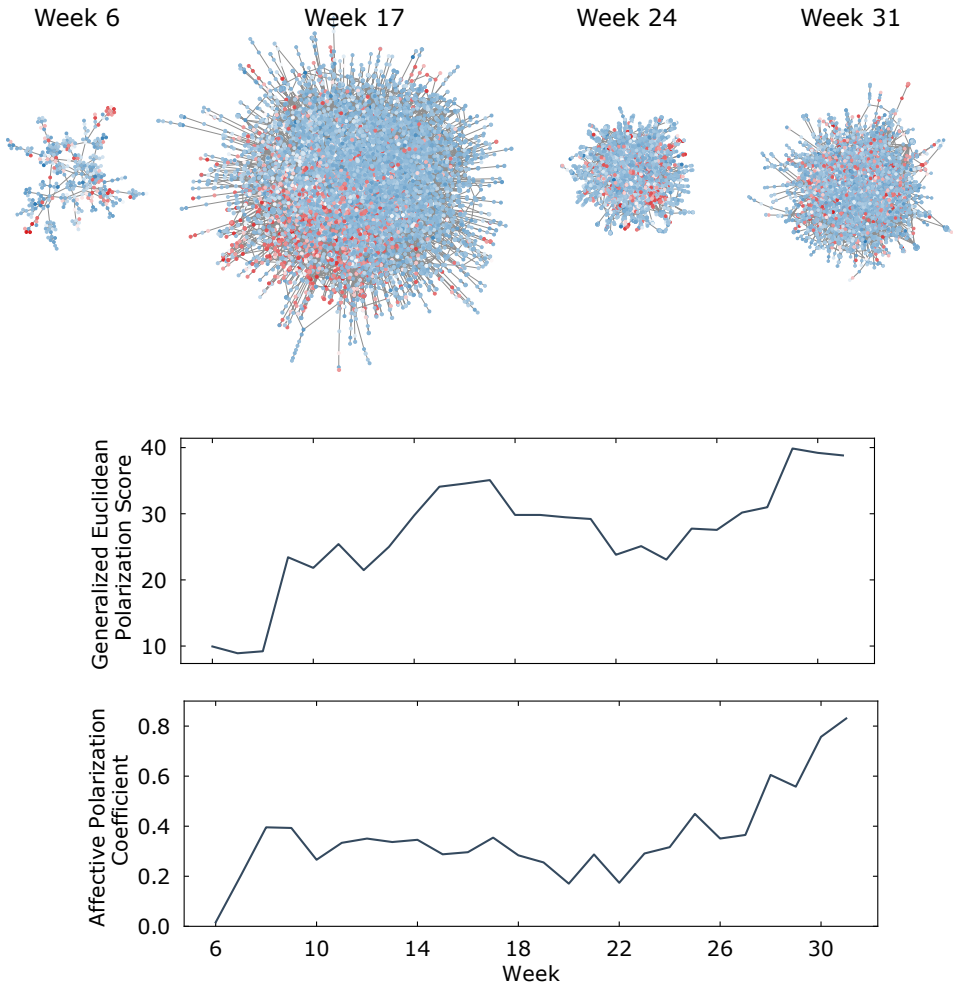


Figure 5.4: Ideological and affective polarization in the debate about Covid-19 restrictions on Twitter. The first row shows four examples of network topologies; node color reflects ideological leaning (blue for liberals, red for conservatives). The networks are visualized using a force-directed layout algorithm. The line plot in the second row summarizes the Generalized Euclidean Polarization Index and the third row shows the Affective Polarization Coefficient.

The third row in Figure 5.4 shows the development of the Affective Polarization Coefficient in the Twitter sample. To recap, this score captures the relation between ideological differences and the use of offensive language towards other users. If there is no relation, i.e., users speak similarly (un)offensive to others regardless of whether they share the same ideological leaning or not, then

the score is 0. If there is a perfect relation, that is like-minded users only use unoffensive language and disagreeing users only use offensive language to address each other, then the score is 1.

According to the measure proposed in this paper, there is no affective polarization in week 6. This is to be expected since the share of offensive tweets was smaller than 5% in this week and all users therefore communicated with each other using unoffensive language regardless of whom they talked to. This changes in early March 2020 when the Affective Polarization Coefficient increases to approximately 0.3. The score remains at this level until June 2020 when it starts increasing considerably more. According to this measurement, affective polarization was very high throughout June and especially July 2020 as the highest score of approximately 0.8 is reported in the last week in July. The keyword analysis shows that, apart from the terms *test* and *mask*, the keyword *school* is among the three most frequently used words in the sample at this point in time. This coincides with an intense public debate about whether or not to reopen schools after the summer break (Grossmann et al., 2021), and the results presented here might be an indicator that Twitter users engaged in a heated debate online too.

Lastly, it is important to note that the fluctuations in the ideological and affective polarization scores are not merely the result of differences in the size of the networks. As I show in the Companion Paper Section 8, both measures do not take the number of nodes in the network into account and they are thus size-independent.

6 Discussion

In this paper, I set out to quantify ideological and affective polarization in a social network and apply those measures to a case study of the Covid-19 debate on Twitter. The contribution of this project lies in the theoretical, methodological and analytical work presented here. To begin, I derive a theoretical conceptualization and distinction of ideological and affective polarization from the literature. These insights are then used to select and discuss a network-based measure to quantify ideological polarization. Furthermore, I propose a new network-based affective polarization measure that addresses some shortcomings of other methods currently in use.

I conclude that the Generalized Euclidean Polarization Index and the Affective Polarization Coefficient are suitable measures to quantify ideological and affective polarization in social networks (RQ1).

Nevertheless, the measures presented here are subject to several limitations that could be addressed in future studies. First, both measures rely on one-dimensional ideological scaling. When applied to other countries than the United States, a different ideological scale that, for example, accounts for a multi-party system would be needed. Second, the affective dimension quantified here solely focuses on hostility. There are other types of social ties such as trust, respect or support which could be explored instead.

While the synthetic experiments show that the Affective Polarization Index captures all relevant dimensions in a controlled setting, further verification of the measure and how it behaves when applied to real-world social media data is still needed. This question could be tackled by analyzing social media samples collected during election versus non-election times. Since elections are moments of intense partisan conflict, these events are likely to be accompanied by increased levels of affective polarization (Hansen & Kosiara-Pedersen, 2017; Hernández et al., 2021).

A second contribution of this study concerns the social media data analyzed as I collect and examine a large-scale data set on Twitter interactions surrounding the Covid-19 debate in the United States. The analysis leverages state-of-the-art computational methods, in particular natural language processing using transfer learning and algorithms for network analysis, to generate insights about the ideological and affective polarization during the Covid-19 pandemic (RQ2). I find that the levels of both ideological and affective polarization were low in early February and then increased during the subsequent weeks. The affective polarization measure shows moderately high scores throughout spring 2020 and then increases starkly in June and throughout July; a development that coincided with a heated debate about school reopenings in the United States. For ideological polarization, I observe a peak in mid-April, a time when the policy regarding the use of face masks changed. There are further interesting questions that the present study does not answer; for instance, the review of literature on political polarization could not conclusively clarify how ideological and affective polarization are related. The data

used here is not suited to approach this question, but future studies should shed light on the relation between the two types of polarization on social media.

Lastly, there are limitations associated with the data analysis. Especially the skewed distribution of user leanings calls into question whether the sample collected here is a valid snapshot of the Covid-19 debate on Twitter. The strong liberal leaning of the sample is most likely due to the keyword list which was used as part of the TBCOV data set, based on which I collected the sample. This list comprises keywords such as *Covid-19* or *coronavirus* which are terms used by people who acknowledge that the pandemic is a real and serious public health emergency. It follows that tweets by users who ridicule or reject this idea are not included in the data set. This issue could be addressed in future work by including more keywords and hashtags that are used by pandemic-skeptic individuals.

Moreover, due to the different preprocessing and filtering steps, only a small fraction of users and tweets are considered in the final analysis of ideological and affective polarization. This problem is common to many studies of polarization on social media, and further insight is needed on how representative the results are for the social media user base at large. While it is important to not overinterpret platform-related and sample-specific results, the findings nevertheless complement previous studies well and they represent a first, exploratory approach to ideological and affective polarization in the pandemic debate on Twitter.

7 References

Abramowitz, A. I., & Saunders, K. L. (2008). Is polarization a myth? *The Journal of Politics*, *70*(2), 542–555. https://doi.org/10.1017/S0022381608080493

Ahler, D. J., & Sood, G. (2018). The parties in our heads: Misperceptions about party composition and their consequences. *The Journal of Politics*, *80*(3), 964–981. https://doi.org/10.1086/697253

Baldassarri, D., & Page, S. E. (2021). The emergence and perils of polarization. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(50). https://doi.org/10.1073/pnas.2116863118

Barberá, P. (2015). Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political Analysis*, *23*(1), 76–91. https://doi.org/10.1093/pan/mpu011

Barbieri, F., Camacho-Collados, J., Espinosa Anke, L., & Neves, L. (2020). TweetEval: Unified benchmark and comparative evaluation for tweet classification. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1644–1650. https://doi.org/10.18653/v1/2020.findings-emnlp.148

Bernacer, J., García-Manglano, J., Camina, E., & Güell, F. (2021). Polarization of beliefs as a consequence of the COVID-19 pandemic: The case of Spain. *PLOS ONE*, *16*(7), 1–22. https://doi.org/10.1371/journal.pone.0254511

Bougher, L. D. (2017). The Correlates of Discord: Identity, Issue Alignment, and Political Hostility in Polarized America. *Political Behavior*, *39*(3), 731–762. https://doi.org/10.1007/s11109-016-9377-1

Bruine de Bruin, W., Saw, H. W., & Goldman, D. P. (2020). Political polarization in US residents' COVID-19 risk perceptions, policy preferences, and protective behaviors. *Journal of Risk and Uncertainty*, *61*, 177–194. https://doi.org/10.1007/s11166-020-09336-3

Carothers, T., & O'Donohue, A. (2019). Introduction. In T. Carothers & A. O'Donohue (Eds.), *Democracies divided* (pp. 1–13). Brookings Institution Press.

Centers for Disease Control and Prevention. (2022). CDC museum COVID-19 timeline. https://www.cdc.gov/museum/timeline/covid19.html

Chen, Q., Wang, W., Huang, K., & Coenen, F. (2021). Zero-shot text classification via knowledge graph embedding for social media data. *IEEE Internet of Things Journal*, 1–1. https://doi.org/10.1109/JIOT.2021.3093065

Cinelli, M., De, G., Morales, F., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The echo chamber effect on social media. https://doi.org/https://doi.org/10.1073/pnas.2023301118

Coscia, M. (2021). Pearson correlations on complex networks. *Journal of Complex Networks*, *9*(6). https://doi.org/10.1093/comnet/cnab036

Dias, N., & Lelkes, Y. (2022). The nature of affective polarization: Disentangling policy disagreement from partisan identity. *American Journal of Political Science*. https://doi.org/https://doi.org/10.1111/ajps.12628

Druckman, J. N., Klar, S., Krupnikov, Y., Levendusky, M., & Ryan, J. B. (2021a). Affective polarization, local contexts and public opinion in America. *Nature Human Behavior*, 28–38. https://doi.org/https://doi.org/10.1038/s41562-020-01012-5

Druckman, J. N., Klar, S., Krupnikov, Y., Levendusky, M., & Ryan, J. B. (2021b). How affective polarization shapes americans' political beliefs: A study of response to the COVID-19 pandemic. *Journal of Experimental Political Science*, *8*(3), 223–234. https://doi.org/10.1017/XPS.2020.28

Druckman, J. N., & Levendusky, M. S. (2019). What do we measure when we measure affective polarization? *Public Opinion Quarterly*, *83*(1), 114–122. https://doi.org/10.1093/poq/nfz003

Fiorina, M. P., & Abrams, S. J. (2008). Political polarization in the American public. *Annual Review of Political Science*, *11*(1), 563–588. https://doi.org/10.1146/annurev.polisci.11.053106.153836

Garimella, K., Morales, G. D. F., Gionis, A., & Mathioudakis, M. (2018). Quantifying controversy on social media. *Trans. Soc. Comput.*, *1*(1). https://doi.org/10.1145/3140565

Goetz, S. J., Heaton, C., Imran, M., Pan, Y., Tian, Z., Schmidt, C., Qazi, U., Ofli, F., & Mitra, P. (2022). Food insufficiency and Twitter emotions during a pandemic. *Applied Economic Perspectives and Policy*. https://doi.org/https://doi.org/10.1002/aepp.13258

Green, J., Edgerton, J., Naftel, D., Shoub, K., & Cranmer, S. J. (2020). Elusive consensus: Polarization in elite communication on the COVID-19 pandemic. *Science Advances*, *6*(28). https://doi.org/10.1126/sciadv.abc2717

Grossmann, M., Reckhow, S., Strunk, K. O., & Turner, M. (2021). All states close but red districts reopen: The politics of in-person schooling during the covid-19 pandemic. *Educational Researcher*, *50*(9), 637–648. https://doi.org/10.3102/0013189X211048840

Hansen, K. M., & Kosiara-Pedersen, K. (2017). How campaigns polarize the electorate: Political polarization as an effect of the minimal effect theory within a multi-party system. *Party Politics*, *23*(3), 181–192. https://doi.org/10.1177/1354068815593453

Hernández, E., Anduiza, E., & Rico, G. (2021). Affective polarization and the salience of elections. *Electoral Studies*, *69*, 102203. https://doi.org/https://doi.org/10.1016/j.electstud.2020.102203

Hohmann, M., Devriendt, K., & Coscia, M. (2022). Quantifying political polarization on a network using Generalized Euclidean distance [manuscript submitted for publication].

Imran, M., Qazi, U., & Ofli, F. (2022). TBCOV: Two billion multilingual COVID-19 tweets with sentiment, entity, geo, and gender labels. *Data*, *7*(1). https://doi.org/10.3390/data7010008

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (2019). The origins and consequences of affective polarization in the United States. https://doi.org/https://doi.org/10.1146/annurev-polisci-051117-073034

Iyengar, S., Sood, G., & Lelkes, Y. (2012). Affect, not ideology: A social identity perspective on polarization. *Public Opinion Quarterly*, *76*(3), 405–431. https://doi.org/10.1093/poq/nfs038

Jia, S., Chen, Q., & Wang, W. (2021). Covid-19 tweets analysis with topic modeling. *4th International Conference on Computing and Big Data*, 68–74. https://doi.org/10.1145/3507524.3507536

Jiang, J., Chen, E., Yan, S., Lerman, K., & Ferrara, E. (2020). Political polarization drives online conversations about COVID-19 in the United States. *Human Behavior and Emerging Technologies*, *2*(3), 200–211. https://doi.org/https://doi.org/10.1002/hbe2.202

Jiang, X., Su, M.-H., Hwang, J., Lian, R., Brauer, M., Kim, S., & Shah, D. (2021). Polarization over vaccination: Ideological differences in Twitter expression about COVID-19 vaccine favorability and specific hesitancy concerns. *Social Media + Society*, *7*(3). https://doi.org/10.1177/20563051211048413

Kerr, J., Panagopoulos, C., & van der Linden, S. (2021). Political polarization on COVID-19 pandemic response in the United States. *Personality and Individual Differences*, *179*. https://doi.org/https://doi.org/10.1016/j.paid.2021.110892

Kubin, E., & von Sikorski, C. (2021). The role of (social) media in political polarization: a systematic review. *Annals of the International Communication Association*, *45*(3), 188–206. https://doi.org/10.1080/23808985.2021.1976070

Lang, J., Erickson, W. W., & Jing-Schmidt, Z. (2021). #Maskon! #maskoff! Digital polarization of mask-wearing in the United States during COVID-19. *PLOS ONE*, *16*(4), 1–25. https://doi.org/10.1371/journal.pone.0250817

Lazer, D., Ruck, D. J., Quintana, A., Shugars, S., Joseph, K., Grinberg, N., Gallagher, R. J., Horgan, L., Gitomer, A., Bajak, A., A., M., Ognya, K., Qu, H., H, W. R., McCabe, S., & Green, J. (2020). The state of the nation: A 50-state COVID-19 survey report #18: COVID-19 fake news on Twitter.

Lelkes, Y. (2021). Policy over party: Comparing the effects of candidate ideology and party on affective polarization. *Political Science Research and Methods*, *9*(1), 189–196. https://doi.org/10.1017/psrm.2019.18

Lewis, J. B., Poole, K., Rosenthal, H., Boche, A., Rudkin, A., & Sonnet, L. (2022). Voteview: Congressional roll-call votes database. https://voteview.com/

Marchal, N. (2022). "Be nice or leave me alone": An intergroup perspective on affective polarization in online political discussions. *Communication Research*, *49*(3), 376–398. https://doi.org/10.1177/00936502211042516

Mason, L. (2016). A cross-cutting calm: How social sorting drives affective polarization. *Public Opinion Quarterly*, *80*(S1), 351–377. https://doi.org/10.1093/poq/nfw001

Mentzer, K., Fallon, K., Prichard, J., & Yates, D. J. (2020). Measuring and unpacking affective polarization on Twitter: The role of party and gender in the 2018 Senate races. *Proceedings of the 53rd Hawaii International Conference on System Sciences*, 2459–2468.

Mønsted, B., & Lehmann, S. (2022). Characterizing polarization in online vaccine discourse: A large-scale study. *PLOS ONE*, *17*(2). https://doi.org/10.1371/journal.pone.0263746

Orr, L. V., & Huber, G. A. (2020). The policy basis of measured partisan animosity in the United States. *American Journal of Political Science*, *64*(3), 569–586. https://doi.org/https://doi.org/10.1111/ajps.12498

Park, J., Warner, B. R., McKinney, M. S., Kearney, C., Kearney, M. W., & Kim, G.-E. (2021). Partisan identity and affective polarization in presidential debates. *American Behavioral Scientist*. https://doi.org/10.1177/00027642211046551

Pennycook, G., McPhetres, J., Bago, B., & Rand, D. G. (2021). Beliefs about COVID-19 in Canada, the United Kingdom, and the United States: A novel test of political polarization and motivated reasoning. *Personality and Social Psychology Bulletin*. https://doi.org/10.1177/01461672211023652

Poole, K. T., & Rosenthal, H. L. (1997). *Congress: A political-economic history of roll call voting*. Oxford University Press.

Qazi, U., Imran, M., & Ofli, F. (2020). GeoCoV19: A dataset of hundreds of millions of multilingual COVID-19 tweets with location information. https://crisisNLP.qcri.org/covid19

Reiter-Haas, M., Klösch, B., Hadler, M., & Lex, E. (2022). Polarization of opinions on COVID-19 Measures: Integrating Twitter and survey data. *Social Science Computer Review*. https://doi.org/10.1177/08944393221087662

Simchon, A., Brady, W. J., Bavel, J. J. V., & Mutz, D. (2022). Troll and divide: The language of online polarization. *PNAS Nexus, 1*(1). https://www.pnas.org/doi/abs/10.1093/pnasnexus/pgac019

Trad, R., & Spiliopoulou, M. (2021). Juxtaposing 5G coronavirus tweets with general coronavirus tweets during the early months of coronavirus outbreak. *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, 597–602. https://doi.org/10.1109/CBMS52027.2021.00107

Tyagi, A., Uyheng, J., & Carley, K. (2021). Heated conversations in a warming world: Affective polarization in online climate change discourse follows real-world climate anomalies. *Social Networks Analysis and Mining, 11*(87). https://doi.org/10.1007/s13278-021-00792-6

Vandeweerdt, C. (2021). Someone like you: False consensus in perceptions of Democrats and Republicans. *Journal of Elections, Public Opinion and Parties*, 1–11. https://doi.org/10.1080/17457289.2021.1942891

Webster, S. W., & Abramowitz, A. I. (2017). The Ideological Foundations of Affective Polarization in the U.S. Electorate. *American Politics Research, 45*(4), 621–647. https://doi.org/10.1177/1532673X17703132

Wilson, A. E., Parker, V., & Feinberg, M. (2020). Polarization in the contemporary political and media landscape. https://doi.org/10.1016/j.cobeha.2020.07.005

Yarchi, M., Baden, C., & Kligler-Vilenchik, N. (2021). Political polarization on the digital sphere: A cross-platform, over-time analysis of interactional, positional, and affective polarization on social media. *Political Communication, 38*(1-2), 98–139. https://doi.org/10.1080/10584609.2020.1785067

Zhunis, A., Lima, G., Song, H., Han, J., & Cha, M. (2022). Emotion bubbles: Emotional composition of online discourse before and after the COVID-19 Outbreak. *Proceedings of the ACM Web Conference 2022*, 2603–2613. https://doi.org/10.1145/3485447.3512132

Part II

COMPANION PAPER

The companion paper outlines additional technical details and analyses which are referred to in the main article. I first describe how the Generalized Euclidean Polarization Index and the Affective Polarization Coefficient are formally defined and I show that the measures are scale invariant. Next, I provide formal definitions of the other alternative polarization measures and I describe how the synthetic network data in the experiments is generated. The subsequent sections detail the Twitter data collection and preprocessing steps, and I present robustness checks of the analysis presented in the main article. Summary statistics of the Twitter networks used in the analysis conclude the companion paper.

8 Network Polarization Measures

8.1 Generalized Euclidean Polarization Index

The Generalized Euclidean Polarization Index, which I will refer to as $\delta_{G,o}$, requires two inputs (Hohmann et al., 2022): First, a network $G = (V, E)$ with the set of nodes defined as V and the

set of edges E. The network needs to be connected and it cannot contain any edges connecting a node with itself (self-loops). Second, the measure requires the ideological leaning of each user as a numeric value between $[-1, 1]$. In the case study above, -1 indicates extremely liberal Twitter users, 0 indicates moderates, and $+1$ indicates extremely conservative users. The leanings are represented in a vector o which is split into two vectors of length $|V|$: on the one hand, there is o^+ which contains all positive opinions and zero otherwise and on the other hand, there is o^- which contains the absolute value of all negative opinions and zero otherwise.

Given these two inputs, the Generalized Euclidean Polarization Index $\delta_{G,o}$ quantifies the ideological polarization of a social network by returning a numeric value. This value is in between 0 (no polarization) and an arbitrary positive number (the higher this number, the higher the polarization in the network). The measure is defined as (Hohmann et al., 2022):

$$\delta_{G,o} = \sqrt{(o^+ - o^-)^T L^\dagger (o^+ - o^-)}$$

where o^+ is the positive opinion vector and o^- is the negative opinion vector (as outlined above), and $L^\dagger$ is the pseudoinverse of the Laplacian matrix of G. The Laplacian matrix captures the structure of the graph as a matrix representation. It is formally defined as $L = D - A$ where A is the adjacency matrix of G (a matrix showing which nodes are connected by an edge) and D is the degree matrix of G (a matrix capturing the number of edges of each node). This results in an $n \times n$ matrix which contains each node's degree on the diagonal, -1 if two nodes i and j are connected, and 0 otherwise. The Generalized Euclidean Polarization index uses the Moore-Penrose pseudoinverse $L^\dagger$ of this Laplacian matrix.

8.2 Affective Polarization Coefficient

The Affective Polarization Coefficient is based on the recently introduced network correlation which requires three inputs (Coscia, 2021): a network $G = (V, E)$ and two vectors x and y. The

measure uses those inputs to calculate a correlation between the vectors x and y while taking the structure of the network into account.

The network correlation can only calculate correlations for node attributes which poses a problem since the ideological difference and hostility between two nodes are modelled as edge attributes in the networks I analyze here. To solve this problem, I use the line graph representation which turns all edges and their attributes into nodes and associated node attributes. The line graph is an established network science method to, for instance, find communities among edges rather than nodes (Deng et al., 2015; Evans & Lambiotte, 2009). Each edge in G is represented as a node in the line graph G' and each node that is common to two edges in G is represented by an edge in G' (see Figure 8.1).

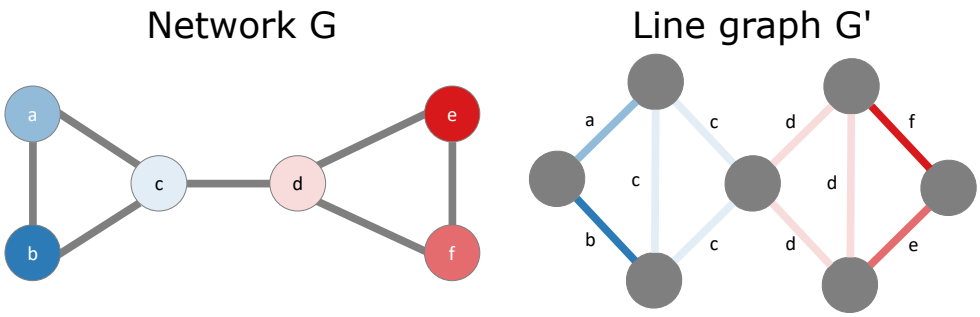


Figure 8.1: Example of a network G and its line graph G'.

The Affective Polarization Coefficient takes the line graph G', a vector x specifying the ideological difference between each pair of users, and a vector y recording the hostility between user pairs. From this line graph, I retrieve a weight matrix W that contains the exponentiated shortest path distances for each node pair in G'. Moreover, I calculate the means $\bar{x}$ and $\bar{y}$ of the respective vectors and then determine their centred versions; i.e., $\hat{x} = (x - \bar{x})$ and $\hat{y} = (y - \bar{y})$. Next, I calculate the (network) standard deviation as $\sigma_{x,W} = \sqrt{\sum W \times (\hat{x} \otimes \hat{x})}$ and $\sigma_{y,W} = \sqrt{\sum W \times (\hat{y} \otimes \hat{y})}$.

The measure is defined as:

$$\rho_{x,y,G'} = \frac{\sum W \times (\hat{x} \otimes \hat{y})}{\sigma_{x,W}\sigma_{y,W}}$$

where $\hat{x}$ and $\hat{y}$ are the centred vectors, W is the weight matrix, $\times$ is an element-wise product operation, and $\otimes$ is the outer product operation.

8.3 Scale Invariance

I argue that the ideological and affective polarization measures are scale invariant. In other words, they report the same value for two networks with the same topology even if the number of nodes in the two networks are different. To verify this property, I generate a network with two separated cliques, a blue one and a red one, which are connected by a few edges (see Figure 8.2). The ideological leaning values are randomly drawn from a uniform distribution. The edges *between* the cliques are always 5% of the edges *within* each of the cliques. I then grow the network by adding nodes and edges within the cliques, and adjusting the number of edges between the cliques. I generate hostility values that are highly aligned with the ideological difference between two nodes. For all edges within the blue and red clique, the hostility values are low, while I assign a high hostility value to each (newly added) edge between the cliques.

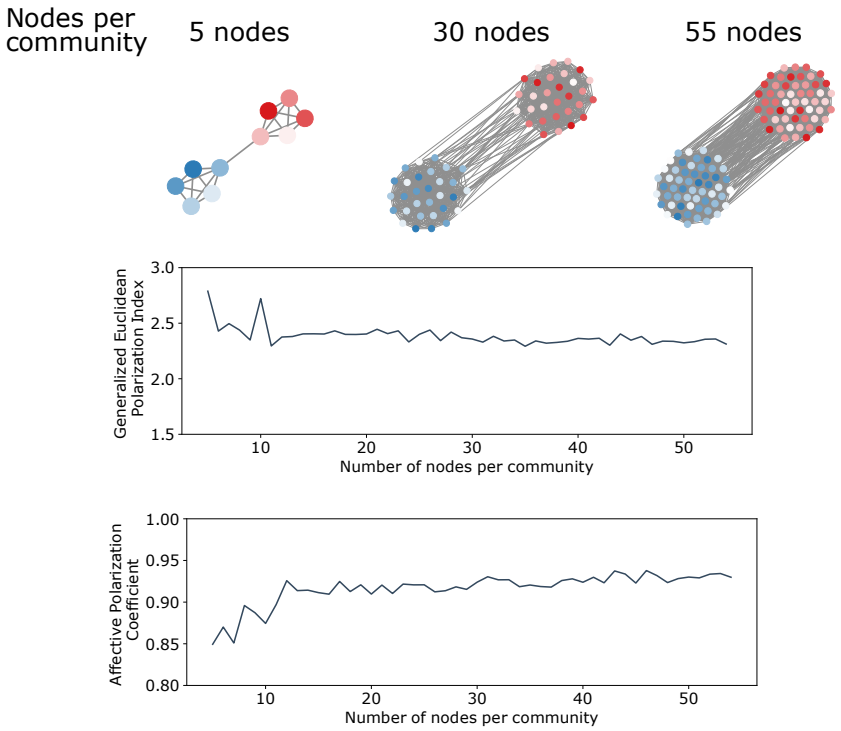


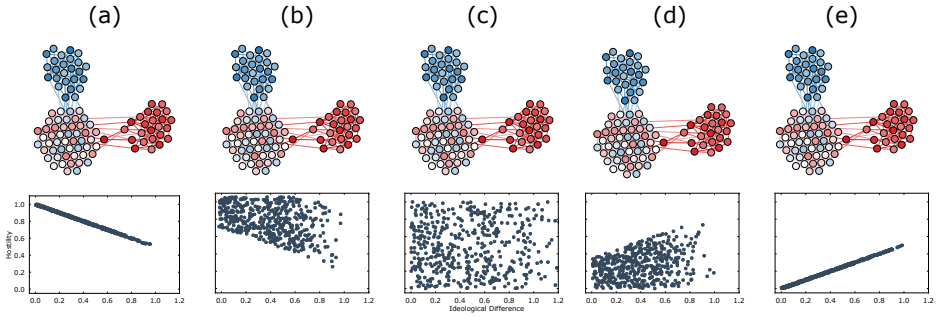
Figure 8.2: Scale invariance experiment. Exemplary network visualizations (first row), Generalized Euclidean Polarization Index (second row), and Affective Polarization Coefficient (third row row) for topologically similar networks of different sizes.

Figure 8.2 shows that as the network grows, the scores approach a limit. This is to be expected since the Generalized Euclidean Polarization Index is a relative measure of the distance between node attributes and the Affective Polarization Coefficient is derived from the scale invariant

Pearson correlation. I conclude that the measures are indeed scale invariant for networks with similar topologies but different sizes. This is an important property since I can rule out the possibility that the results of the Twitter data analysis are due to the number of nodes in the networks.

8.4 Relation Between the Polarization Measures

Since I apply both measures to the same Twitter networks, it is important to investigate whether variations in one measure determine variations in the other measure by construction. I argue that the two measures are unrelated to each other even when they are calculated on the same network with the same node leanings.



	(a)	(b)	(c)	(d)	(e)
Affective Polarization Coefficient	-1.0	-0.74	0.0	0.74	1.0
Generalized Euclidean Polarization Index	3.77	3.77	3.77	3.77	3.77

Figure 8.3: Relation between the Affective Polarization Coefficient and the Generalized Euclidean Polarization Index. The first row shows the network topology where the node color reflects opinion (blue as liberals, red as conservatives). The scatter plots in the second row show how ideological difference and hostility are related across all node pairs. The rows below report results of 100 iterations of the experiment.

To verify this claim, I repeat the experiments in the Section 4.4 in the main article. As Figure 8.3 shows, the network structure is the same from (a)-(e) but the correlation between the ideological difference and the hostility changes. As expected, the Generalized Euclidean Polarization Index and the Affective Polarization Coefficient are entirely unrelated. While the former one is the same for all networks due to the fixed topology, the latter one changes according to the correlation reported in the scatter plots. This confirms that the values reported by one measure are not determined by the other measure.

9 Alternative Polarization Measures

9.1 Assortativity Coefficient

The Assortativity Coefficient measures whether node pairs share similar values with respect to a given attribute (Newman, 2003). As a measure of ideological polarization, the Assortativity Coefficient captures if nodes are surrounded by other like-minded nodes (Mønsted & Lehmann, 2022).

For each node in a node pair, the values of the given attribute are recorded as x (for the first node) and y (for the second node). I summarize these value pairs (x, y) in a matrix e that captures the fraction of edges connecting the values x and y in the network in each entry e_{xy}. To put it differently, the matrix shows how often the different combinations of the values x and y appear in the network relative to the total number of edges in the network. Furthermore, I calculate the row sums, i.e., the fraction of edges starting from a node with value x, as a_x ($\sum_y e_{xy} = a_x$) and the column sums, i.e., the fraction of edges arriving in a node with value y, as b_y ($\sum_x e_{xy} = b_y$).

The Assortativity Coefficient is then defined as the Pearson correlation between x and y (Newman, 2003):

$$\rho_{G,o} = \frac{\sum_{x,y} xy(e_{xy} - a_x b_y)}{\sigma_a \sigma_b}$$

where σ_a and σ_b are the standard deviations of a_x and b_y. The measure can return values between -1 (perfectly disassortative network) to $+1$ (perfectly assortative network).

9.2 Random Walk Controversy

The Random Walk Controversy measure quantifies how easily the boundary between two communities in a network can be crossed. When applied to ideological polarization, this measure can show how closely connected two communities in the network are.

The measure partitions the graph into two communities using a community discovery algorithm. In the experiments presented in the main article, I use the Kernighan-Lin algorithm (Kernighan & Lin, 1970). Once the two communities C_1 and C_2 are identified, the measure simulates several thousand random walks[1] for which half of the walkers start in community C_1 and the other half start in C_2. From each community, 10% of nodes are randomly drawn and saved in two community-specific subsets. Each walk ends as soon as the walker reaches a node in either of those subsets.[2] The results of the random walks are summarized in four probabilities p_{C_1,C_2}, p_{C_2,C_1}, p_{C_1,C_1}, and p_{C_2,C_2} that capture in which community the random walker started and ended. For instance, p_{C_1,C_2} is the probability of the walker starting in C_1 and ending in C_2. The Random Walk Controversy measure is then calculated as (Garimella et al., 2018):

$$RWC_G = p_{C_1,C_1}p_{C_2,C_2} - p_{C_1,C_2}p_{C_2,C_1}$$

This measure ranges from -1 (all random walkers end in the other community) to $+1$ (all random walkers end within their own community).

9.3 Pearson Correlation Coefficient

The Pearson Correlation Coefficient captures the linear relationship between two variables x and y. In the case of affective polarization, the correlation between ideological difference and

[1] In particular, the measure randomly chooses 10% of the nodes in the network, simulates random walks for all of the nodes within the subset, and repeats this process 1,000 times.

[2] Note that in the paper itself, Garimella et al. (2018) argue that their measure stops as soon as the random walker reaches a node with high degree in either community. Since the code implementation made available by the authors works slightly differently, I describe what the code does here.

hostility is indicative of the in-group versus out-group hostility. To calculate the measure, the mean value $\bar{x}$ and $\bar{y}$ of the two variables is obtained. The sample Pearson correlation coefficient is then defined as (Hogg et al., 2019):

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y}$$

where x_i and y_i are sample points of the variables, $\bar{x}$ and $\bar{y}$ are the sample means, and σ_x as well as σ_y are the standard deviations of x and y.

9.4 Earth Mover's Distance

I also compare my measure to the affective polarization measure proposed by Tyagi et al. (2021) which is based on the Earth Mover's Distance, a measure of the distance between two distributions, as well as a network measure named Krackhardt's E/I index. The premise of this measure is that the nodes in the network can be divided into two stance groups k and k'; e.g., climate change believers and disbelievers in the case of Tyagi et al. (2021). In the experiments presented in the main article, I use 0 as a threshold to split the nodes into a liberal (< 0) and a conservative group (> 0). Next, the weight of each edge w_{ij} is determined by the hostility of the user interaction. In particular, w_{ij}^{+} denotes all *non-hostile* edge weights, and w_{ij}^{-} denotes all *hostile* edge weights.[3]

The measure consists of two components that are determined separately, the valence (sign) of the affective polarization score and its magnitude. Moreover, it is important to note that the measure is calculated for each stance group k separately.

[3] Tyagi et al. (2021) rely on a sentiment analysis to estimate sentiment values between ± 1 as an indicator for hostility. In their measure, they therefore consider positive (> 0) and negative (< 0) interactions. Since the hostility values I generate in the experiments are between $[0, 1]$, I change the measure to non-hostile (< 0.5) and hostile interactions (> 0.5).

Affective polarization valence. To retrieve the sign of the score, the E/I index for the subgraph containing all non-hostile interactions G^+ is calculated as follows:

$$P_k^+ = \frac{E_k^+ - I_k^+}{E_k^+ + I_k^+}$$

where E_k^+ is the sum of all non-hostile *out-group* edges and I_k^+ is the sum of all non-hostile *in-group* edges. Similarly, the E/I index for the subgraph of all hostile interactions G^- is calculated as:

$$P_k^- = \frac{E_k^- - I_k^-}{E_k^- + I_k^-}$$

where E_k^- is the sum of all hostile *out-group* edges and I_k^- is the sum of all hostile *in-group* edges. Finally, the sign of the affective polarization score is determined by:

$$P_k = \frac{P_k^- - P_k^+}{2}$$

If P_k results in a positive value, then the out-group interactions are disproportionally hostile and affective polarization is therefore high, while values close to 0 indicate low levels of affective polarization. Lastly, negative values of P_k indicate that the in-group interactions are especially hostile.

Affective polarization magnitude. To determine the extent to which a network is affectively polarized, the measure considers the distribution of out-group hostility u_k and the distribution of in-group hostility v_k. Next, it calculates the Earth Mover's Distance to determine the difference between these two distributions. Finally, Tyagi et al. (2021) define their measure l_k as:

$$l_k = \begin{cases} -\int_{-\infty}^{+\infty} |U_k - V_k| & : P_k < 0 \\ \int_{-\infty}^{+\infty} |U_k - V_k| & : P_k \geq 0 \end{cases}$$

where U_k and V_k are the cumulative distribution functions of u_k and v_k respectively. Intuitively, the Earth Mover's Distance captures how different the in-group hostility values are compared to the out-group hostility values. Moreover, P_k determines the sign of the final measure: if $P_k \geq 0$

then the measure has a positive sign which indicates that affective polarization is high. Conversely, if $P_k < 0$, the measure has a negative sign which indicates that there is more in-group than out-group hostility.

A few remarks about this measure should be noted. First, as I show in the main article, it does not appropriately account for one of the dimensions of affective polarization, namely the interplay between ideological differences and hostility. Second, the measure assumes that two groups of individuals can always be clearly distinguished in the network which proves difficult for cases in which ideological leaning is modelled as a continuous variable, for instance between ± 1. I solved this problem in the experiments by enforcing a split at 0. As a consequence, the measure is more coarse since it treats nodes with an ideological leaning smaller (larger) than but close to 0 the same as nodes with extreme leanings close to -1 ($+1$). It follows that a measure that takes continuous values into account, such as the Affective Polarization Coefficient proposed here, is preferred.

10 Synthetic Data Generation

I generate synthetic network data for the experiments comparing the different polarization measures. Each network G is generated using a stochastic block model for which I specify the number of nodes per community, and a block probability matrix which defines the connection probabilities of these communities. In this matrix, the edge probability for nodes to connect to other nodes within their community (p_{in}) is shown on the diagonal; and all other matrix entries show the probability of edges between the communities (p_{out}). All networks in the experiments presented in section 3 have n=100 nodes. I use a force-directed layout algorithm to visualize all networks presented in the main article.

Ideological Polarization Experiments. I create a stochastic block model with four communities (I - IV) of 25 nodes each and edge probability matrices as specified in the Table 10.1. The ideological leaning values for the nodes in these networks are randomly drawn from a normal

distribution in Figure 4.3(a) in the main article, and bimodal distributions in Figures 4.3(b) and 4.3(c).

	I	II	III	IV
I	0.136	0.136	0.136	0.136
II	0.136	0.136	0.136	0.136
III	0.136	0.136	0.136	0.136
IV	0.136	0.136	0.136	0.136

	I	II	III	IV
I	0.256	0.0024	0.0024	0.0024
II	0.0024	0.256	0.0024	0.0024
III	0.0024	0.0024	0.256	0.0024
IV	0.0024	0.0024	0.0024	0.256

	I	II	III	IV
I	0.269788	0.001208	0.0	0.0
II	0.001208	0.269788	0.001208	0.0
III	0.0	0.001208	0.269788	0.001208
IV	0.0	0.0	0.001208	0.269788

Table 10.1: Block probability matrices for the networks used for the experiments on the ideological polarization measures. Left: random network where $p_{in} = p_{out}$ in Figure 4.3 and 4.4(a) in the main article. Middle: network with four communities which are all connected to each other in Figure 4.4(b). Right: network with four consecutively connected communities in Figure 4.4(c).

Affective Polarization Experiments. For the experiments on the affective polarization measures, I generate stochastic block models with four communities (I - IV) of 25 nodes each. To assign ideological leanings to the nodes, I draw random values from a uniform distribution which results in a dark blue, light blue, light red, and dark red community. Due to how I choose to connect the nodes (see Table 10.2), it is not possible to visually distinguish between all four communities in Figure 4.7 in the main article as the light blue and light red community are tightly connected in Figure 4.7(a)-(c).

	I	II	III	IV
I	0.0625	0.1250	0.0000	0.0000
II	0.1250	0.0625	0.5000	0.0000
III	0.0000	0.5000	0.0625	0.1250
IV	0.0000	0.0000	0.1250	0.0625

	I	II	III	IV
I	0.2375	0.0125	0.0000	0.0000
II	0.0125	0.0000	0.5500	0.0000
III	0.0000	0.5500	0.0625	0.1250
IV	0.0000	0.0000	0.1250	0.0625

	I	II	III	IV
I	0.2375	0.0125	0.0	0.0
II	0.0125	0.0	0.62	0.0
III	0.0	0.62	0.0	0.0125
IV	0.0	0.0	0.0125	0.2375

Table 10.2: Block probability matrices for the networks used for the experiments on the affective polarization measures. Left: network without a clear community structure (Figure 4.7(a) in the main article). Middle: network with a dark blue and a mixed community (Figure 4.7(b)). Right: network a dark blue, dark red, and a mixed community (Figure 4.7(c)).

Next, I calculate the ideological difference for each node pair in the network which can take values in between 0 (no difference) and 2 (maximum difference). For each ideological difference value, I determine an interval from which the hostility value is randomly drawn. Then, the resulting hostility values are adjusted to be in a range between 0 (no hostility) to 1 (maximum hostility).

11 Data Collection

After having discussed the generation of the synthetic experiment data, I now turn to the Twitter data collection process. The first part of the chapter discusses legal considerations that play a role when processing social media data and the second part then describes the data collection and available geo-tags in the data set.

11.1 Legal Considerations

I start by briefly outlining how I assured that the data collection and processing is in line with the legal guidelines set out by the General Data Protection Regulation (GDPR).[1] While some of the tweets in the data set were authored by organizations such as companies or news media outlets, the majority of the tweets were posted by private Twitter users. Within the European legal framework, these users are referred to as *natural persons*. Since the data set contains personally identifiable data such as user names relating to natural persons (GDPR article 1) and the processing of this data takes place within the European Union (GDPR article 3(1)), the data collection and processing activities are subject to the GDPR.

To begin, a legal basis for collecting the data needs to be defined. As outlined in article 6 and article 9 of the GDPR, "scientific research purposes in the public interest" constitute a valid legal basis for collecting personal data (GDPR article 4) as well as sensitive personal data (GDPR article 9). I use scientific research purposes as the legal basis and, in accordance with the data protection guidelines for master theses by the University of Copenhagen, I obtained the confirmation that this project indeed qualifies as a research project from the thesis supervisors prior to collecting the data.

[1] The reference to the legal text can be found under The European Parliament and the Council of European Union (2016). I will refer to this source simply as 'GDPR' in order to improve readability.

Furthermore, the GDPR outlines that special precautions need to be taken when storing personally identifiable data (GDPR article 5(1)(f)). The raw data, i.e., the tweet objects which I collect from the Twitter API, are stored on a password protected external hard drive to prevent unauthorized access by third parties. Further restrictions apply to all files that specify the ideological leaning of the users in the data set since they contain information about the political opinions held by the users and thus qualify as sensitive personal data (GDPR article 9). Consequently, any files containing ideological scores which were inferred from the data are stored on the University of Copenhagen's secure drive to ensure additional data protection.

Lastly, I only analyze the tweets at an aggregate level. I do not infer or describe any information that relates to individually identifiable natural persons in the data set and I do not publish the tweet information or make it otherwise available as part of the thesis.

11.2 Tweet Locations

As argued in the main article, I use the tweet IDs available as part of the TBCOV data set to collect the full tweet object including meta-data (Imran et al., 2022). The authors of the TBCOV data set used a list of manually curated Covid-19 keywords and hashtags to collect tweets. They then inferred geo-locations from the tweet or the meta-data available for each tweet. There are five different types of location tags available in the TBCOV data set:

Geo-coordinates. For a small fraction of tweets, the GPS location from which the tweet was sent is specified. This information is only available for users who actively enable GPS tracking in their privacy setting. Among the different location tags, the GPS coordinates are the most reliable since they point towards the exact location of the user. When available, I use the geo-coordinates to determine if a tweet belongs to a US-based user. If this information is not available, one of the other tags below is used instead.

Place bounding box. When users tweet, they can choose to use a GPS location *in* their tweet. These place bounding boxes point towards a larger area that the user selected as their current location (Figure 11.1 shows an example). They are different from the geo-coordinates mentioned

above, since the user can actively choose the area of the place bounding box themselves. Those areas always refer to real-world places, but a user could choose any place around the world regardless of whether they are currently there or not.

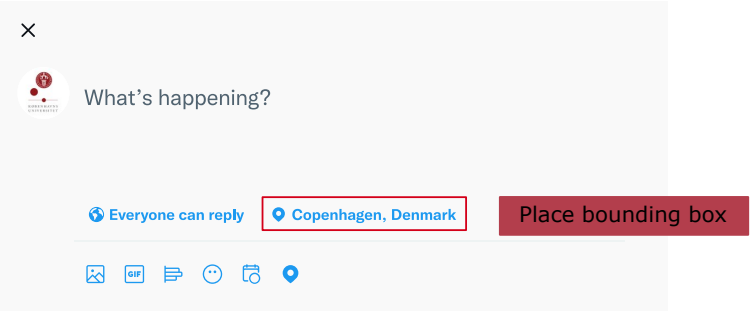


Figure 11.1: Example of a place bounding box used to tag a tweet.

Profile description. Moreover, Imran et al. (2022) rely on the user profile descriptions to infer the geo-location of a user (see Figure 11.2). Since this description is in a free-text field, users are free to refer to any (or no) place here regardless of whether it exists or not and inferring the location is therefore less reliable in these cases.

User location. Public profiles on Twitter can opt to specify a place at which the account is usually located (see Figure 11.2). This user location is a free text field and users can therefore put any kind of information there, including fictional places. Similar to the profile description, this information is less accurate than the GPS coordinates.

Figure 11.2: Example of a profile description and a user location in a Twitter profile.

Finally, Imran et al. (2022) use the text of a tweet since some tweets mention a place or region. However, I do not use the tags inferred from the text of a tweet since they might not reflect the actual location of a user. For instance, a user located in Denmark might tweet about the US Covid-19 restrictions and the geo-tag inferred from the tweet text might thus be 'US' although the user is actually located in Europe.

As argued above, the different approaches vary in how reliably they can detect the actual location of a Twitter user. The authors of the TBCOV data set therefore assess the quality of the geo-tagging results (see Table 11.1). First, they compare the place bounding box information with the actual geo-coordinates for the profiles where GPS coordinates were available. Second, they manually annotate tweets to check the user location and the user profile. Based on the manually annotated data, several performance metrics are calculated. *Precision* shows how many of the inferred locations are relevant, while *recall* indicates how many of the relevant locations were inferred. Moreover, the *F1-score* determines the balance between precision and recall. As the table shows, the geo-tagging approaches performed well overall and there are no major issues reported.

	Precision	Recall	F1-score
Place Bounding Box	0.988	NA	NA
User Location	0.866	1.0	0.929
User Profile	0.888	0.732	0.803

Table 11.1: Performance of the geo-tagging approach deployed by Imran et al. (2022) on the TBCOV data set.

12 Data Preprocessing

Apart from determining the subset of US-based tweets, I apply further preprocessing steps which are described in the following sections.

12.1 English Language Detection

The first preprocessing step consists of detecting English-language tweets. This is necessary because I use a deep learning model to classify offensive tweets in the subsequent analysis. Since this offensiveness detection model has been pre-trained on English language tweets, it is important to ensure that the tweets in my data set are English-speaking to avoid misclassifications due to different tweet languages.

To detect English-language tweets, I use a pre-trained version of the fastText language model (Joulin et al., 2016; Joulin et al., 2017). FastText builds on the idea that the morphological structure of each word matters. In practice, this means that each word is split into smaller parts, the character n-grams of a word. For instance, the n-gram of the word 'pandemic' with $n = 3$ are <pan, and, nde, dem, emi, mic>. The model then uses a hashing function to assign a numeric vector to each n-gram. The word vector for the entire word, e.g. 'pandemic', is calculated as the sum of all the n-gram vectors. To put it differently, the vector representation of each word is a summary of all the sub-word parts and thus accounts for the morphology of each word.

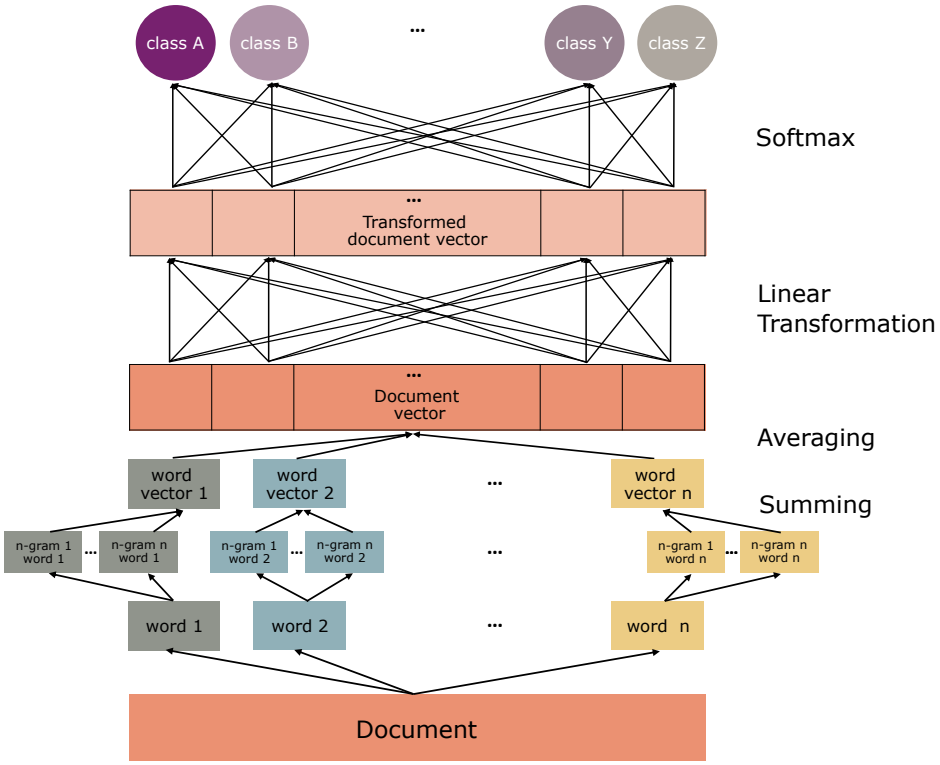


Figure 12.1: FastText architecture. The figure is recreated from Zhou et al. (2020, p. 3).

Next, the average of all the word vectors in a document is calculated (see Figure 12.1). The resulting document vector passes through a linear hidden layer and a softmax function is finally applied to calculate the class probabilities. In the case of a language detection task, the classes correspond to the different languages which can be detected and the class probabilities specify how (un)certain the classification is.

Figure 12.2 shows the distribution of these probability scores for the detection of English-language tweets in the Twitter sample. I choose a relatively high threshold of 0.8 above which I

consider tweets to be English-speaking in order ensure that the further analysis, in particular the offensiveness classification, is not distorted by non-English tweets.

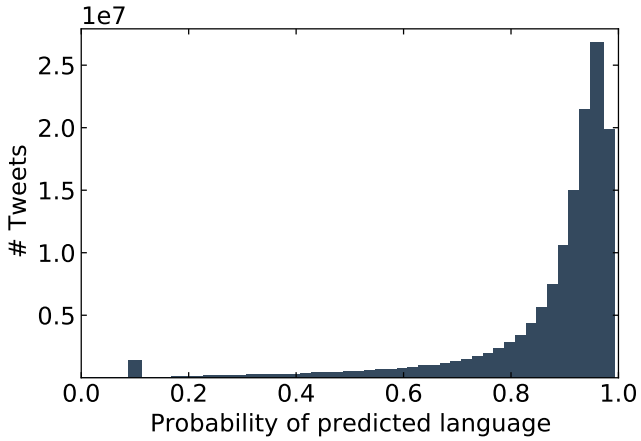


Figure 12.2: Distribution of the language detection probabilities.

12.2 Keyword List: Covid-19 Restrictions

The second preprocessing step involves filtering the data set to only contain tweets which mention the Covid-19 restrictions. The keywords that Imran et al. (2022) used to obtain the TBCOV data set contain broad Covid-19-related terms such as *coronavirus*, *lockdown* or *disaster* and I further restrict this data set to ensure that the results measure polarization within a topically bounded Twitter public that discusses restrictions as a sub-topic of the broader Covid-19 discourse.

Since defining a suitable list of keywords is a challenging task, I train a word2vec model on the tweet corpus which allows me to explore the corpus and the Twitter-specific terms that the users mention in their tweets (Mikolov, Sutskever, et al., 2013; Mikolov, Chen, et al., 2013). The conceptual premise of a word2vec model is that one can infer the similarities between words from the context they appear in. Each word in the corpus is represented as a numerical vector of size n that captures different features. The semantic meaning of words can be described in terms of how these features differ since word synonyms are represented by similar vectors. In practice, cosine similarity is used to calculate the similarity between vectors.

Whereas the fastText vectors are based on the sub-word structure of each word, word2vec learns vector embeddings for each word regardless of its morphological structure. To define

the word2vec vector representations, a neural network is trained. The labelled training data that the model requires is retrieved by defining a word window that moves over the text of each tweet. I use the Continuous Bag of Words (CBOW) technique here in which the model takes the context of each word as the input and then predicts the word corresponding to the context $w(t)$ (see Figure 12.3). A neural network with one hidden layer of size n is then trained on all labelled examples and once the training is finished, the weights of the hidden layer are used as the numerical vectors.

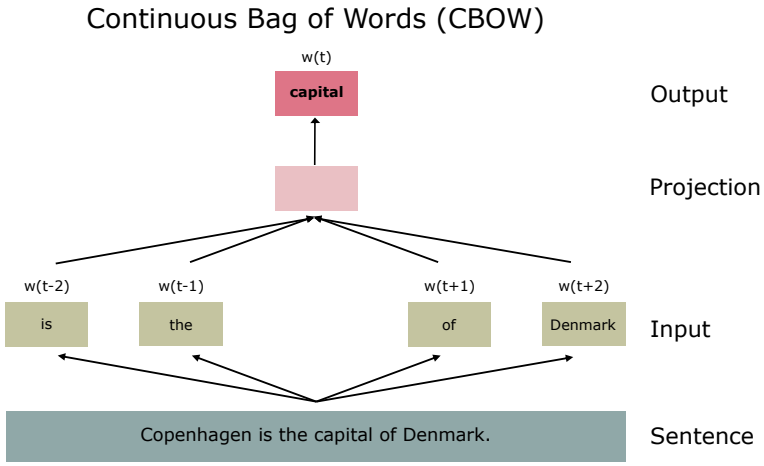


Figure 12.3: CBOW word2vec architecture. The visualization is recreated from Mikolov, Chen, et al. (2013).

As a last step, I manually collect a list of keywords on Covid-19 restrictions and use the trained word2vec model to find synonyms and further relevant terms to add to the keyword list. In this list, I combine keywords and Twitter-specific hashtags. To include Covid-19-skeptical attitudes in the data set, I also use the word2vec model to define a list of keywords used by pandemic-skeptical individuals. Since these users do not seem to use terms that specifically discuss the restrictions, but rather reject the idea of a pandemic in general, the Covid-19-skeptical keywords are broader. The final list of stemmed keywords and hashtags is shown below. The list contained three emojis that were frequently used by users discussing Covid-19 restrictions and those were therefore included in the keyword matching (see Figure 12.4).

Keywords on Covid-19 restrictions. #alonetogether, #asktheexperts, #avoidcrowds, #coveryourcough, #coveryourface, #dontbeacovidiot, #fightcovid, #flattenthecurve, #keepingyousafe, #quaranteam, #quaratinelife, #shelterinplace, #sixfeetapart, #stayingsafe, #staysafeeveryone, #staysafestayhealthy, #stoppingthespread, #stopthespreadofcovid, #togetherapart, abid, adher,

antigen, ban, clorox, clos, complianc, contact, curb, curfew, decontamin, deterg, disinfect, dispens, distanc, guidanc, guidelin, home, hygein, hygien, isol, lockdown, lysol, mask, measur, minim, mitig, pcr, plexiglass, precaut, prevent, procedur, protect, protocol, quarantin, quarentin, recommend, requir, respons, restrict, result, sanat, sanit, school, sheild, shield, test, touchless, touchpoint, trace, wash, wear

Figure 12.4: Emojis used to match tweets discussing Covid-19 restrictions.

Covid-19-skeptical keywords. #americawakeup, #covidlies, #covidscamdemic, #falseflag, #fauxnews, #medialies, #mediascum, #msmistheenemyofthepeople, #nonewnormal, #openamericanow, #plandemic, #plannedemic, #qanons, #scummedia, #sheeple, #thegreatawakening, #wearethenewsnow, #wethepeople, #wwgwga, brainwash, cherrypick, chinaviru, conspirac, debunk, dishonest, fascism, fake, fauci, fearmong, fraudci, hoax, honkler, hype, hyster, kungflu, lamestream, mouthpiec, parrot, propagan, puppet, scaremong, strawman

13 Estimating Ideological User Leanings

The final filtered and English-language data set contains approximately 47 million tweets which had been posted by 4.1 million Twitter users over the course of half a year. As outlined in the main article, the next step of the data analysis involves estimating the ideological leanings of the Twitter users.

13.1 Alternative Approaches

Other studies have proposed various approaches to estimating ideological leaning on Twitter. I review them here to demonstrate why the approach I choose and describe in the following section is the most appropriate in this case.

First, it is possible to estimate ideological leaning based on a piece of text since certain words or discursive techniques can be indicative of the ideological leaning of the speaker (Slapin & Proksch, 2008). Although this is an established approach, the dictionary which this method is based on dates back to 2008 and it works best for inferring ideology from longer texts instead of short Twitter messages. Moreover, all users rely on a similar vocabulary to discuss the Covid-19 restrictions and it is therefore not possible to meaningfully detect ideology based on word usage.

Second, prior studies have used URL sharing behavior as an indicator of ideology. For instance, Cinelli et al. (2021) determine which news outlets users on social media link to in their posts and they use the bias scores provided by mediabiasfactcheck.com (MBFC) to estimate the ideological user leaning. However, it is questionable whether this approach captures the true ideological leaning of users well. For instance, Lazer et al. (2020) analyze the Covid-19 tweets of registered voters in the United States and they specifically investigate the most shared URLs by Democrat and Republican voters. Their findings suggest that Republican Twitter users also share liberal sources such as cnbc.com or cnn.com and that Democrats also link to conservative news outlets such as nypost.com or foxnews.com. It follows that URL sharing behavior is not a reliable indicator of ideological leaning.

A third approach to estimating Twitter user ideology relies on a Bayesian ideal point estimation introduced in Barberá (2015) and Barberá et al. (2015). In this estimation method, a list of elite Twitter accounts comprising politicians, media outlets, pundits, and experts is identified. This method draws on the same idea as multidimensional scaling techniques: left-leaning users are more likely to follow left-leaning elite accounts, whereas right-leaning users prefer right-leaning accounts. Once the scores for the elite accounts are determined, this method estimates scores for other Twitter users by considering how many left-leaning and right-leaning accounts they follow.

Importantly, the elite scores used in Barberá (2015) and Barberá et al. (2015) date back to 2015 and they thus do not capture the current political and media landscape any longer. Recently, other scholars have proposed updated elite scores but their distribution seems skewed as the Democratic elites are grouped closely together and the Republican elites are spread out (McCabe

et al., 2022). Moreover, the scores presented by McCabe et al. (2022) are only preliminary as of now and might be subject to further changes in the future.

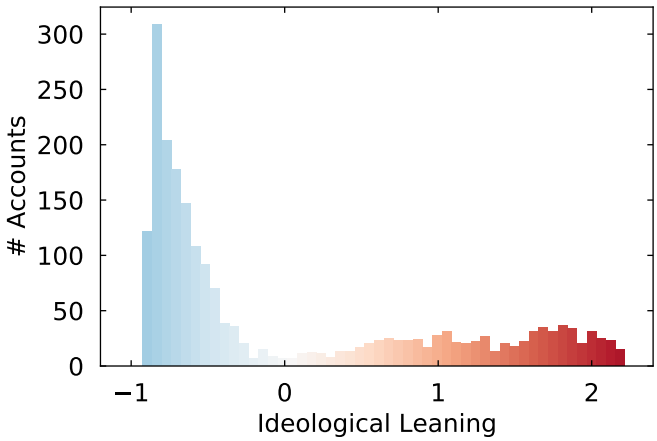


Figure 13.1: Distribution of elite scores in McCabe et al. (2022).

13.2 Retweet-Based Ideological Scaling

As I argue above, it is questionable whether these approaches could accurately capture the ideological leaning of Twitter users in the Covid-19 debate on Twitter. I therefore develop the scaling approach that is based on how often users retweet a list of political and media baseline accounts.

To match the DW-NOMINATE scores of each politician in the 118th US Congress (2019-2021) to a Twitter account, I use a comprehensive list of US politician Twitter accounts provided in McCabe et al. (2022).[1] Moreover, to find the accounts of media outlets on Twitter, I use the names of news outlets available on the MBFC website as a starting point. I then search the Twitter API in an automated manner to look for Twitter profiles that contain the name of the news outlets. For each match, the script verifies that the profile descriptions indeed link to the URL associated with the given media outlet. All ambiguous matches are manually verified.

The MBFC website provides a visual classification of the leaning of each news outlet like the one shown in Figure 13.2. I use a script to determine where the yellow dot is relative to the outermost

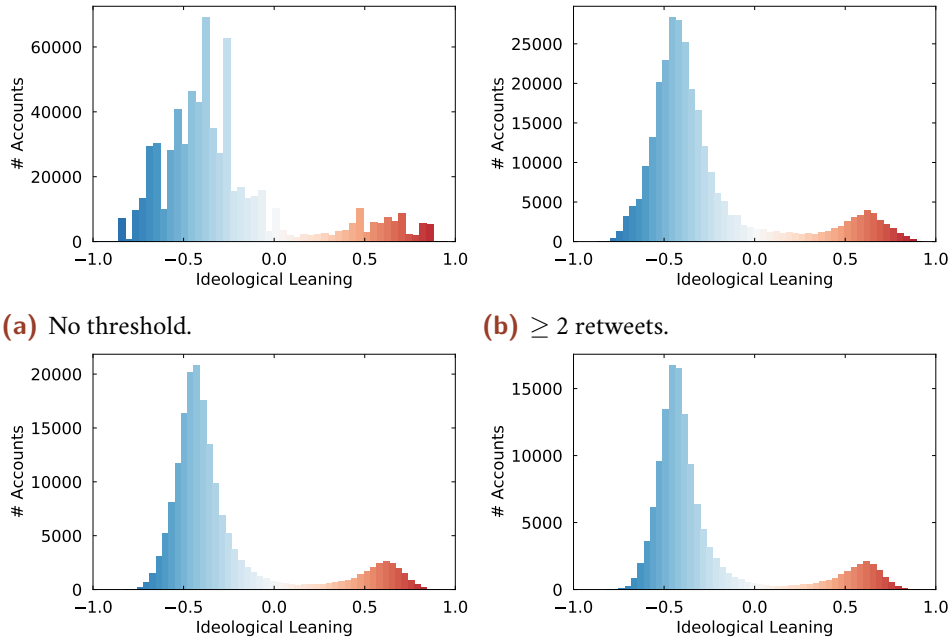
[1]The data is available in the associated GitHub repository.

points of the arrow because this relative position can then be translated into a numerical score between -1 and $+1$.[2]

Figure 13.2: MFBC leaning score for the news outlet Foxnews. Based on the image, I determine the leaning score of Foxnews to be 0.67.

13.3 Robustness Checks

The final list of baseline accounts contains the Twitter handle and user ID of all available political and media accounts as well as an ideological leaning score. As argued in the main article, I only consider users who have retweeted at least 5 posts by the baseline accounts. As the plots below show, the results are robust regardless of the threshold that is chosen.



(a) No threshold. **(b)** $\geq$ 2 retweets.

(c) $\geq$ 3 retweets. **(d)** $\geq$ 4 retweets.

Figure 13.3: Distribution of user scores for different retweet thresholds.

[2]The list of news outlets and leaning scores had been scraped from the website for a previous project and they therefore date back to September 2021 (Hohmann et al., 2022).

14 Offensive Language Classification

To measure affective polarization, I not only need to estimate the ideological leaning of the Twitter users, but also the level of hostility of their interactions. As argued in the main article, I use offensive language as an indicator for hostility. I use use a pre-trained classification model proposed by Barbieri et al. (2020) to detect offensive tweets. The model returns a binary classification of tweets as either not offensive or offensive. This section outlines which architecture underlies the model by Barbieri et al. (2020) and how it was trained.

As a starting point, the authors take a pre-trained RoBERTa-base model, and re-train it on a large corpus of 60 million English-language tweets that had been posted on Twitter in between May 2018 and August 2019. RoBERTa (Robustly Optimized BERT Pre-training Approach) is a version of the commonly used BERT model (Bidirectional Encoder Representations from Transformers). The basic idea of these language models is that they learn contextualized representations from large text corpora in an unsupervised manner. As opposed to the word2vec word representations that I discuss above, these models additionally capture contextual information relating to each word. For instance, the word2vec vector representation of *overlook* is the same regardless of whether it refers to *failing to notice something* or *having a view from above*. However, the BERT or RoBERTa representations for a sentence using the word *overlook* assigns different vectors to the given word since they are able to account for contextual differences.

In order to train the RoBERTa model or use it for classification tasks, the words in each document (tweet) in the corpus are tokenized and an initial input embedding is calculated (for further details see Devlin et al., 2019; Liu et al., 2019). A fraction of the tokens (usually 15%) are then randomly masked and the model is tasked with predicting the hidden tokens based on their context. To do so, RoBERTa relies on a transformer architecture with 24 encoders that each use self-attention mechanisms to quantify the relationships between the input words (see Figure 14.1).

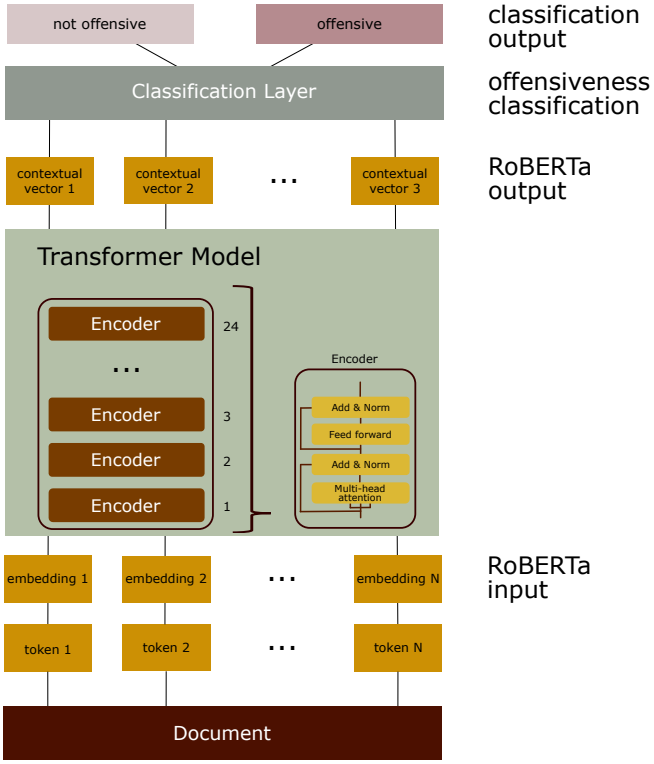


Figure 14.1: Architecture of the RoBERTa-base classification model proposed by Barbieri et al. (2020).

Similar to how humans perceive and remember information, self-attention mechanisms determine which words are most important and should thus be focused on in the further steps of the model. To put it differently, the model captures how important each word in the input sequence is for the other words in the sequence. RoBERTa uses multi-headed self-attention, i.e., each self-attention mechanism is computed multiple times in parallel. The resulting 8 attention vectors are summarized in one vector and passed through a feed-forward neural network before the vector enters the next self-attention layer. The RoBERTa-base model is pretrained on a massive corpus of English-language pieces of text from books, Wikipedia, news sources, and Reddit. Barbieri et al. (2020) then re-train this model on 60 million English-language tweets. In a last step, they use a supervised learning approach to fine-tune the model for offensive language detection. In particular, the authors add a dense layer to reduce RoBERTa's last layer to only two labels and they train the model using the labelled data set provided by Zampieri et al. (2019). They report an F1-score of 81.6 and conclude that the model performs well on the classification task.

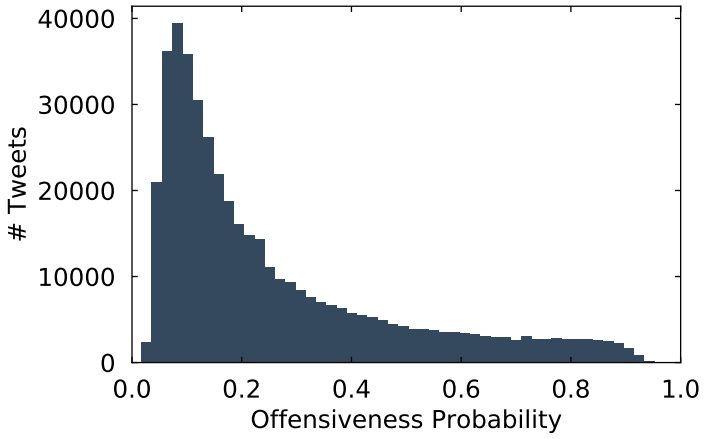
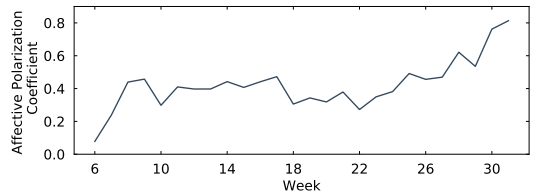


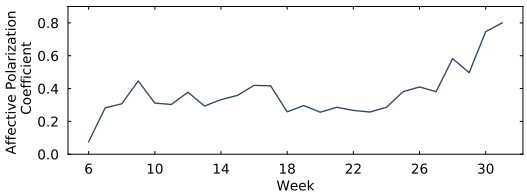
Figure 14.2: Distribution of probabilities for the offensiveness classification.

I use this fine-tuned model to classify tweets as not offensive (0) or offensive (1). In line with the preprocessing steps outlined in Barbieri et al. (2020), I remove all line breaks and URL links, and I replace the user mentions in the text by an anonymized place holder. For each tweet, the model predicts a class and its associated probability (see Figure 14.2). In line with how Barbieri et al. (2020) introduce the use of their model, I choose 0.5 as the threshold above which I consider tweets to contain offensive language. Importantly, the results are robust across different thresholds as shown in the Figure 14.3.

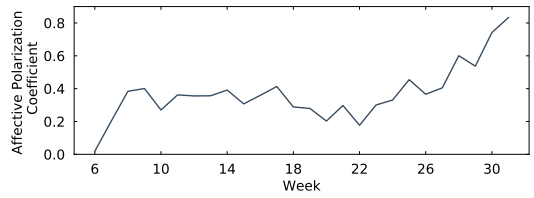
Furthermore, as I argue in the main article, I only draw one undirected edge between two individuals and therefore need to summarize the offensiveness scores for multiple interactions which I do by calculating the average offensiveness score. Other approaches could have entailed summing all offensiveness scores, or taking the maximum score instead of the average and I therefore conduct further robustness checks. Figure 14.3 shows that regardless of which approach I choose, the results are robust as there are no vast differences between them.



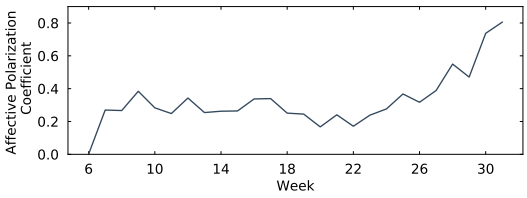
(a) Threshold 0.4; max edge attributes.



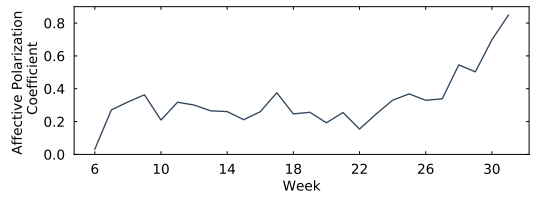
(b) Threshold 0.4; max edge attributes.



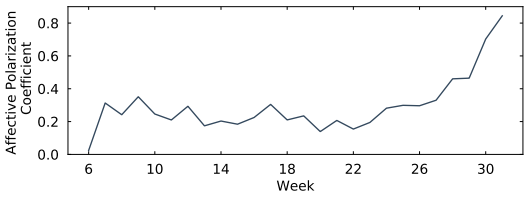
(c) Threshold 0.5; max edge attributes.



(d) Threshold 0.5, summed edge attributes.



(e) Threshold 0.6; max edge weights.



(f) Threshold 0.6; summed edge weights.

Figure 14.3: Affective polarization scores based on the binary offensiveness classes with 0.4 (first row), 0.5 (second row), and 0.6 (third row) as threshold. The plots on the left show the Affective Polarization Score when the edge attributes are calculated as maximum offensiveness scores, and the right plots show the Affective Polarization Score for edge attributes that are summed offensiveness scores.

15 Summary Statistics

I conclude the companion paper by reporting summary statistics for the series of Twitter networks analyzed here. For each network, Table 15.1 shows the number of nodes n, the number of edges m, and the density D of the network which describes the number of actual connections relative to all connections which are theoretically possible. As the table shows, the networks are very sparse as there are only very few connections compared to the number of edges.

Moreover, the table reports the average shortest path length l. A shortest path between two nodes is the path with the fewest number of edges connecting them and the average shortest path length summarizes all these path lengths over all node pairs in the network. Lastly, the table shows

Week	n	m	D	l	Q
6	348	378	0.0063	6.2484	0.1004
7	345	380	0.0064	7.6019	0.1079
8	487	559	0.0047	5.8837	0.0478
9	4458	7024	0.0007	4.7452	0.0163
10	4961	7878	0.0006	4.8274	0.0106
11	6802	12114	0.0005	4.6510	0.0002
12	4171	6022	0.0007	5.3276	0.0137
13	6413	10523	0.0005	4.9655	0.0210
14	8104	13333	0.0004	4.8739	0.0411
15	10297	17006	0.0003	4.8960	0.0524
16	11435	19256	0.0003	4.8994	0.0429
17	10910	18424	0.0003	4.8862	0.0530
18	7161	11359	0.0004	4.9495	0.0532
19	7645	11840	0.0004	5.1327	0.0456
20	7047	10950	0.0004	4.9948	0.0404
21	6427	9551	0.0005	5.1769	0.0480
22	5207	7355	0.0005	5.3698	0.0629
23	3603	4817	0.0007	5.7702	0.0629
24	3928	5152	0.0007	5.6979	0.0389
25	4804	6516	0.0006	5.6444	0.0438
26	4930	7282	0.0006	5.1982	0.0139
27	6313	8974	0.0005	5.3813	0.0177
28	7046	10634	0.0004	5.1794	0.0175
29	8472	13416	0.0004	4.9580	0.0262
30	6463	9916	0.0005	5.0978	0.0220
31	6903	10500	0.0004	4.9309	-0.0001

Table 15.1: Summary statistics for the Twitter networks. For each network considered in the analysis, the table shows the number of nodes n, the number of edges m, the density D, the average shortest path length l, and the modularity Q. To calculate Q, I partition the network into a community of liberal nodes (ideological leanings < 0) and a community of conservative nodes (ideological leanings > 0).

the modularity value Q which indicates how divided a network is into communities based on a given attribute. To calculate Q, I partition the network into a liberal group and a conservative group based on the ideological leaning values. The modularity value then compares the expected fraction of edges if they were randomly distributed to the actual observed fraction of edges within each group. Q can return values between -0.5 (disassortative community structure) and $+1$ (strongly assortative community structure). The table shows that the networks do not exhibit a community structure, i.e., liberal and conservative users are not divided into clear communities as is also shown by the visualizations in Figure 5.4 in the main article.

16 References

Barberá, P. (2015). Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political Analysis*, *23*(1), 76–91. https://doi.org/10.1093/pan/mpu011

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science*, *26*(10), 1531–1542. https://doi.org/10.1177/0956797615594620

Barbieri, F., Camacho-Collados, J., Espinosa Anke, L., & Neves, L. (2020). TweetEval: Unified benchmark and comparative evaluation for tweet classification. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1644–1650. https://doi.org/10.18653/v1/2020.findings-emnlp.148

Cinelli, M., De, G., Morales, F., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The echo chamber effect on social media. https://doi.org/https://doi.org/10.1073/pnas.2023301118

Coscia, M. (2021). Pearson correlations on complex networks. *Journal of Complex Networks*, *9*(6). https://doi.org/10.1093/comnet/cnab036

Deng, X., Li, G., & Dong, M. (2015). Finding overlapping communities with random walks on line graph and attraction intensity. In K. Xu & H. Zhu (Eds.), *Wireless algorithms, systems, and applications* (pp. 94–103). Springer International Publishing.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *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)*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Evans, T. S., & Lambiotte, R. (2009). Line graphs, link partitions, and overlapping communities. *Phys. Rev. E*, *80*, 016105. https://doi.org/10.1103/PhysRevE.80.016105

Garimella, K., Morales, G. D. F., Gionis, A., & Mathioudakis, M. (2018). Quantifying controversy on social media. *Trans. Soc. Comput.*, *1*(1). https://doi.org/10.1145/3140565

Hogg, R. V., McKean, J. W., & Craig, A. T. (2019). *Introduction to mathematical statistics* (8th ed.). Pearson Education.

Hohmann, M., Devriendt, K., & Coscia, M. (2022). Quantifying political polarization on a network using Generalized Euclidean distance [manuscript submitted for publication].

Imran, M., Qazi, U., & Ofli, F. (2022). TBCOV: Two billion multilingual COVID-19 tweets with sentiment, entity, geo, and gender labels. *Data*, *7*(1). https://doi.org/10.3390/data7010008

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., & Mikolov, T. (2016). Fasttext.zip: Compressing text classification models. https://arxiv.org/abs/1612.03651

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2017). Bag of tricks for efficient text classification. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 427–431. https://aclanthology.org/E17-2068

Kernighan, B. W., & Lin, S. (1970). An efficient heuristic procedure for partitioning graphs. *The Bell System Technical Journal*, *49*(2), 291–307. https://doi.org/10.1002/j.1538-7305.1970.tb01770.x

Lazer, D., Ruck, D. J., Quintana, A., Shugars, S., Joseph, K., Grinberg, N., Gallagher, R. J., Horgan, L., Gitomer, A., Bajak, A., A., M., Ognya, K., Qu, H., H, W. R., McCabe, S., & Green, J. (2020). The state of the nation: A 50-state COVID-19 survey report #18: COVID-19 fake news on Twitter.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. https://doi.org/10.48550/ARXIV.1907.11692

McCabe, S., Green, J., Wan, A., & Lazer, D. (2022). New tweetscores: Or, did Donald Trump break tweetscores? *Midwestern Political Science Assocation, Chicago, IL*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, 3111–3119. https://doi.org/10.48550/arXiv.1310.4546

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In Y. Bengio & Y. LeCun (Eds.), *1st International Conference on Learning*

Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings. http://arxiv.org/abs/1301.3781

Mønsted, B., & Lehmann, S. (2022). Characterizing polarization in online vaccine discourse: A large-scale study. *PLOS ONE*, *17*(2). https://doi.org/10.1371/journal.pone.0263746

Newman, M. E. (2003). Mixing patterns in networks. *Physical review E*, *67*(2).

Slapin, J. B., & Proksch, S.-O. (2008). A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, *52*(3), 705–722. http://www.jstor.org/stable/25193842

The European Parliament and the Council of European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN

Tyagi, A., Uyheng, J., & Carley, K. (2021). Heated conversations in a warming world: Affective polarization in online climate change discourse follows real-world climate anomalies. *Social Networks Analysis and Mining*, *11*(87). https://doi.org/10.1007/s13278-021-00792-6

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). *Proceedings of the 13th International Workshop on Semantic Evaluation*, 75–86. https://doi.org/10.18653/v1/S19-2010

Zhou, F., Yang, X. J., & Zhang, X. (2020). Takeover transition in autonomous vehicles: A YouTube study. *International Journal of Human–Computer Interaction*, *36*(3), 295–306. https://doi.org/10.1080/10447318.2019.1634317