Journal of Computational Social Science (2023) 6:1147–1178
https://doi.org/10.1007/s42001-023-00227-6

RESEARCH ARTICLE

A high-dimensional approach to measuring online polarization

Samantha C. Phillips[1] · **Joshua Uyheng**[1] · **Kathleen M. Carley**[1]

Received: 26 May 2023 / Accepted: 25 September 2023 / Published online: 25 October 2023

Abstract

Polarization, ideological and psychological distancing between groups, can cause dire societal fragmentation. Of chief concern is the role of social media in enhancing polarization through mechanisms like facilitating selective exposure to information. Researchers using user-generated content to measure polarization typically focus on direct communication, suggesting echo chamber-like communities indicate the most polarization. However, this operationalization does not account for other dimensions of intergroup conflict that have been associated with polarization. We address this limitation by introducing a high-dimensional network framework to evaluate polarization based on three dimensions: social, knowledge, and knowledge source. Following an extensive review of the psychological and social mechanisms of polarization, we specify five sufficient conditions for polarization to occur that can be evaluated using our approach. We analyze six existing network-based polarization metrics in our high-dimensional network framework through a virtual experiment and apply our proposed methodology to discussions around COVID-19 vaccines on Twitter. This work has implications for detecting polarization on social media using user-generated content, quantifying the effects of offline divides or de-polarization efforts online, and comparing community dynamics across contexts.

Keywords Polarization · High-dimensional · Multi-dimensional · Measurement · Simulation · Social networks

✉ Samantha C. Phillips
samanthp@cs.cmu.edu

Joshua Uyheng
juyheng@cs.cmu.edu

Kathleen M. Carley
kathleen.carley@cs.cmu.edu

[1] Software and Societal Systems, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15231, USA

Introduction

Polarization refers to ideological and psychological distancing between opposed groups through an interplay of social and cognitive processes. There has been growing concern regarding the consequences of polarization in recent years,[1] as it has been linked to a range of negative societal consequences. Research shows that polarization can lead to decreased social cohesion, increased intergroup conflict, and decreased trust in democratic institutions [1–3]. Furthermore, polarization can contribute to the spread of misinformation and propaganda, as individuals may become more susceptible to cognitive biases [2, 4].

There is evidence that polarization is enhanced in online contexts, where users are exposed to a high volume of information and can easily find and share content that aligns with their pre-existing beliefs [5]. Yet the complex relationship between social media and polarization remains largely obfuscated due to constantly evolving platform and algorithmic factors, numerous offline influences, and individual-level differences [6, 7]. Direct interactions between users provide limited information about the complex sharing of information and attitudes online. Therefore, developing tools to evaluate polarization beyond direct communication between communities on social media is essential to improve understanding of the effect of contextual factors [8, 9] and de-polarization interventions [10] on ideological and psychological division.

Many scholars intuitively think of polarization as a distanced bimodal distribution of opinions of group members, often called ideological or issue polarization [11, 12]. In a survey, researchers can directly ask participants about the valence and salience of their opinion towards issues to form such a distribution [13]. Working with observable user-generated content at the scale afforded by social media complicates this process, requiring the development of new theories and methods to estimate the ideological and psychological division between groups of users without having direct access to self-reported attitudes.

Our measurement approach is based on a collective narrative conceptualization of group-level polarization, where collective narratives are shared perspectives or cognitions about social reality [14]. We posit that collective narratives are formed and represented through three dimensions: social, knowledge and source. The social dimension represents communication between users to share and shape collective narratives. Knowledge consists of the ideas, arguments, and other information that form collective narratives, while knowledge sources include any opinion leaders or organizations that group members cite.

Synthesizing existing literature on social and psychological mechanisms of polarization, we propose the following properties constitute sufficient conditions for polarization to occur:

[1] https://carnegieendowment.org/2019/10/01/how-to-understand-global-spread-of-political-polarization-pub-79893

1. Group membership: definition of group membership of two mutually exclusive groups holding opposing ideologies towards a topic or set of topics
2. Awareness: ideologically opposed groups are aware of the collective narratives of the outgroup
3. Social dimension: high levels of direct communication within groups and low levels between groups
4. Knowledge dimension: high levels of shared ideas, arguments, and phrases (referred to as knowledge bits) within groups and **low** levels between groups
5. Source dimension: high levels of shared opinion leaders, media, and organizations (referred to as knowledge sources) within groups and low levels between groups

In this work, we establish a high-dimensional network approach to assess polarization that detects the presence of these conditions in online discourse (Sect. 3). Each dimension of the network represents a different dimension of polarization. Furthermore, we evaluate six existing network-based measures and three network aggregation procedures in our framework through a virtual experiment (Sects. 4 and 5). This work leads to a recommendation of the W/B index or average I/E index applied to the lossy intersection of the social, knowledge, and source networks. We also demonstrate applying the proposed methodology applied to discussions surrounding the COVID-19 vaccine on Twitter over time (Sect. 6), before discussing the implications and path forward (Sect. 7).

Polarizing properties and processes

In this section, we define polarization in terms of collective narratives and expound on the proposed sufficient conditions for polarization to occur. Polarization broadly refers to the state or trend of increasing division between two or more groups [15]. In social psychology, polarization has classically been described as an *intragroup* process in which discussion among group members shifts their views to be more extreme in the same direction as the average pre-existing views [16]. The social identity approach to conceptualizing polarization, specifically self-categorization theory (SCT), introduces the role of *intergroup* context in enhancing and enabling polarization [17, 18].

SCT suggests that polarization occurs when group members conform to some group norm that is induced through internal communication and deduced from the broader context and relative to other groups. According to this theory, the group norm is shifted further from the outgroup to appear more extreme than the actual average view of the group when groups are distinct and hold opposing ideologies [17]. As group members hold more extreme views and further psychologically separate from other groups, polarization occurs.

Building on the idea that intragroup social-psychological processes are enhanced by the intergroup context, we turn to a recent conceptualization of polarization based on shared collective narratives (i.e., shared cognitions and perspectives) of social reality [14]. Bliuc and colleagues propose that polarization occurs when there are groups

with opposing collective narratives, which can be thought of as opposing ideological groups (group membership) [14]. These collective narratives are defined in opposition to an alternative collective narrative, requiring intergroup awareness (awareness).

When members of a group endorse a collective narrative, they have the same basis for social identity formation that informs subsequent behaviors, beliefs, and attitudes. Collective narratives contain ideas, beliefs, and values that influence the adopted norms and intergroup attitude. Analogously, differences in information (and crucially, misinformation) contributing to collective narratives can result in entirely different perceptions of reality and behavioral effects [19, 20]. Therefore, salient shared collective narratives (like identities) within groups, defined in opposition to alternative narratives, promotes further ideological and psychological division between groups.

We aim to detect the presence (or absence) of opposed collective narratives between online communities to indicate polarization is occurring, rather than analyzing the degree of belief or attitude directly. We suggest three dimensions that dictate a social structure in which there are shared collective narratives within groups and different narratives between groups. The first dimension is social and represents direct communication between group members. This is a way narratives are shared or disputed in the context of social relationships. As suggested by social identity theory, social influence shapes how people perceive information and members of other groups [17, 18, 21]. Thus, more communication within ideologically-defined communities than between indicates more consideration of ideas from in-group members than out-group, which contributes to polarization (social dimension).

An affordance of social media is the possibility of shaping the narrative around an issue through indirect communication. Posts do not necessarily directly endorse or reply to another user. Rather, the content in posts can represent ideas, arguments, and other knowledge that form collective narratives. More common knowledge usage within a group relative to between groups indicates different collective narratives adopted by each community, enhancing division (knowledge dimension) [21].

Shared knowledge sources indicate similar underlying knowledge is informing collective narrative and identity formation. Sources includes opinion leaders, such as politicians and media, that have agenda-setting power in public discourse [21]. Furthermore, polarization among influential people and organizations can incite mass polarization, especially if people feel a strong sense of identification with that leader's group or ideology [22, 23]. Hence, common sources used within groups and not between designates opposing collective narratives as well (source dimension).

Given the preceding review of psychological and social theories of polarization, we establish five sufficient conditions for polarization to occur in List 1. These conditions were put together to evaluate polarization on social media, but they extend to offline contexts as well. The main differentiating factor is the data access afforded by social media platforms.

Note that we purposefully use "high" and "low" instead of quantitative values when describing conditions because 1) different measures report polarization on different scales and 2) our approach to assessing polarization is not necessarily meant to be informative in isolation. Instead, polarization of online groups can be compared across contexts and time.

Effects of social media

On social media, algorithmic factors can compound cognitive influences, like selective exposure, to create highly fragmented social networks where people almost exclusively communicate with others who hold the same identity and ideology [24]. These self-contained groups are often referred to as echo chambers. There is evidence that echo chambers facilitate stronger group identities and more extreme views online, and therefore, collective narrative formation [5, 25]. The true prevalence of echo chambers on social media is highly contested [26] with some studies suggesting access to social media increases exposure to diverse views [27]. Most likely, this effect depends on the social media platform and any number of decisions made by the user [24, 28].

Even if people do see opposing views, the impact on belief and identity is not clear. Some studies report a negative relationship between discussions along users with opposing views and polarization, indicating such interactions do in fact moderate opinions [29, 30]. Other work empirically shows the opposite occurs—communication across groups increases polarization between ideological groups as people defend their pre-existing position [31, 32]. Yet other scholars suggest people merely tolerate other viewpoints without being impacted further [33].

We assume communication and information sharing across groups indicates a weaker bias towards like-minded users, following research that shows people selectively interact with in-group users and content that aligns with their pre-existing beliefs [34, 35]. Although users may see posts containing opposing views, selective communication and information sharing indicates strong in-group favoritism and identification, which reveals preferential collective narrative development. At the same time, polarization requires some degree of awareness of ideas held by other groups to provide intergroup context that informs intragroup collective narrative development.

In sum, we suggest that reported deviations in the prevalence of echo chambers on social media and their impact on polarization may arise due to differences in detection and influence measures. Having a shared collective narrative within a group requires discussion and sharing of ideas and information sources, suggesting high levels of communication, shared knowledge, and shared knowledge sources within a group is a necessary condition for polarization. When members of different groups communicate and share ideas between them as much as within their respective groups, there is evidence of a common collective narrative or perspective across communities. Therefore, fewer connections between ideologically opposed groups relative to within indicates opposing collective narratives and therefore, polarization.

High-dimensional network approach

In our high-dimensional approach, we apply existing network-structure based measures of polarization to networks representing social, knowledge, and source dimensions. Remarkably, most of these methods have only successfully been applied to uni-dimensional networks, typically interaction networks where edges indicate one user re-posts, mentions, or follows another [36–42]. This neglects the knowledge and source dimensions of polarization, overlooking the impact of indirect

communication on collective narrative development and division. Alternatively, a high-dimensional representation of communication is required to detect the sufficient conditions of polarization described in List 1.

The proposed methodology broadly entails four steps:

1. Data collection
2. Generate high-dimensional and aggregated networks
3. Partition users into ideologically opposed groups
4. Measure polarization

We focus on improving the second and fourth steps in this article. The network generation step is described in the following subsections, where we define high-dimensional networks and associated measures, as well as four aggregation techniques. Polarization measure application is discussed in Sect. 4 and 5, where we specify and evaluate six existing network-based measures that fit within our framework. We demonstrate all required steps and interpretation of results in a case study in Sect. 6.

High-dimensional network definitions

High-dimensional networks are sometimes referred to as multi-dimensional, multi-layer, multi-plex, or meta-network, depending on the field and context [43–46]. Features that differentiate these terms include the number of sets of nodes and dimensions. All networks generated in this work have the same set of nodes in each dimension; we use “high-dimensional" throughout the paper for consistency.

Definition 1 *(High-dimensional network) Define a weighted, directed high-dimensional network by* $G = (V, E, L)$, *where* V *is the set of nodes,* L *is the set of labels for layers,* E *is the set of edges denoted by* (u, v, l, w) *where* u, v *are nodes,* l *is the label for the dimension, and* w *is the edge weight [*44]. *Each edge weight* w *is a non-negative integer.*

The high-dimensional network used here contains three dimensions: social, knowledge, and source. The social dimension is typically directed, where edges indicate one node is directing communication towards another. However, the shared knowledge and shared information source networks are inherently undirected—each edge denotes two nodes using the same knowledge bit or information source, respectively. To preserve as much information as possible, we make all networks directed before applying measures whenever possible. For undirected networks, this entails replacing each undirected edge with bidirectional, directed edges.

For the measure that requires undirected networks (spectral segregation index—SSI), we make all networks undirected. Directed edges between each pair of nodes are summed to weight the replacement undirected edges. Because we consider both directed and undirected forms of the high-dimensional

representation of connections between users, we define overall density, within community density, and external community density for each.

Definition 2 *(Overall density) We define the density of an (un)directed high-dimensional network analogously to the traditional definition of density for uni-dimensional networks, based on the ratio of existing and possible edges. Consider an (un)directed high-dimensional network G with $|V|$ nodes, $|E|$ total edges across all layers and $|L|$ layers. There are, by definition, no edges between dimensions. For the directed case, there are at most $2 \cdot |L|$ edges between nodes. For the undirected case, there are at most $|L|$ edges between nodes. The density of an (un)directed high-dimensional network, d, is calculated by*

$$d = \frac{|E|}{k \cdot |L| \cdot |V|(|V| - 1)}$$

where $k = 2$ for the directed case and $k = 1$ for the undirected case.

Definition 3 *(Within Community Density) Consider a subset of nodes forming a community C in an (un)directed high-dimensional network G with $|V_C|$ nodes, $|E_C|$ edges within the community across all layers and $|L|$ layers. The density of a community in an (un)directed high-dimensional network, d_C, is calculated by*

$$d_C = \frac{|E_C|}{k \cdot |L| \cdot |V_C|(|V_C| - 1)}$$

where $k = 2$ for the directed case and $k = 1$ for the undirected case.

Definition 4 *(External Community Density) Consider a subset of nodes forming a community C in an (un)directed high-dimensional network G with $|V_{\sim C}|$ nodes in the overall network that are not contained in C, $|E_{\sim C}|$ edges external to the community across all layers and $|L|$ layers. The density external to a community in an (un) directed high-dimensional network, $d_{\sim C}$, is calculated by*

$$d_{\sim C} = \frac{|E_{\sim C}|}{k \cdot |L| \cdot |V_C| \cdot |V_{\sim C}|}$$

where $k = 2$ for the directed case and $k = 1$ for the undirected case.

Aggregation techniques

There are two possible paths to account for multiple dimensions at the same time when measuring polarization using a high-dimensional network. The first option is to aggregate measures after applying them to each dimension separately. This approach is typically used if the dimensions are conceptually distinct, and therefore merging them does not make sense [47]. Additionally, this preserves the information

contained in each layer separately. In this case, we simply *average* the output of the measure applied to each dimension.

The other option is to aggregate the dimensions of the network prior to applying measures. This is a well-studied process [48, 49] due to the wealth of applications of high-dimensional networks, such as social and organizational systems [46, 50, 51]. We focus on methods that aggregate dimensions without removing nodes. Rather, edge weights between nodes that have an edge in any layer are updated according to a mapping.

There are several possible mappings to generate a uni-dimensional network from a high-dimensional network [49, 52]. We define a family of thresholded networks that are generated by setting a threshold $L^* \leq |L|$ to keep (or disregard) edges. An edge between nodes u and v is included in the aggregated network if there exists an edge between u, v for at least L^* layers of the high-dimensional network and 0 otherwise. In particular, we have the *union* network, where $L^* = 1$, and *intersection* network, where $L^* = |L|$. Additionally, we define L^*-edges networks where $1 < L^* < |L|$. Since we have three dimensions, the only other possible value of L^* is 2. We refer to this network as the *lossy intersection* network.

The edge weights of thresholded networks are the sum of included edges (and 0 otherwise). If the edge weights of the input networks are binary, so are the edge weights of the aggregated thresholded networks.

In sum, we investigate four aggregation techniques in our methodology: average measures, union network, intersection network, and lossy intersection network.

Comparison of existing measures

Methods applied to characterize online polarization are typically described as content-based, network-based, or a combination of the two. Purely content-based methods use the information contained in posts and do not account for the underlying structure of communication. These methods often rely on language and/or domain-dependent natural language processing tools [53]. Alternatively, researchers have turned to manually labelled keywords [54] or social media users [55] to estimate the degree of polarization between communities.

We incorporate content into our approach without necessarily requiring language models or manual labelling of terms. Rather, we associate knowledge bits and information sources with users previously assigned stances towards the specified topic using a bipartite network. Then we project the bipartite network to obtain an undirected network between users where a link indicates the users used the same term or domain. Because we describe the degree of shared narratives, our method is not purely a network-based.

In the following subsection, we describe existing network-based polarization measures and our selection process for our analysis. Next, we specify the six metrics we analyze and their relevant properties.

Network-based polarization measures

Existing network-based measures of online polarization can broadly be characterized as either traditional measures of group structure or polarization-specific. Several of the measures selected have also been described as measures of segregation [39]. Segregation, polarization, and homophily are closely related, but distinct conceptually.

At a general level, segregation is “the degree to which two or more groups live separately from one another" [56]. Homophily, the tendency to have social ties with those most similar to oneself, is a process that can lead to segregation and polarization [5, 57]. Yet it is not necessarily sufficient for either to occur [58, 59]. Thus, the extent to which the groups are divided is encapsulated by segregation [56]. However, segregation is not necessarily concerned with the degree of homogeneity *within* groups.

Our selection process involved an in-depth literature review of prominent methods [39, 41, 42, 60]. Notably, measures based on the work of Esteban and colleagues use the distribution of an attribute of the population of interest, such as income or opinion [12, 61, 62]. Since we are working with network representations of communication and content usage, we do not consider measures based on a uni-dimensional distribution.

Well-established measures of group structure not initially developed for quantifying polarization include the E/I index [63], modularity [64], segregation matrix index (SMI) [65], and spectral segregation index (SSI) [66]. These measures display a variety of attributes and applications. The E/I index and modularity have been used to assess echo chamber-ness of polarized groups [37] and degree of network structure of political groups online [36], respectively. Another similar measure of intergroup connectedness that has been used to measure polarization compares the number of edges between groups and total number of edges, described by Rajabi and colleagues [67].

The SMI is a measure of group cohesiveness, where a cohesive group is “a social group of actors who prefer to interact with one another more than with others and reveal a highly self-preference segregative attitude" [65]. This definition aligns with our ideal properties of polarization we established. Similar to the E/I index, the SMI measures the relative number and intensity of edges within and between groups. Finally, the SSI was established to measure school and residential segregation using social interactions [66]. Notably, the SSI returns individual and group-level segregation assessments.

More recently, metrics were introduced specifically for measuring polarization on social media. Some scholars presented measures that highlight the role of boundary nodes and edges between stance groups, like boundary connectivity controversy and edge betweenness controversy [38, 41]. Other techniques include random walks to determine the likelihood of a member of one group interacting with a member of another group and mapping the network to a low-dimensional embedding [41]. Garimella and colleagues find random walk controversy and embeddings controversy most reliably distinguish between controversial and non-controversial topics. Later studies modified random walk controversy to account for the distance of

the random walk from the initial user and weights of edges with reliable results [8, 42]. On the other hand, embedding-based approaches have been shown to produce unstable results [8]. To encapsulate multiple approaches established to assess online polarization, we consider the boundary connectivity controversy and random walk controversy [38, 41].

Description of selected measures

Crucially, some measures report polarization for each group, while others describe polarization of the network as a whole. We aim to measure polarization of both groups at the same and adjust measures as needed, described in the following sections. While all the measures can incorporate weighted edges, they are not necessarily designed for directed networks. Whenever possible, we retain the information about the direction of interactions. We discuss any adjustments for each measure.

In sum, the following measures require: (1) a (high-dimensional) network and (2) labelled group membership (partitioning the network) for each node.

W/B index

The W/B index is the percent difference in edges within (W) and between (B) all groups. It is a network-level extension of the group-level E/I index and SMI [63, 65]. Suppose we have two groups denoted X and Y. Let l_{XX} and l_{YY} be the number of edges within group X and Y, respectively. Let l_{XY} be the number of links from group X to Y and l_{YX} be the opposite.

$$W/B = \frac{l_{XX} + l_{YY} - l_{XY} - l_{YX}}{l_{XX} + l_{YY} + l_{XY} + l_{YX}} \tag{1}$$

The W/B index is bounded between -1 and 1. In order for the W/B index to equal 1, the groups must have no links between them. However, the awareness condition is not satisfied in this case, and therefore we would not assume polarization is occurring. The threshold of interactions, shared knowledge, and shared sources between groups that designate awareness is not necessarily constant and thus, requires case-by-case consideration.

A value closer to -1 indicates low polarization, where groups are more interconnected than intra-connected. The W/B index equals -1 if all links are between groups, which certainly does not provide evidence of *opposed* collective narratives held by each group.

In sum, both the minimum and maximum value of the W/B index describes unpolarized groups. Other than the extremes, a higher W/B index denotes more polarization. We note that the W/B index is appropriate for both undirected and directed networks.

Average I/E index

In addition to creating an entirely new network-level measure, we assess the average percent difference in edges internal (I) and external (E) to each group. This is inspired by the E/I index and SMI, but we use the number of edges instead of density and reverse the use of internal and external links. Suppose we have two groups denoted X and Y. Let l_{XX} and l_{YY} be the number of edges within group X and Y, respectively. Similarly, l_{XY} is the number of links from group X to Y and l_{YX} is the opposite.

$$\text{Avg. } I/E = \frac{l_{XX}l_{YY} - l_{XY}l_{YX}}{l_{XX}l_{YY} + l_{XY}l_{YX} + l_{XX}l_{YX} + l_{YY}l_{XY}} \tag{2}$$

The average I/E index is also between -1 and 1, where a larger value denotes more polarization (unlike the E/I index). An average I/E index of 1 indicates the density between groups is 0, while an average I/E index of -1 indicates the density within groups is 0. Again, if there is no evidence of awareness of other groups, we assume polarization is not occurring. Thus, the minimum and maximum average I/E values represent unpolarized communities. However, as the average I/E index increases, polarization also increases until there is evidence of a lack of awareness of other groups. The appropriate threshold of awareness is dependent on the context and thus requires case-by-case consideration. The average I/E index is appropriate for both undirected and directed networks.

Modularity

Modularity is a measure of community structure that compared the actual and expected number of edges within communities [64]. It is calculated by summing the difference between actual and expected number of edges of pairs of nodes within the same group.

Let m denote the total number of edges in the network. Then let i, j be distinct nodes, A_{ij} denote the number of edges between the i and j, and k_i, k_j be the degrees of node i and j, respectively. Modularity, Q, is defined as

$$Q = \frac{1}{4m}\sum_{ij}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta_{ij} \tag{3}$$

where $\delta_{ij} = 1$ if i and j are in the same group and $\delta_{ij} = 0$ otherwise.

A score close to 1 denotes well-defined groups and thus a high level polarization, while modularity close to 0 indicates the number of edges within groups relative to between is the same as it would be expected if the edges were random. A score close to -1 means the groups are more likely to communicate between them than would occur due to random chance. Like the W/B and average I/E index, modularity can be applied to both directed and undirected networks.

Spectral segregation index

The spectral segregation index (SSI) was originally developed in the context of racial segregation to capture the connectedness of members within groups using social interactions [66]. The SSI is defined on a group level, but can be averaged across groups to determine network-level polarization.

This method requires a normalized adjacency matrix A and defined groups, X and Y. Without loss of generality, fix group X. Suppose B is a sub-matrix of A with only nodes in X. Next, identify the set of connected components in B, C_K. On a component level, SSI_c is equal to the largest eigenvalue. Since we are concerned with network-level polarization, we do not break down the SSI into values for each individual node. However, we note that a key attribute of this measure is that it can be measured on an individual level and direct readers to [66].

The group-level SSI for group X is

$$SSI_X = \frac{1}{N_c} \sum_{c \in X} SSI_c \tag{4}$$

where N_c is the total number of components in group X. We average SSI_X and SSI_Y to obtain the SSI for the network.

The range of SSI values is 0 to 1, where 1 represents the most polarization. Like the preceding measures, a network with an SSI of 1 has no edges between groups. A network with an SSI of 0 has only edges between groups. Given the necessity of intergroup awareness for polarization to occur between groups, the most extreme cases are both not polarized. Generally, a higher SSI represents more polarization.

Since SSI was established to measure geographical separation, it is intended for undirected networks. It is well-known that symmetric matrices with real elements only have real eigenvalues, which does not hold for non-symmetric matrices with real elements. Matrices of undirected networks with integer edge weights fit the criteria to have real eigenvalues. Distances must be (positive) real numbers, so we make the input network *undirected* for this measure.

Boundary connectivity controversy

Boundary connectivity controversy (BCC) is based on the structure of community boundaries between stance groups [38]. The authors posit that boundary nodes of polarized groups are more likely to connect with internal nodes than boundary nodes of the opposing group, as hostile interactions decrease the number of popular nodes in both groups.

Let X and Y denote groups of users. Define the set of boundary nodes B_X and B_Y of groups X and Y, respectively, as follows. A node n in either group is a boundary node if it satisfies 2 conditions: (1) $n \in X$ has at least one edge connecting to a node in Y; (2) $n \in X$ has at least one edge connecting to another node in X that is not connected to Y. Define the internal nodes of each group as the remaining nodes in X and Y that are not boundary nodes: $I_X = X - B_X$, $I_Y = Y - B_Y$. Let B be the union of B_X and B_Y and I be the union of I_X and I_Y. Then

$$BCC = \frac{1}{|B|} \sum_{n \in B} \frac{d_I(n)}{d_B(n) + d_I(n)} - 0.5 \tag{5}$$

where $d_I(n)$ is the number of edges between node n and internal nodes I and $d_B(n)$ is the number of edges between node n and boundary nodes B. BCC is bounded between -0.5 and 0.5. If BCC is close to 0.5, then the groups are highly polarized with more boundary nodes connected to internal nodes of their respective group than other boundary nodes. A BCC close to -0.5 indicates the opposite.

If there are no boundary nodes, meaning there are no edges between groups, BCC is undefined. This aligns with our intuition that awareness of other groups is a necessary condition for polarization. Finally, BCC can be applied to directed or undirected networks.

Random walk controversy

Random walk controversy (RWC) measures the likelihood of a random user on each side of a controversial discussion being exposed to authoritative content from the other side [41]. The intuition is that members of more polarized groups are less likely to interact with key actors in other groups.

We first identify the k most authoritative users by selecting those with the highest total degree centrality scores. Each random walk starts from one of the groups (chosen randomly) and terminates once an authority node is reached (on either side). Let $P_{XY} = \text{P}[\text{start in group } X | \text{ end in group } Y]$, the probability a random walk started in group X given that it ended in group Y. Then

$$RWC = P_{XX}P_{YY} - P_{XY}P_{YX} \tag{6}$$

A score close to 1 indicates a low probability of exposure to content in the other group, while a score close to 0 denotes a similar likelihood of a node reaching the other group and not. An RWC closer to -1 represents a low level of polarization and a higher likelihood of exposure to content in the opposing group than in their respective group. When RWC equals 1, there are no edges between groups. When RWC equals -1, there are no edges within groups. Both represent a lack of polarization, although overall higher RWC indicates more polarization. RWC is designed for directed networks.

Virtual experiment

Simulations are a well-established tool to compare and evaluate metrics applied to networks with varying parameters [68–70]. Several studies use controlled virtual experiments to investigate the effect of varying parameters, like density and agent influence, on network measures, such as centrality [68] and segregation [70].

Virtual experiments allow us to go beyond arbitrarily choosing empirical cases for evaluation, as ground truth for these datasets is often difficult to discern or quantify and thus prevents systematic analysis. We use a virtual experiment to assess

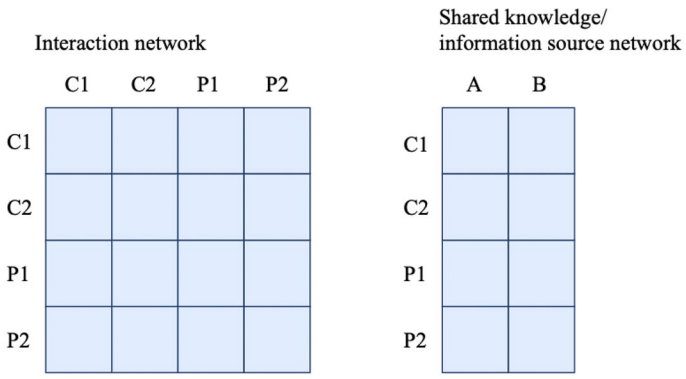

Fig. 1 Stochastic block model (SBM) representation of synthetic networks. Left SBM applies to interaction networks. Right SBM applies to shared knowledge/source networks. Note: C1: user Group 1 core; P1: Group 1 periphery; G2: user Group 2 core; P2: Group 2 periphery; A: knowledge/source Group A; B: knowledge/source Group B

how existing network-based measures perform on projected bipartite networks (shared knowledge, shared source) and aggregated networks (union, lossy intersection, intersection). In particular, we evaluate the following characteristics of polarization measures:

1. More polarized as density within groups increases
2. Less polarized as density between groups increases
3. Distribution of degree of polarization supports identifying differences across domains, times, and platforms
4. Computational efficiency

Parameters and synthetic network generation

We simulate three types of networks: interactions, shared knowledge, and shared sources. Generated networks have global community structure (two stance groups, referred to as Group 1 and Group 2 henceforth) and local core-periphery structure, as introduced by Borgatti and Everett [71–73]. Figure 1 contains examples of the stochastic block models (SBM) used to generate both types of networks, where each block is assigned a density.

Core-periphery structure has been shown to describe communities on social media engaging in collective action [74] and discussion surrounding specified topics like agriculture [75] and national attitudes [76]. It is characterized by two distinct types of nodes: a dense “core" of highly connected nodes surrounded by less connected “periphery" nodes [77].

Generating the interaction network requires directly setting the density between agents in each stance group and core/periphery. To generate shared knowledge and source networks, we set the density of edges between each subgroup of users (i.e., Group 1 core, Group 1 periphery, Group 2 core, Group 2

Table 1 Synthetic interaction network parameters

Parameter	Number	Values
Agent population, N	1	1,000
Topology	1	core-periphery
Relative agent group size, $n1$, $n2$	1	50/50
Relative core and periphery group sizes, $c1/p1$, $c2/p2$	3	50/50, 75/25, 25/75
Within group core-core density, d_{c1c1}, d_{c2c2}	1	0.1
Within group core-periphery density, $d_{c1p1}, d_{p1c1}, d_{c2p2}, d_{p2c2}$	2	0.01, 0.001
Within group periphery-periphery density, d_{p1p1}, d_{p2p2}	1	0.001
Between group core-core density, d_{c1c2}, d_{c2c1}	1	0.0001
Between group core-periphery density, d_{c1p2}, d_{c2p1}	1	0.0001
Between group periphery-core density, d_{p1c2}, d_{p2c1}	2	0.001, 0.0001
Between group periphery-periphery density, d_{p1p2}, d_{p2p1}	2	0.1, 0.0001

periphery) and each group of knowledge bits/sources (without loss of generality, referred to as Group A and Group B henceforth). This creates bipartite user to knowledge bit and source networks, respectively.

The shared knowledge bit and source networks are then generated by projecting these bipartite networks. For the purposes of synthetic network generation, the shared knowledge and shared source networks are the same.

Of course, when working with real data we do not necessarily have knowledge bits and sources designated for each stance group. For the purposes of this virtual experiment, we use these groups for the stochastic block model representation. The idea is that members of each stance group will use knowledge bits/sources in Group A and B to varying degrees. Without loss of generality, we assign Group 1 to share more content from Group A and Group 2 to share more content from Group B.

To bound the set of parameters, we maintain a highly dense core for both stance groups. Then, we vary the relative size of the set of core and periphery nodes, as well as the density of periphery nodes within and between groups. Table 1 contains the control and independent variables for interaction network generation. Table 2 contains variables for the shared knowledge/source networks. In total, there are 36,864 unique sets of parameters. For simplicity, we use binary edge weights when initializing networks.

Results

More polarized as density within groups increases

We evaluate the effect of average within group density on each measure applied to each network by calculating the partial correlation [78]. This way we can measure the relationship between average within group density and polarization level

Table 2 Synthetic shared knowledge and shared source network parameters

Parameter	Number	Values
Number of knowledge bits, K	1	500
Number of knowledge sources, S	1	100
Relative knowledge group size, $k1$, $k2$	1	50/50
Relative knowledge source group size, $s1$, $s2$	1	50/50
Within group core-knowledge, d_{c1k1}, d_{c2k2}	1	0.005
Within group core-source, d_{c1k1}, d_{c2k2}	1	0.001
Within group periphery-knowledge, d_{p1k1}, d_{p2k2}	2	0.001, 0.00001
Within group periphery-source, d_{p1k1}, d_{p2k2}	2	0.001, 0.00001
Between group core-knowledge, d_{p1k2}, d_{p2k1}	1	0.0001
Between group core-source, d_{p1k2}, d_{p2k1}	1	0.00001
Between group periphery-knowledge, d_{p1k2}, d_{p2k1}	2	0.00001, 0.00005
Between group periphery-knowledge, d_{p1k2}, d_{p2k1}	2	0.00001, 0.00005

Table 3 Partial correlation between measures and average within group density controlling for between group density

Measure	Interaction	Shared knowledge	Shared source	Union	Lossy intersection	Intersection
W/B index	0.201^{***}	0.697	0.069^{***}	0.75^{***}	0.533^{***}	0.137^{***}
Avg. I/E index	0.224^{***}	0.773^{***}	0.176^{***}	0.807^{***}	0.627^{***}	0.236^{***}
Modularity	0.556^{***}	0.675^{***}	0.286^{***}	0.862^{***}	0.651^{***}	0.368^{***}
RWC	0.048^{***}	0.44^{***}	-0.198^{***}	0.341^{***}	0.153^{***}	-0.007
BCC	0.073^{***}	-0.496^{***}	-0.078^{***}	-0.281^{***}	0.031^{***}	-0.359^{***}
SSI	0.231^{***}	0.82^{***}	0.221^{***}	0.318^{***}	0.325^{***}	0.112^{***}

$^{*}p < 0.05, ^{**}p < 0.01, ^{***}p < 0.001$

while controlling for between group density. Table 3 contains partial correlation coefficients and significance.

We expect positive partial correlations for each measure and network type. This largely holds. In particular, the W/B index, average I/E index, SSI, and modularity have significantly positive partial correlations for all six network types.

The projected bipartite networks, shared knowledge and shared source, seem to alter the behavior of RWC and BCC. This is evident by the negative partial correlations for those measures and networks. For these networks, when the density within groups increases (controlling for between groups density), RWC and BCC report less polarization.

Moreover, projected bipartite networks tend to be more dense than the traditional uni-dimensional interaction network. Hence, the projected bipartite networks may overwhelm the structure of the strict union network. Similarly, the interaction network can be overly constrained by the interaction network. RWC and BCC both have

Table 4 Partial correlation between measures and between group density controlling for average within group density

Measure	Interaction	Shared knowledge	Shared source	Union	Lossy intersection	Intersection
W/B index	-0.994^{***}	-0.945^{***}	-0.786^{***}	-0.948^{***}	-0.926^{***}	-0.888^{***}
Avg. I/E index	-0.981^{***}	-0.915^{***}	-0.715^{***}	-0.939^{***}	-0.936^{***}	-0.298^{***}
Modularity	-0.947^{***}	-0.794^{***}	-0.677^{***}	-0.945^{***}	-0.659^{***}	0.022^{***}
RWC	-0.888^{***}	-0.586^{***}	-0.839^{***}	-0.534^{***}	-0.707^{***}	-0.971^{***}
BCC	0.365^{***}	-0.342^{***}	-0.218^{***}	-0.276^{***}	-0.037^{***}	0.359^{***}
SSI	0.013^{*}	0.742^{***}	0.745^{***}	0.228^{***}	0.186^{***}	0.195^{***}

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

negative partial correlations for the intersection network, while only BCC is negatively (partially) correlated for the union network.

When communities are highly dense or sparse, there may be cases where there are very few (or even no) boundary nodes satisfying the conditions described in Section 4.2.5. for BCC. Furthermore, projected networks are inherently undirected. For all measures except the SSI, we transform the undirected shared knowledge and source networks to be directed. Then every edge between users is reciprocal, changing the behavior of the random walk in RWC.

Less polarized as density between groups increases

We evaluate the effect of between group density on each measure applied to each network by calculating the partial correlation [78]. This way we can measure the relationship between between group density and polarization level while controlling for average within group density. Table 4 contains partial correlation coefficients and significance.

Now we expect negative partial correlations for each measure and network type. This largely holds. In particular, the W/B index, average I/E index, RWC have significantly negative partial correlations for all six network types.

Interestingly, modularity reports more polarization when the density between groups is higher for the intersection network. BCC also reports less polarization when the between group density increases for the interaction and intersection network.

Finally, the SSI completely defies the expected relationship between increasing between group density and polarization. Notably, the SSI can be interpreted as a measure of which information spreads within groups [66]. Hence, changing the number of edges between groups does not necessarily alter the SSI as we expect.

Distribution of values

To compare the behavior of the proposed measures, we produce violin plots of the values they attain throughout simulation runs. All values are linearly re-scaled so

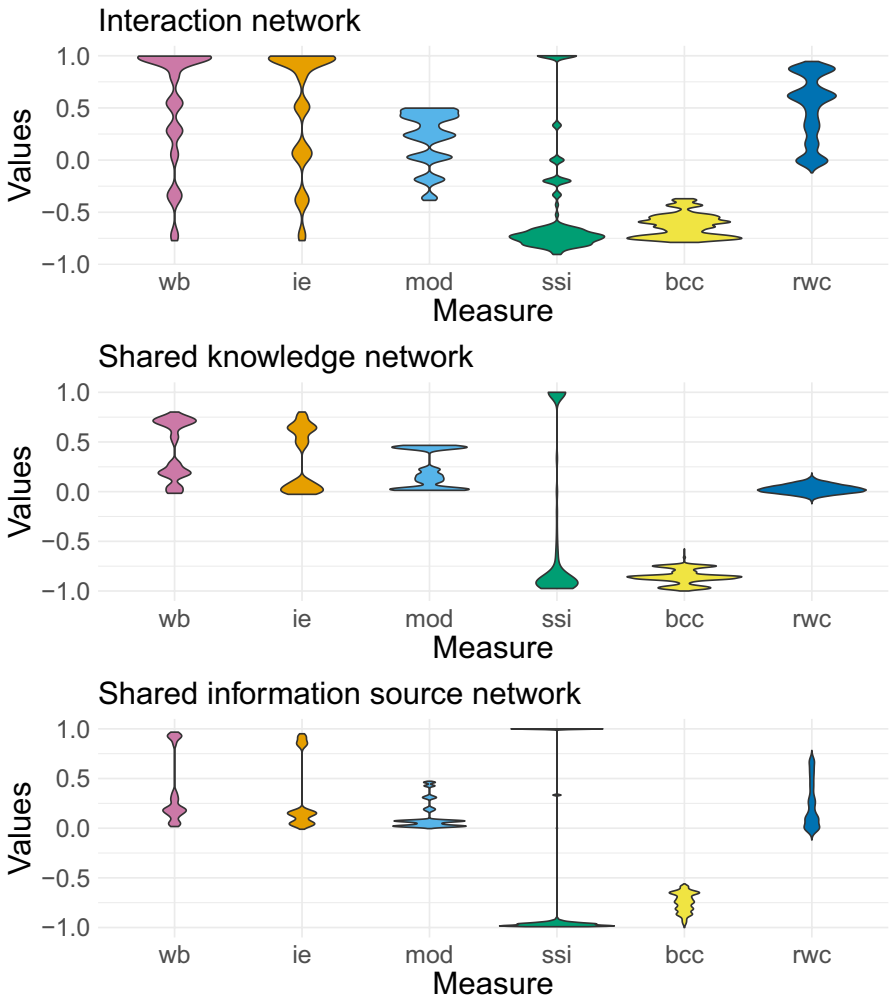

Fig. 2 Distribution of measure values for each measure applied to interaction, shared knowledge, and shared source networks

that they fell between -1 and +1, where -1 denotes the lowest possible polarization, and +1 indicates the highest possible polarization.

Figures 2 and 3 contain the distribution of values for each measure applied to each network type, in addition to each measure averaged across the interaction, shared knowledge, and shared source networks.

The W/B index and average I/E index follow similar patterns, which is unsurprising given their nearly identical formulas. The range of modularity values is more narrow than W/B index and average I/E index across networks. We see the SSI is highly biased towards the extremes, limiting the ability to compare polarization using the SSI values across contexts.

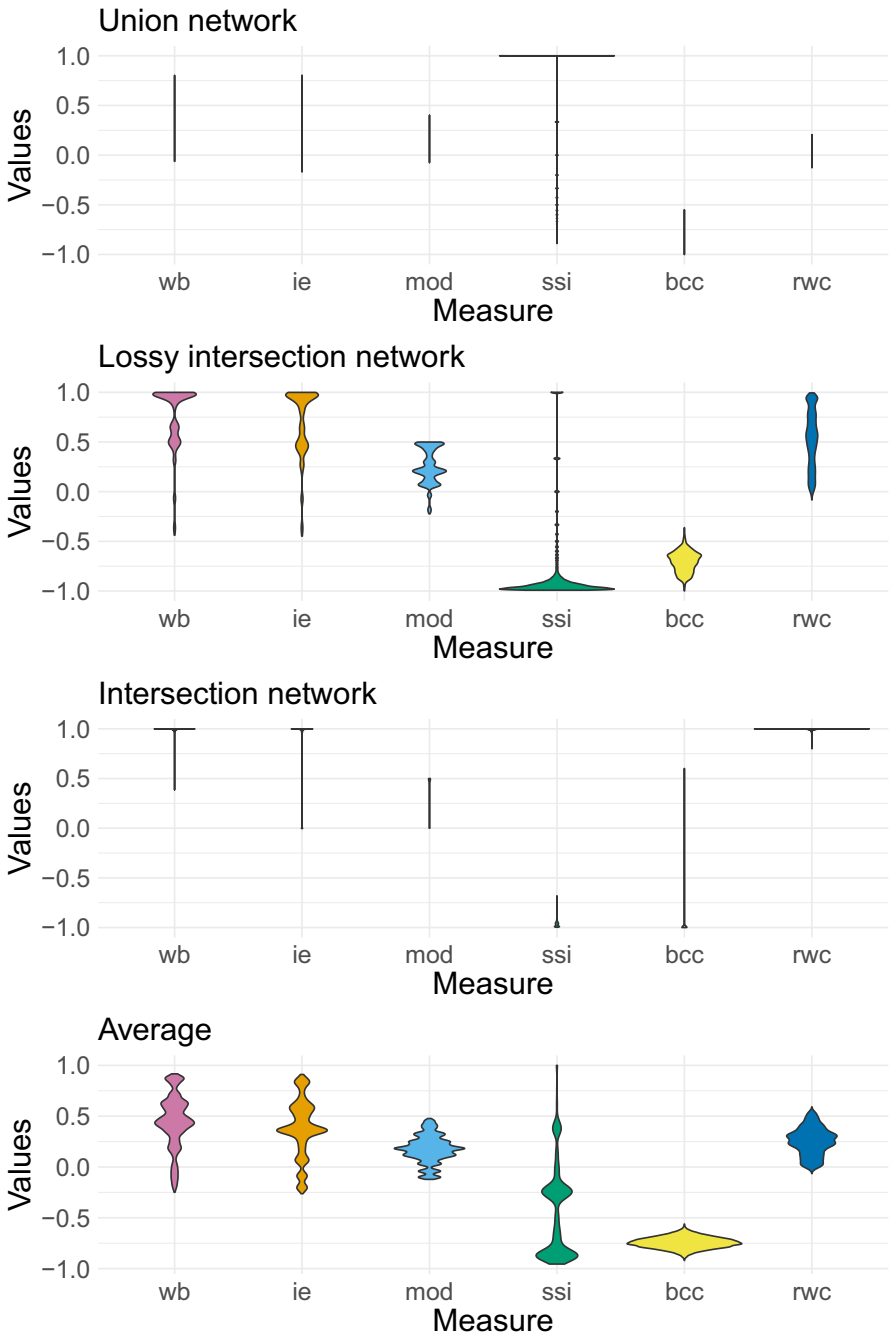

Fig. 3 Distribution of measure values for each measure applied to aggregated networks and average

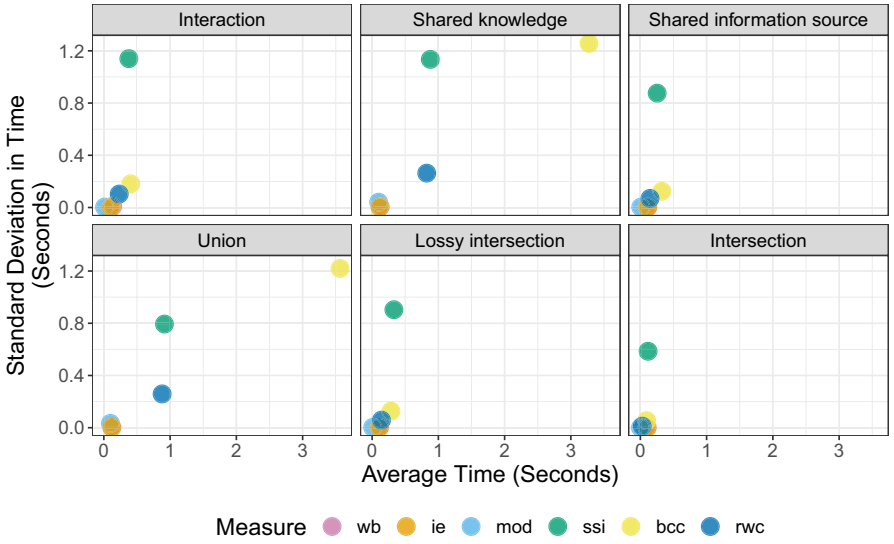

Fig. 4 Average and standard deviation in time (seconds) to apply each measure to each network type

BCC has a relatively narrow range and is less than 0 across networks. This emphasizes the need to investigate the range of actual values each measure reports (rather than simply the theoretical bounds). In addition, measures cannot necessarily be directly compared. A BCC value of 0.3 should be interpreted very differently than a W/B index value of 0.3.

RWC maintains a range appropriate for comparisons for the interaction network. For highly dense projected bipartite networks, like the shared knowledge network, the spread of RWC narrows considerably.

Computational efficiency

To assess the selected polarization measures, we also consider the time taken to run them across simulation experiments. More practical algorithms should run in less time and with less variance in the time taken. Figure 4 shows the arithmetic mean and standard deviation of time in seconds for each run of each measure. Overall, we favor measures featuring lower mean times and lower standard deviations, corresponding with faster and more stable measures. The W/B index, average I/E index, and modularity are consistently the fastest across network types. RWC is also relatively fast, but varies more for highly dense networks. Finally, BCC and the SSI are the slowest and most variable across networks.

Recommendations

We incorporate measure features with the results from the virtual experiment to assess the ability of each measure and aggregation technique to detect the sufficient

conditions for polarization to occur in List 1, support comparisons across contexts, and compute efficiently.

Beginning with the awareness condition, the W/B index, average I/E index, and SSI are equal to 1 if the communities have no edges between them, indicating a lack of awareness of opposition. This is not encapsulated by modularity, which does not equal 1 even when there are no edges between groups.

Technically, BCC is not applicable if there are no boundary nodes. The first condition for a node to be a boundary node is connection to the opposing group. The second condition is connection to at least one internal node within the node's group. Because of the second condition, no boundary nodes does not necessarily mean there are no connections between groups. Thus, BCC does not encapsulate awareness without further analysis of groups. In addition, RWC of 1 may denote a lack of awareness but requires more analysis due to randomness.

We turn now to polarization evaluation given changes in interactions, shared knowledge, and shared sources within and between groups. The W/B index and average I/E index directly assess the number of edges within communities relative to edges between communities, reflected by positive partial correlation with average within group density and negative partial correlation with between group density across non-aggregated and aggregated networks in Table 3 and Table 4.

Modularity, which compares the number of edges within communities to the expected number of edges due to random chance, also maintains positive partial correlation with average within group density and negative partial correlation with between group density across all non-aggregated networks and most aggregated networks. The exception is positive association with between group density for the intersection network. We also see in Fig. 2 that the range of modularity values is smaller than the range of W/B or average I/E index values, limiting comparisons across contexts.

RWC behaves as expected for interaction networks, but deviates for projected shared knowledge and shared source networks. Figure 2 shows highly dense and projected networks tend to greatly reduce the range of RWC values. Finally, SSI is biased towards extremes across network types and does not behave as expected as the density between groups changes.

All the measures are relatively fast (average ≤ 4 seconds) for the simulated networks. Compared to many real datasets, the simulated networks are very small with only 1000 nodes. Hence, the slowest measures (BCC and SSI) may be fast enough for small networks but could pose issues for networks with more nodes and edges.

In sum, the W/B index and average I/E index encapsulate awareness, consistently report more polarization when there are more connections within groups and fewer connections between groups, return a range of polarization values that support comparisons across contexts, and run quickly (average ~ 1 second).

Considering the aggregation techniques, a disadvantage of the averaging approach is that it does not speed up the polarization assessment process since measures must be applied to each dimension separately. Both union and intersection network aggregation result in large shifts in density. Typically, the projected bipartite networks (shared knowledge, shared source) are more dense than the interaction network. Thus, the interaction network can substantially constrain the intersection

network, while the projected networks can overwhelm the union. The lossy intersection seems to mitigate these extremes and limit shifts in density that impact the behavior of polarization measures, as seen in Fig. 2 and 3.

Case study

The following case study demonstrates our high-dimensional approach to assessing polarization between online communities. We apply the methodology to discussions surrounding COVID-19 vaccines on Twitter over time (before and during governmental emergency authorization in the U.S.). We proceed by applying the W/B and average I/E indices that we established are most appropriate for our framework through the virtual experiment. In addition, we use lossy intersection aggregation, as recommended.

Data collection

On December 11, 2020, the U.S. Food & Drug Administration issued the first (emergency) authorization of the Pfizer-BioNTech COVID-19 vaccine. We analyze the discussion on Twitter surrounding COVID-19 vaccines leading up to the initial emergency authorization of the Pfizer-BioNTech COVID-19 vaccine from December 1, 2020 thru December 14, 2020. Our data was collected via keyword searches using Twitter v1 API. Selected tweets contain at least one term referring to COVID-19 (coronaravirus, coronavirus, wuhan virus, wuhanvirus, 2019nCoV, NCoV, NCoV2019, covid-19, covid19, covid 19) *and* one term referring to vaccines (vaccine, vax, mRNA, autoimmuneencephalitis, vaccination, vaccinate, getvaccinated, covidisjustacold, autism, covidshotcount, dose1, dose2, VAERS, GBS, believemothers, mybodymychoice, thisisourshot, killthevirus, proscience, immunization, gotmyshot, igottheshot, covidvaccinated, beatcovid19, moderna, astrazeneca, pfizer, johnson & johnson, j &j, johnson and johnson, jandj).

In total, we have 436,474 users (346,329 pro-vaccine, 90,045 anti-vaccine), 12,979 knowledge bits (hashtags), 252,610 sources (URLs and @-mentions), and 2,959,920 tweets distributed across the 14 days of interest. Summary statistics for each day in the dataset can be found in Table 5 in Appendix A.

Generate high-dimensional and aggregated networks

In this step, we designate how edges are generated for each dimension: interaction, knowledge, and source. We generate a high-dimensional network for each day from December 1, 2020 thru December 14, 2020.

The interaction dimension represents users who retweet other users. We exclude other types of interactions available on Twitter, such as replies, because retweets typically indicate endorsements [41]. By retweeting a tweet, users are amplifying the ideas in that tweet to their audience.

Fig. 5 The W/B index and average I/E index applied to COVID-19 vaccine discussion from December 1-14 2020. The vertical light grey dashed line represents the day of the emergency authorization of the Pfizer-BioNTech COVID-19 vaccine

Our methodology purposely defines knowledge and sources broadly because of the variability in platform affordances and norms. On Twitter, hashtags serve a variety of communicative purposes, such as participating in a discussion or social movement or elaborating on the text in the tweet [79, 80]. We identify the set of hashtags that designate engagement with COVID-19 vaccine discourse as knowledge bits.

Second, we identify links to external websites contained in tweets. This indicates what media, government, and other entities users refer to for information, creating one set of sources. Another source of information is other Twitter users, so we use @-mentions as sources as well. We generate the shared source network with both types of sources at once. This demonstrates how to incorporate multiple types of sources, which is an analogous process for knowledge bits.

Finally, we aggregate the interaction, shared knowledge, and shared source dimensions by taking the lossy intersection.

Partition users into ideologically opposed communities

We use a semi-supervised stance detection algorithm that entails labelling the stance of hashtags, n-grams, URLs, and/or domains [81]. The two general stances in this dataset are pro or anti vaccines. Of course, many have much more nuanced views of the COVID-19 vaccines and public health measures generally. However, given the limited amount of information we have about each user, detecting a general stance towards COVID-19 vaccines is most appropriate.

Labels for our stance detection model come from previously validated terms [82]. Pro-vaccine hashtags include #Vaccines4All and #Iwillgetvaccinated. Anti-vaccine hashtags include #antivaxx and #RejectWeaponizedVaccines. We apply the stance detection method all 14 days at once.

Measure polarization of communities

Figure 5 provides the value of the average I/E index and W/B index applied to the interaction, shared knowledge, shared source, and lossy intersection networks on each day of our dataset.

Notably, the average I/E index is nearly always less than the corresponding W/B index. This is likely due to the effect of unequal group sizes. When one group is much larger than the other, it likely has many more edges internally in total than the smaller group. In the averaged measure, the relative number of edges within and between groups of both groups are weighted equally. Unequal groups skew the W/B index more.

The number of vaccine supporting users increases throughout the days preceding the vaccine authorization, while anti-vaccine users do not at the same rate. Thus, the communities become increasingly disparate in size and there is less agreement between the W/B index and average I/E index.

We focus on the average I/E index because it controls for group sizes. From December 1 thru December 7, pro-vaccine and anti-vaccine users are consistently highly polarized. They are not largely communicating, using similar knowledge, or using similar knowledge sources across camps. As the date gets close to December 11th, polarization decreases between groups due to more interactions and shared knowledge sources between groups. This trend continues for knowledge sources until the end of our dataset on December 14th, but not interactions or knowledge usage.

Broadly, the pro-vaccine group consists of public health officials and organizations, as well as members of the public supporting the vaccine rollout. The anti-vaccine group at this time was not as established. One reason may be that it did not have governmental support already in place. The emergency authorization drew attention towards COVID-19 vaccines becoming available to the public, but December 2020 was still relative early into the COVID-19 pandemic and COVID-19 vaccine rollout. When opinion leaders, such as official U.S. government accounts, made the monumental decision to give emergency authorization for the Pfizer-BioNTech COVID-19 vaccine, people who were excited and skeptical alike came to Twitter to comment. Even if people continue to use different hashtags closer to the authorization, they begin to use more similar URLs and mention the same users as the conversation converges.

By considering multiple dimensions of direct and indirect communication, we encapsulate the collapse in different knowledge sources across groups as pro and anti-vaccine users discuss the official announcement of the first emergency authorization of a COVID-19 vaccine. At the same time, both camps continue to use different knowledge in their posts, as their position towards the common knowledge sources are opposed.

Note: we do not suggest this analysis is representative of the world population's discourse around COVID-19 vaccines. Rather, it is a case study of English-language discourse around COVID-19 vaccines on Twitter in early December 2020.

Discussion

In this work, we introduce a high-dimensional approach to assess polarization online such that differences in communication, knowledge usage, and knowledge source usage within and between ideologically opposed groups is encapsulated. Through a virtual experiment, we evaluate six existing measures of network structure in our framework. The measures are applied to over 36,000 synthetic networks representing each dimension, as well as three aggregated networks (union, lossy intersection, intersection). We then assess their ability to efficiently capture polarization according to the definition laid out in Sect. 2.

Ultimately, the W/B index and average I/E index, the measures that directly assess the relative difference in connections within and between groups, consistently report more polarization when density within groups increases and between groups decreases such that awareness is accounted for and comparison across contexts is accessible. Additionally, these measures are consistently fast.

Furthermore, we recommend using the lossy intersection method of aggregating the social, shared knowledge, and shared source networks to avoid large density shifts, which we showed can cause measures to behave differently than expected. This technique limits the extent to which the interaction network constrains the aggregated network (as with intersection) or projected networks overwhelm the aggregated network (as with union). Our recommendation aligns with previous work that found aggregating layers using the AND (OR) operation is beneficial for dense (sparse) networks [49].

Crucially, the high-dimensional methodology supports evaluation of all five criteria of polarization established in List 1. Following the data collection step, high-dimensional network generation entails representing each dimension of collective narrative formation (social, knowledge, source) in network form. The measures of polarization, W/B index and average I/E index, assess the relative density of interactions, knowledge sharing, and knowledge source sharing within and between groups. These measures require the users are partitioned into distinct groups. For our purposes, these partitions are generated using some form of stance detection, thus incorporating ideologically opposed groups.

By encapsulating these criteria, our approach approximates the degree of ideological and psychological distancing between communities more directly than existing measures of online polarization.

Finally, we demonstrate that applying established measures to networks in this novel way reveals aspects of polarization in terms of content without relying on domain- or language-dependent methods. We show pro-vaccine and anti-vaccine users in the discussion surrounding the emergency authorization of Pfizer-BioNTech COVID-19 vaccine in early December 2020 become less polarized as announcements from the government and related organizations provide common knowledge sources to comment on. Yet both camps continue to use different knowledge in their posts, as their position towards the common knowledge sources are opposed.

The divergence in the level of polarization across dimensions emphasizes the importance of considering multiple ways division occurs through a

high-dimensional approach. Each dimension can be affected by events differently, which has implications for understanding drivers of online discussion dynamics. In this case, we see official communication collapses the online vaccine discourse around the same sources of information. As people develop their attitudes towards the COVID-19 vaccine, they interact with an ideologically diverse set of users about the (relatively) little information available at that time. Therefore, pro-vaccine and anti-vaccine users display less polarization overall during the days surrounding the authorization announcement despite consistently using different knowledge in their posts. Interpreting the polarization of each dimension and overall requires an understanding of contextual factors, such as the recentness of the debated issue.

Polarization is necessarily a dynamic phenomenon. The sufficient conditions of polarization described in List 1 can only arise due to social and cognitive processes that occur over time. Groups and collective narratives do not develop or disappear instantaneously. In this study, we see polarization as salient when division is systematically reproduced across group members.

That said, these measures can be applied over multiple time points to determine the extent to which there is consistent polarization, as we did in the case study. Alternatively, researchers can generate networks based on communication, knowledge and knowledge sources used within a range of days rather than a single day to detect persistent patterns of division.

The proposed methodology is flexible enough to allow for modular adjustments to the analytical steps. In the case study, we use hashtags as knowledge bits. However, researchers can use topics identified through topic modeling or qualitative analyses, keywords, or any other categorization of content in posts to characterize knowledge use within and between groups. Similarly, alternative definitions of communication between users and knowledge sources, as well as group membership, can be incorporated into our framework.

Our aim is to assist analyses of the wealth of data afforded by social media and other online platforms that provide new opportunities to understand how discourse and group dynamics evolve. Digital technologies are in constant evolution and require ongoing empirical investigations to capture changes in polarizing effects [7]. With the proposed high-dimensional framework to evaluate online polarization, researchers can investigate how communication, knowledge usage, and knowledge source citation within and between ideologically opposed communities is affected by related events [8] and de-polarization efforts (e.g., altering the social network through algorithmic bridging of users [10, 83]). Moreover, assessing polarization in terms of multiple dimensions could reveal how certain communities or topics are vulnerable to polarization. This would inform proactive interventions, rather than reactive ones, to improve resilience to division and extremism.

Limitations

Our choice of data, network representations of communication and shared knowledge/source usage, prevents direct comparison of opinion extremity and sentiment

expressed within and between groups. These are worthwhile future additions to our measure depending on the goal of the researchers. For example, the sentiment of an interaction between users provides additional information about the nature of their relationship (e.g., antagonistic, friendly, neutral) [84]. It also often requires language and domain specific knowledge to detect.

Moreover, we qualify our claims of language and domain independence as follows. Although most stance detection requires some level of supervision [85], more methods are being developed where manual labelling is not necessary [86, 87]. We expect these unsupervised tools will be developed further, but do not address the language and domain dependence of many existing stance detection methods in this work. The claim of language and domain dependence solely applies to the polarization measurement following the identification of ideologically opposing groups of users. However, stance detection is an essential step to assign users to ideologically opposing groups.

Appendix A. Daily summary statistics for case study

See Table 5 here.

Table 5 Summary statistics by day for case study

Day	Num. agents (pro/anti)	Num. knowledge bits	Num. knowledge sources	Num. tweets
1	29,735 (24,608/5,127)	1154	11,613	39,026
2	148,662 (125,795/22,867)	2960	32,931	217,071
3	134,119 (97,089/37,030)	3059	32,647	193,599
4	98,478 (67,176/31,302)	2812	28,286	138,318
5	74,779 (49,920/24,859)	2214	19,943	103,103
6	64,994 (42,033/22,961)	2116	18,469	89,166
7	90,699 (67,195/23,504)	2169	21,811	126,771
8	218,678 (190,184/28,494)	4646	58,346	339,309
9	174,855 (137,774/37,081)	4602	53,993	251,810
10	157,078 (117,757/39,321)	4194	50,507	226,507
11	149,258 (129,883/19,375)	4062	48,448	214,403
12	101,024 (88,791/12,233)	3501	36,444	139,595
13	132,637 (118,339/14,298)	3227	37,119	187,659
14	171,942 (156,955/14,987)	3984	49,620	249,248

Author contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by SCP with assistance from JU. The first draft of the manuscript was written by SCP and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Open Access funding provided by Carnegie Mellon University. The research for this paper was supported in part by the Knight Foundation (G-2019-58792) and the Office of Naval Research under the following projects MURI:FACTIONS:Near Real Time Assessment of Emergent Complex Systems of Confederates (N000141712675); Minerva-Multi-Level Models of Covert Online Information Campaigns (N000142112765); Group Polarization in Social Media (N000141812106). This paper is also supported by the center for Informed Democracy and Social-cybersecurity (IDeaS) and the Center for Computational Analysis of Social and Organizational Systems (CASOS) at Carnegie Mellon University.

Availability of data and materials The datasets analysed during the current study are available from the corresponding author on reasonable request.

Code availability Code will be made available upon acceptance.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (2019). The origins and consequences of affective polarization in the united states. *Annual review of political science, 22*, 129–146. https://doi.org/10.1146/annurev-polisci-051117-073034
2. Tucker, J. A., Guess, A., Barbera, P., Vaccari, C., Siegel, A., Sanovich, S., Stukal, D., & Nyhan, B. (2018). Social media, political polarization, and political disinformation: A review of the scientific literature (March 19, 2018). *Political polarization, and political disinformation: a review of the scientific literature March*.https://doi.org/10.2139/ssrn.3144139
3. Kingzette, J., et al. (2021). How affective polarization undermines support for democratic norms. *Public Opinion Quarterly, 85*(2), 663–677. https://doi.org/10.1093/poq/nfab029
4. Vicario, M. D., Quattrociocchi, W., Scala, A., & Zollo, F. (2019). Polarization and fake news: Early warning of potential misinformation targets. *ACM Transactions on the Web, 13*(2), 1–22. https://doi.org/10.1145/3316809
5. Sunstein, C. R. *# Republic* (Princeton university press, 2018).
6. 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
7. Barberá, P. (2020). Social media, echo chambers, and political polarization. *Social media and democracy. The state of the field, prospects for reform*.https://doi.org/10.1017/9781108890960

8. Darwish, K. (2019). Quantifying polarization on twitter: the kavanaugh nomination. *Proceedings of International Conference on Social Informatics*. https://doi.org/10.1007/978-3-030-34971-4_13
9. 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
10. Garimella, K., De Francisci Morales, G., Gionis, A., & Mathioudakis, M. (2017). Reducing controversy by connecting opposing views. *ACM International Conference on Web Search and Data Mining, 10*, 81–90. https://doi.org/10.24963/ijcai.2018/731
11. Fiorina, M. P., Abrams, S. J., et al. (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
12. Esteban, J.-M., & Ray, D. (1994). On the measurement of polarization. *Econometrica: Journal of the Econometric Society*. https://doi.org/10.2307/2951734
13. Lelkes, Y. (2016). Mass polarization: Manifestations and measurements. *Public Opinion Quarterly, 80*(S1), 392–410. https://doi.org/10.1093/poq/nfw005
14. Bliuc, A.-M., Bouguettaya, A., & Felise, K. D. (2021). Online intergroup polarization across political fault lines: An integrative review. *Frontiers in Psychology, 12*, 641215. https://doi.org/10.3389/fpsyg.2021.641215
15. DiMaggio, P., Evans, J., & Bryson, B. (1996). Have american's social attitudes become more polarized? *American journal of Sociology, 102*(3), 690–755. https://doi.org/10.1086/230995
16. Myers, D. G., & Lamm, H. (1976). The group polarization phenomenon. *Psychological bulletin, 83*(4), 602. https://doi.org/10.1037/0033-2909.83.4.602
17. Hogg, M. A., & Reid, S. A. (2006). Social identity, self-categorization, and the communication of group norms. *Communication theory, 16*(1), 7–30. https://doi.org/10.1111/j.1468-2885.2006.00003.x
18. Tajfel, H., Turner, J. C., Austin, W. G., & Worchel, S. (1979). An integrative theory of intergroup conflict. *Organizational identity: A reader, 56*(65), 9780203505984–16.
19. Barua, Z., Barua, S., Aktar, S., Kabir, N., & Li, M. (2020). Effects of misinformation on covid-19 individual responses and recommendations for resilience of disastrous consequences of misinformation. *Progress in Disaster Science, 8*, 100119. https://doi.org/10.1016/j.pdisas.2020.100119
20. An, J., Quercia, D., & Crowcroft, J. (2014). Partisan sharing: Facebook evidence and societal consequences. *Proceedings of the second ACM conference on online social networks*. https://doi.org/10.1145/2660460.2660469
21. Jost, J. T., Baldassarri, D. S., & Druckman, J. M. (2022). Cognitive-motivational mechanisms of political polarization in social-communicative contexts. *Nature Review Psychology*. https://doi.org/10.1038/s44159-022-00093-5
22. Baldassarri, D., & Gelman, A. (2008). Partisans without constraint: Political polarization and trends in american public opinion. *American Journal of Sociology, 114*(2), 408–446. https://doi.org/10.1086/590649
23. Flores, A., et al. (2022). Politicians polarize and experts depolarize public support for COVID-19 management policies across countries. *Proceedings of the National Academy of Sciences, 119*(3), e2117543119. https://doi.org/10.1073/pnas.2117543119
24. Spohr, D. (2017). Fake news and ideological polarization: Filter bubbles and selective exposure on social media. *Business information review, 34*(3), 150–160. https://doi.org/10.1177/0266382117722446
25. Du, S., & Gregory, S. (2016). The echo chamber effect in twitter: does community polarization increase? *International workshop on complex networks and their applications*. https://doi.org/10.1007/978-3-319-50901-3_30
26. Guess, A., Nyhan, B., Lyons, B., & Reifler, J. (2018). Avoiding the echo chamber about echo chambers. *Knight Foundation, 2*(1), 1–25.
27. Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on facebook. *Science, 348*(6239), 1130–1132. https://doi.org/10.1126/science.aaa1160
28. Wilson, A. E., Parker, V. A., & Feinberg, M. (2020). Polarization in the contemporary political and media landscape. *Current Opinion in Behavioral Sciences, 34*, 223–228. https://doi.org/10.1016/j.cobeha.2020.07.005
29. 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

30. Heatherly, K. A., Lu, Y., & Lee, J. K. (2017). Filtering out the other side? cross-cutting and like-minded discussions on social networking sites. *New Media & Society, 19*(8), 1271–1289. https://doi.org/10.1177/1461444816634677
31. Bail, C. A., et al. (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences, 115*(37), 9216–9221. https://doi.org/10.1073/pnas.1804840115
32. Suhay, E., Bello-Pardo, E., & Maurer, B. (2018). The polarizing effects of online partisan criticism: Evidence from two experiments. *The International Journal of Press/Politics, 23*(1), 95–115. https://doi.org/10.1177/1940161217740697
33. Garrett, R. K., et al. (2014). Implications of pro-and counterattitudinal information exposure for affective polarization. *Human communication research, 40*(3), 309–332. https://doi.org/10.1111/hcre.12028
34. Mosleh, M., Martel, C., Eckles, D., & Rand, D. G. (2021). Shared partisanship dramatically increases social tie formation in a twitter field experiment. *Proceedings of the National Academy of Sciences, 118*(7), e2022761118. https://doi.org/10.1073/pnas.2022761118
35. Rathje, S., Van Bavel, J. J., & Van Der Linden, S. (2021). Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences, 118*(26), e2024292118. https://doi.org/10.1073/pnas.2024292118
36. Conover, M., et al. (2011). Political polarization on twitter. *Proceedings of the international aaai conference on web and social media, 5*(1), 89–96. https://doi.org/10.1609/icwsm.v5i1.14126
37. Uyheng, J. & Carley, K.M. (2020) Bot impacts on public sentiment and community structures: Comparative analysis of three elections in the Asia-Pacific. *International conference on social computing, behavioral-cultural modeling and prediction and behavior representation in modeling and simulation* 12–22. https://doi.org/10.1007/978-3-030-61255-9_2 .
38. Guerra, P., Meira, W., Jr., Cardie, C., & Kleinberg, R. (2013). A measure of polarization on social media networks based on community boundaries. *Proceedings of the international AAAI conference on web and social media, 7*(1), 215–224. https://doi.org/10.1609/icwsm.v7i1.14421
39. Bojanowski, M., & Corten, R. (2014). Measuring segregation in social networks. *Social networks, 39*, 14–32. https://doi.org/10.1016/j.socnet.2014.04.001
40. Coletto, M., Garimella, K., Gionis, A., & Lucchese, C. (2017). Automatic controversy detection in social media: A content-independent motif-based approach. *Online Social Networks and Media, 3*, 22–31. https://doi.org/10.1016/j.osnem.2017.10.001
41. Garimella, K., Morales, G. D. F., Gionis, A., & Mathioudakis, M. (2018). Quantifying controversy on social media. *ACM Transactions on Social Computing, 1*(1), 1–27. https://doi.org/10.1145/3140565
42. Emamgholizadeh, H., Nourizade, M., Tajbakhsh, M. S., Hashminezhad, M., & Esfahani, F. N. (2020). A framework for quantifying controversy of social network debates using attributed networks: biased random walk (BRW). *Social Network Analysis and Mining, 10*(1), 1–20. https://doi.org/10.1007/s13278-020-00703-1
43. Škrlj, B., & Renoust, B. (2020). Layer entanglement in multiplex, temporal multiplex, and coupled multilayer networks. *Applied Network Science, 5*(1), 1–34. https://doi.org/10.1007/s41109-020-00331-w
44. Berlingerio, M., Coscia, M., Giannotti, F., Monreale, A., & Pedreschi, D. (2013). Multidimensional networks: foundations of structural analysis. *World Wide Web, 16*, 567–593. https://doi.org/10.1007/s11280-012-0190-4
45. Carley, K. M., Lee, J.-S., & Krackhardt, D. (2002). Destabilizing networks. *Connections, 24*(3), 79–92.
46. Benigni, M. C., Joseph, K., & Carley, K. M. (2019). Bot-ivistm: assessing information manipulation in social media using network analytics. *Emerging research challenges and opportunities in computational social network analysis and mining*. https://doi.org/10.1007/978-3-319-94105-9_2
47. Berlingerio, M., Pinelli, F., & Calabrese, F. (2013). Abacus: frequent pattern mining-based community discovery in multidimensional networks. *Data Mining and Knowledge Discovery, 27*, 294–320. https://doi.org/10.1007/s10618-013-0331-0
48. Interdonato, R., Magnani, M., Perna, D., Tagarelli, A., & Vega, D. (2020). Multilayer network simplification: approaches, models and methods. *Computer Science Review, 36*, 100246. https://doi.org/10.1016/j.cosrev.2020.100246

49. Taylor, D., Shai, S., Stanley, N., & Mucha, P. J. (2016). Enhanced detectability of community structure in multilayer networks through layer aggregation. *Physical review letters, 116*(22), 228301. https://doi.org/10.1103/PhysRevLett.116.228301
50. Fiori, K. L., Smith, J., & Antonucci, T. C. (2007). Social network types among older adults: A multidimensional approach. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 62*(6), P322–P330. https://doi.org/10.1093/geronb/62.6.P322
51. Diesner, J. & Carley, K. M. (2004) Using network text analysis to detect the organizational structure of covert networks. *Proceedings of the North American Association for Computational Social and Organizational Science (NAACSOS) Conference* **3**.
52. Berlingerio, M., Coscia, M. & Giannotti, F. (2011) Finding and characterizing communities in multidimensional networks. *2011 international conference on advances in social networks analysis and mining* 490–494. https://doi.org/10.1109/ASONAM.2011.104 .
53. Serrano-Contreras, I.-J., García-Marín, J., & Luengo, Ó. G. (2020). Measuring online political dialogue does: polarization trigger more deliberation? *Media and Communication, 8*(4), 63–72. https://doi.org/10.17645/mac.v8i4.3149
54. Belcastro, L., Cantini, R., Marozzo, F., Talia, D., & Trunfio, P. (2020). Learning political polarization on social media using neural networks. *IEEE Access, 8*, 47177–47187. https://doi.org/10.1109/ACCESS.2020.2978950
55. Morales, A. J., Borondo, J., Losada, J. C., & Benito, R. M. (2015). Measuring political polarization: Twitter shows the two sides of venezuela. *Chaos: An Interdisciplinary Journal of Nonlinear Science, 25*(3), 033114. https://doi.org/10.1063/1.4913758
56. Massey, D. S., & Denton, N. A. (1988). The dimensions of residential segregation. *Social forces, 67*(2), 281–315. https://doi.org/10.1093/sf/67.2.281
57. McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual review of sociology*. https://doi.org/10.1146/annurev.soc.27.1.415
58. Henry, A. D., Prałat, P., & Zhang, C.-Q. (2011). Emergence of segregation in evolving social networks. *Proceedings of the National Academy of Sciences, 108*(21), 8605–8610. https://doi.org/10.1073/pnas.101448610
59. Dandekar, P., Goel, A., & Lee, D. T. (2013). Biased assimilation, homophily, and the dynamics of polarization. *Proceedings of the National Academy of Sciences, 110*(15), 5791–5796. https://doi.org/10.1073/pnas.1217220110
60. Interian, R., Marzo, R., Mendoza, I., & Ribeiro, C. C. (2022). Network polarization, filter bubbles, and echo chambers: an annotated review of measures and reduction methods. *International Transactions in Operational Research*. https://doi.org/10.1111/itor.13224
61. Chartishvili, A. G., Kozitsin, I. V., Goiko, V. L. & Saifulin, E. R. (2019) On an approach to measure the level of polarization of individuals' opinions. *2019 Twelfth International Conference" Management of large-scale system development"(MLSD)* 1–5. https://doi.org/10.1109/MLSD.2019.8911015 .
62. Bramson, A., et al. (2016). Disambiguation of social polarization concepts and measures. *The Journal of Mathematical Sociology, 40*(2), 80–111. https://doi.org/10.1080/0022250X.2016.1147443
63. Krackhardt, D., & Stern, R. N. (1988). Informal networks and organizational crises: An experimental simulation. *Social psychology quarterly*. https://doi.org/10.2307/2786835
64. Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the national academy of sciences, 103*(23), 8577–8582. https://doi.org/10.1073/pnas.0601602103
65. Fershtman, M. (1997). Cohesive group detection in a social network by the segregation matrix index. *Social Networks, 19*(3), 193–207. https://doi.org/10.1016/S0378-8733(96)00295-X
66. Echenique, F., & Fryer, R. G., Jr. (2007). A measure of segregation based on social interactions. *The Quarterly Journal of Economics, 122*(2), 441–485. https://doi.org/10.1162/qjec.122.2.441
67. Rajabi, A., Mantzaris, A. V., Atwal, K. S., & Garibay, I. (2021). Exploring the disparity of influence between users in the discussion of brexit on twitter: Twitter influence disparity in brexit if so, write it here. *Journal of Computational Social Science, 4*, 903–917. https://doi.org/10.1007/s42001-021-00112-0
68. Borgatti, S. P., Carley, K. M., & Krackhardt, D. (2006). On the robustness of centrality measures under conditions of imperfect data. *Social networks, 28*(2), 124–136. https://doi.org/10.1016/S0378-8733(99)00019-2

69. Takesue, H. (2020). From defection to ingroup favoritism to cooperation: simulation analysis of the social dilemma in dynamic networks. *Journal of Computational Social Science, 3*(1), 189–207. https://doi.org/10.1007/s42001-019-00058-4
70. Sasahara, K., et al. (2021). Social influence and unfollowing accelerate the emergence of echo chambers. *Journal of Computational Social Science, 4*(1), 381–402. https://doi.org/10.1007/s42001-020-00084-7
71. Borgatti, S. P., & Everett, M. G. (2000). Models of core/periphery structures. *Social networks, 21*(4), 375–395. https://doi.org/10.1016/S0378-8733(99)00019-2
72. Elliott, A., Chiu, A., Bazzi, M., Reinert, G., & Cucuringu, M. (2020). Core-periphery structure in directed networks. *Proceedings of the Royal Society A, 476*(2241), 20190783. https://doi.org/10.1098/rspa.2019.0783
73. Rombach, M. P., Porter, M. A., Fowler, J. H., & Mucha, P. J. (2014). Core-periphery structure in networks. *SIAM Journal on Applied mathematics, 74*(1), 167–190. https://doi.org/10.1137/120881683
74. Barberá, P., et al. (2015). The critical periphery in the growth of social protests. *PloS one, 10*(11), e0143611. https://doi.org/10.1371/journal.pone.0143611
75. Bastos, M., Piccardi, C., Levy, M., McRoberts, N., & Lubell, M. (2018). Core-periphery or decentralized? topological shifts of specialized information on twitter. *Social Networks, 52*, 282–293. https://doi.org/10.1016/j.socnet.2017.09.006
76. Yang, J., Zhang, M., Shen, K. N., Ju, X., & Guo, X. (2018). Structural correlation between communities and core-periphery structures in social networks: Evidence from twitter data. *Expert Systems with Applications, 111*, 91–99. https://doi.org/10.1016/j.eswa.2017.12.042
77. Gallagher, R. J., Young, J.-G., & Welles, B. F. (2021). A clarified typology of core-periphery structure in networks. *Science Advances, 7*(12), eabc9800. https://doi.org/10.1126/sciadv.abc9800
78. Epskamp, S., & Fried, E. I. (2018). A tutorial on regularized partial correlation networks. *Psychological methods, 23*(4), 617. https://doi.org/10.1037/met0000167
79. Wikström, P. (2014). # srynotfunny: Communicative functions of hashtags on twitter. *SKY Journal of Linguistics, 27*, 127–152.
80. Xiong, Y., Cho, M., & Boatwright, B. (2019). Hashtag activism and message frames among social movement organizations: Semantic network analysis and thematic analysis of twitter during the# metoo movement. *Public relations review, 45*(1), 10–23. https://doi.org/10.1016/j.pubrev.2018.10.014
81. Williams, E. M., & Carley, K. M. (2022). Tspa: Efficient target-stance detection on twitter. *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. https://doi.org/10.1109/ASONAM55673.2022.10068608
82. Blane, J. T., Bellutta, D., & Carley, K. M. (2022). Social-cyber maneuvers during the covid-19 vaccine initial rollout: content analysis of tweets. *Journal of Medical Internet Research, 24*(3), e34040. https://doi.org/10.2196/34040
83. Ravandi, B., & Mili, F. (2019). Coherence and polarization in complex networks. *Journal of Computational Social Science, 2*, 133–150. https://doi.org/10.1007/s42001-019-00036-w
84. Keuchenius, A., Törnberg, P., & Uitermark, J. (2021). Why it is important to consider negative ties when studying polarized debates: A signed network analysis of a dutch cultural controversy on twitter. *PloS one, 16*(8), e0256696. https://doi.org/10.1371/journal.pone.0256696
85. Küçük, D., & Can, F. (2020). Stance detection: A survey. *ACM Computing Surveys (CSUR), 53*(1), 1–37. https://doi.org/10.1145/3369026
86. Darwish, K., Stefanov, P., Aupetit, M., & Nakov, P. (2020). Unsupervised user stance detection on twitter. *Proceedings of the International AAAI Conference on Web and Social Media, 14*, 141–152. https://doi.org/10.1609/icwsm.v14i1.7286
87. Kobbe, J., Hulpuş, I. & Stuckenschmidt, H. (2020) Unsupervised stance detection for arguments from consequences. *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)* 50–60. https://doi.org/10.18653/v1/2020.emnlp-main.4 .