


WILEY

INTERNATIONAL  
TRANSACTIONS  
IN OPERATIONAL  
RESEARCHIntl. Trans. in Op. Res. 30 (2023) 3122–3158  
DOI: 10.1111/itor.13224

# Network polarization, filter bubbles, and echo chambers: an annotated review of measures and reduction methods

Ruben Interian\* , Rusl  n G. Marzo, Isela Mendoza and Celso C. Ribeiro*Institute of Computing, Universidade Federal Fluminense, Niter  i, RJ 24210-346, Brazil**E-mail: rintarian@id.uff.br [Interian]; ruslangm@id.uff.br [G. Marzo]; imendoza@id.uff.br [Mendoza]; celso@ic.uff.br [Ribeiro]*

Received 15 August 2022; accepted 6 October 2022

## Abstract

Polarization arises when the underlying network connecting the members of a community or society becomes characterized by highly connected groups with weak intergroup connectivity. The increasing polarization, the strengthening of echo chambers, and the isolation caused by information filters in social networks are increasingly attracting the attention of researchers from different areas of knowledge such as computer science, economics, and social and political sciences. This work presents an annotated review of network polarization measures and models used to handle the polarization. Several approaches for measuring polarization in graphs and networks were identified, including those based on homophily, modularity, random walks, and balance theory. The strategies used for reducing polarization include methods that propose edge or node editions (including insertions or deletions as well as edge weight modifications), changes in social network design, or changes in the recommendation systems embedded in these networks.

**Keywords:** social networks; network polarization; echo chambers; filter bubbles; polarization measures; polarization reduction; network analysis; network optimization

## 1. Introduction

Polarization is a well-known phenomenon increasingly attracting the attention of the media, politicians, influencers, and researchers. According to the Oxford Dictionary, *polarization is the division of a group (or a society, or a network) into sharply contrasting subgroups, communities, or sets of opinions or beliefs* (Oxford, 2021).

Some degree of polarization is unavoidable in any democratic system. The excess of political homogeneity may eliminate the presence of democratic alternatives (Sunstein, 2003). However, extreme polarization can lead to gridlocks or even violent conflicts (Garcia et al., 2015).

\*Corresponding author.

In his farewell address back in 1796, George Washington predicted that factions, or monolithic parties, would yield political sectarianism (Washington, 1999). In the 19th century, John Stuart Mill claimed that the dialogue across lines of political difference is a key prerequisite for sustaining a democratic citizenry (Mill, 1859). Hannah Arendt asseverated that debate is irreplaceable for forming enlightened opinions that reach beyond the limits of one's own subjectivity to incorporate the standpoints of others (Arendt, 1968). World leaders have often expressed concern about raising social and political polarization (Guterres, 2018). From sociologists to economists, from politicians to the media, many are interested in studying the behavior and interactions in social networks that rule the opinion formation process.

Two of the main factors that shape people's opinions are confirmation bias and social influence (Liu et al., 2021). According to Vicario et al. (2017), the observed polarization of offline and online communities might be the result of the conjugate effect of these two forces.

Confirmation bias is the tendency to process information by seeking or interpreting only those facts consistent with one's existing beliefs (Encyclopedia Britannica, 2021). In short, confirmation bias tends to favor information people are already convinced of. Even though this phenomenon may be largely unintentional, it often results in completely ignoring part of the existing information, causing significantly less contact with contradicting viewpoints. This isolation caused by confirmation bias and other information filters, such as content recommendation systems, is called by the term *filter bubble*. Unlike confirmation bias, social influence is the process under which one's opinions or behaviors are actually affected by others (Gass, 2015; Vicario et al., 2017).

Polarization is characterized by an increasing intragroup agreement (i.e., between individuals with similar beliefs), while, at the same time, there is a deepening intergroup disagreement (i.e., between individuals identified with groups with contrary beliefs) (Buskens et al., 2008). Social networks and mass media are places where this phenomenon manifests itself in a strong way (Interian and Ribeiro, 2018). However, polarization and hostility are increasingly shifting from social media to the real world, as it was demonstrated by several political events, such as the protests of the Yellow Vest movement (France 24, 2019) in France in 2018, the protests after George Floyd's death (The New York Times, 2021) in 2020, the U.S. Capitol attack (The Washington Post, 2021) in 2021, and the convoy protests (BBC News, 2022) in Canada in 2022.

Studies have investigated the echo chamber effect on information spreading, showing that some groups can transmit information, on average, to a larger audience than others (Cota et al., 2019). Polarized groups are often related to the increased spreading of fake news. However, polarization and fake news spreading are different processes. Polarization is about strengthening and isolating groups or communities, while fake news propagation reflects specific information diffusion within these groups. In addition, fake news may propagate even in the absence of polarization (in the form of rumors or misinformation), although high polarization levels facilitate their spreading.

Community detection methods (Newman and Girvan, 2004; Newman, 2006b) are closely related to polarization detection and measurement. Studies and reviews about community detection methods appeared, for example, in Yang et al. (2016) and El-Moussaoui et al. (2019). However, community and polarization detection should not be confused. Community detection amounts to the identification of membership in groups. On the other hand, when detecting or measuring polarization, the attributes reflecting group membership are generally already known or estimated, and one seeks to identify the strength of intergroup and intragroup connections.

In the literature, groups formed around a shared narrative are frequently called echo chambers. As defined by Cinelli et al. (2021), echo chambers are environments in which the opinions or beliefs of people about some topic are reinforced due to repeated interactions with peers or sources having similar tendencies and attitudes. Some social networks show a massive presence of echo chambers, while, in others, their presence is reduced (Morales et al., 2021). The terms *group* (Currarini et al., 2009), *community* (Newman, 2006b), *gated community* (Turow, 1997), *filter bubble* (Spohr, 2017), and *echo chamber* (Cinelli et al., 2021), have different shades of meaning but are often used as synonyms in the literature.

Articles about different topics on the general subject of polarization appear in journals from different areas of knowledge, such as computer science (Garimella et al., 2016; Interian et al., 2021), economics (Currarini et al., 2009; Kawada et al., 2018), or social and political sciences (Maoz and Somer-Topcu, 2010; Flaxman et al., 2016). Researchers refer to phenomena related to polarization using different terms such as *controversy* (Garimella et al., 2016), *disagreement* (Chen and Racz, 2021), *conflict* (Rumshisky et al., 2017), and even *cyberbalkanization* (Bozdog et al., 2014), in addition to polarization itself. We will mostly treat them as synonymous unless otherwise stated.

Polarization manifests mainly in mass media and social, interaction, and collaboration networks. In this review, we are interested in exploring *network polarization* specifically. It is defined as a phenomenon in which the underlying network connecting the members of a society or community is composed of highly connected groups with weak intergroup connectivity (Conover et al., 2012). The polarization of posts or users of a social network may also be assessed independently, without considering the underlying graph or network structure. In this case, the profile of each post or user is evaluated, but the connections between them are not used.

The number of mass media articles, scientific papers, and books about topics such as the increasing polarization and the strengthening of echo chambers and filter bubbles in social networks is growing year by year, as illustrated by Fig. 1. There have been hundreds of publications in the past two decades in this area.

The goal of this annotated review is twofold. First, we identify the most used network polarization measures and the strategies used to handle the polarization problem, together with their main applications. Second, we present a comprehensive and annotated list of publications related to the evaluation of the polarization strength and to strategies used to handle the polarization.

The review is organized as follows. Section 2 describes the methodology applied to collect the publications that propose or use network polarization measures or strategies to handle the polarization. It also includes the queries and digital libraries we used, together with some quantitative results. The polarization measures found in the reviewed papers are presented in Section 3. Different approaches used to handle the polarization problem by using interventions, modifying the recommendation algorithms, or redesigning the network are described in Section 4. Section 5 presents the concluding remarks, which also include short comments about real-life case studies and practical applications of polarization measures and models.

## 2. Methodology

As a first step for creating this review, we conducted a systematic literature mapping (Wohlin et al., 2012). We identified publications that propose or use network polarization measures or strategies to

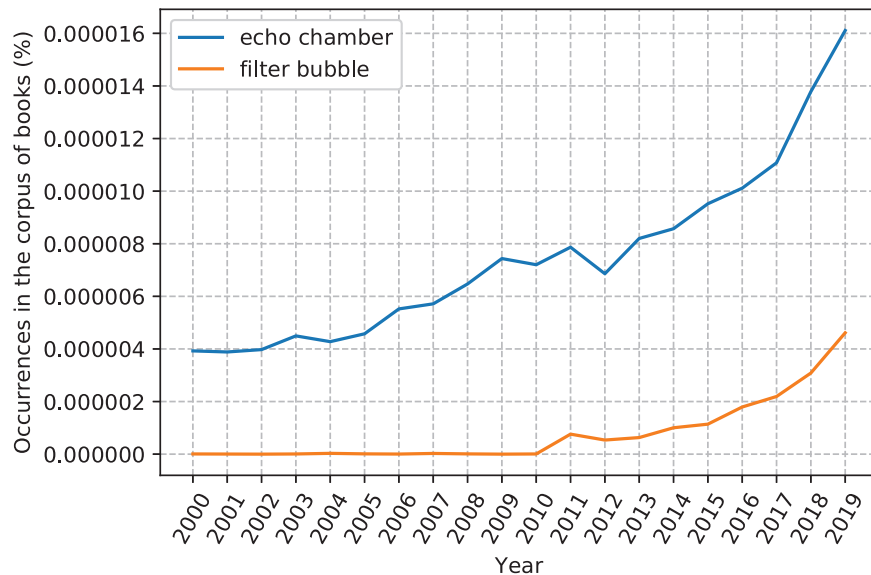


Fig. 1. Evolution of occurrences of the terms “echo chamber” and “filter bubble” from 2000 to 2019 in the corpus of books stored by Google in its digital database. The y-axis shows, of all the bigrams contained in Google’s sample of books written predominantly in English and published in any country, what percentage of them are “echo chamber” or “filter bubble” (source: Google Books Ngram Viewer (Google, 2019)).

handle the polarization. This section presents the systematic mapping planning and the quantitative results we achieved.

The universe of relevant analyzed publications consisted of journal and conference papers about the mathematical or computational modeling of network polarization. From the broad specter of publications dealing with polarization, we are specifically interested in those that present or use polarization measures, as well as in models that use them to handle polarization, from a theoretical or a practical perspective.

Polarization may be understood as an existing phenomenon that can be modeled and measured. However, this concept can also be used to refer to the process of increasing the division between separate groups of individuals, parties, or other entities. In the first case, polarization can be seen as the state of some networks. In the second, the polarization process is dynamic and subject to external modifications. These modifications, commonly known as external interventions (Gillani et al., 2018), aim to bring the network to a new, different state. In most cases, the goal of an intervention is to reduce a specific polarization measure. In this annotated review, we are also interested in publications that model these interventions, reducing or increasing some explicit or implicit polarization measures. Therefore, our two research questions were as follows:

- Q1. What approaches have been proposed for measuring network polarization?
- Q2. What network polarization reduction methods have been suggested?

We chose the following set of initial keywords: polarization, network, measure, intervention, reduce, and increase.

### 2.1. Search string and digital library

The Scopus digital library was chosen for retrieving the publications for this review. It is one of the most used digital libraries and provides access to journals, conference proceedings, and book chapters from ACM, IEEE, Springer, Elsevier, and other publishers. Compared to other digital libraries and search engines such as the Web of Science and Google Scholar, Scopus offers a good balance between a broad coverage of publication venues and their quality. Scopus indexes nearly the entire ScienceDirect database (Elsevier, 2018). Web of Science is more restrictive when choosing the scientific journals it covers. On the other hand, Google Scholar includes many nonpeer-reviewed sources.

The Scopus advanced document search engine allows performing complex search queries using Boolean operators, approximate phrases, and field codes to narrow the scope of the search. Our search string was built using the initial keywords and using the above features, as presented below:

```
TITLE-ABS-KEY (
(polarization OR "echo chamber*" OR "filter bubble*")
AND (graph OR graphs OR network OR networks)
AND (metric* OR measure OR reduce* OR reduction OR increase OR intervention*))

AND ALL (
(polarization OR "echo chamber*" OR "filter bubble*")
AND NOT "cell network*" AND NOT antenna* AND NOT radar AND NOT "cell po-
larization"
AND NOT electromagnetic* AND NOT electric AND NOT optical)

AND (
LIMIT-TO(SUBJAREA, "MATH")
OR LIMIT-TO(SUBJAREA, "COMP")
OR LIMIT-TO(SUBJAREA, "SOCI")
OR LIMIT-TO(SUBJAREA, "MULT"))
AND (EXCLUDE(SUBJAREA, "CHEM"))
AND (EXCLUDE(SUBJAREA, "EART"))
```

In the first component of the search string, the TITLE-ABS-KEY field code is used for finding the most relevant search terms in the title, abstract, or keywords of the retrieved publications. Polarization, echo chamber, and filter bubble are the most used terms when describing the polarization phenomenon since they appear in almost every relevant publication in the field. The analysis and the research questions presented above guided the selection of the rest of the keywords.

Several off-topic terms are enumerated in the second component of the search string. These terms represent different fields of research that generated a large volume of noise in our initial searches. The AND NOT boolean operator is used for removing the uses of the term polarization that are not related to our research questions, such as electromagnetic, electric, or optical polarization. These are terms commonly used in their specific research fields. We also excluded cell, antenna, and radar network mentions. The ALL field code also removes search results when other search fields such as journal title, conference name, or publisher name contain the excluded terms.

In the third and last component of the search string, the LIMIT-TO statement is used to narrow the scope of our search to venues containing computer science, mathematics, social science, or

multidisciplinary (at least one of these) among its subject areas. Using the EXCLUDE statement, we also excluded chemical and earth science subject areas to refine the search results and further exclude noise.

The search string execution returned 405 publications on January 1, 2022. We did not use the publication date as a filtering criterion. However, all these publications appeared between 1986 and 2021.

## 2.2. Selection strategy

Despite the refinement performed by the search string, it still returned many off-topic publications not relevant to this research. It was necessary to perform a selection process among the retrieved publications. To this end, a set of inclusion and exclusion criteria was established.

The inclusion criteria used to select the publications chosen for appearance in the review among the 405 originally retrieved were as follows: (1) the publication venue must be peer reviewed; (2) the publication must meet the research questions; and (3) the publication should involve a polarization measure or model. The exclusion criteria were as follows: (1) white papers, theses, or technical reports were discarded; (2) duplicated studies, that is, work that appeared more than once with the same or similar titles and content (e.g., extended abstracts and full papers); and (3) short abstracts of conference papers, without the full content.

To ensure that the publications indeed met the research questions, we also excluded those that classify posts or texts as polarized or nonpolarized but do not make use of an underlying graph or network structure.

The selection process consisted of two filtering stages. In the first stage, we only considered the title and abstract of each publication. The 405 publications retrieved using the search string were reduced to 91 after applying this first stage of the inclusion and exclusion criteria.

In the second stage, we performed a full reading of all publications filtered by the previous stage. Each publication was reviewed by a second, different author in the second stage, who classified the publication and wrote the annotation. This second and final stage generated a set of 72 publications. The inclusion and exclusion criteria cited above were applied in both filtering stages.

We also added two additional unindexed (by Scopus) extended versions of publications that passed the filters. Lastly, we included four specific publications from 2022 indicated by the referees of this review. This process generated the 78 publications that are included in this review. Figure 2 shows the flow diagram of our methodology.

## 2.3. Categorization of publications

We created two main categories for classifying the publications according to the answers to the research questions: polarization measures and polarization reduction methods. An additional category of publications is formed by case studies and real-life applications involving the use of polarization measures and polarization reduction methods.

One or more categories were assigned to each publication. Figure 3 shows the result of the categorization of the 78 reviewed publications.

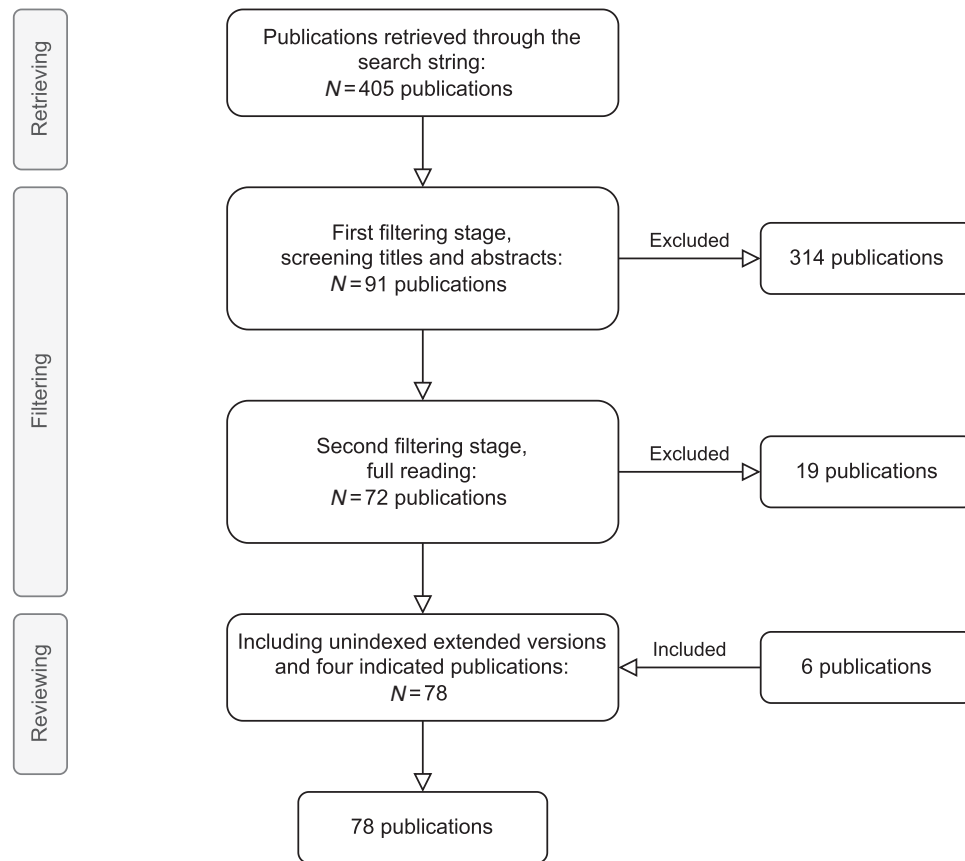


Fig. 2. Flow diagram of the process of retrieving, filtering, and reviewing the publications.

#### 2.4. Quantitative analysis

This section provides a brief quantitative analysis of the 78 publications selected at the end of the two filtering phases, which will be simply referred to as “publications” throughout the remaining of this work. According to the Scopus document classification by type, most publications are journal articles (58.1%), while conference papers represent 40.5%, and the remaining 1.4% correspond to book chapters.

Figure 4 details the number of publications by year since 2006. This number remains low and relatively stable until 2016, when it increases and reaches two peaks in 2018 and 2021, illustrating the increasing relevance and interest in the problem that occurs in parallel with the raising the influence of debates on social networks such as Twitter, Facebook, and Reddit about society’s issues.

Looking at the 30 author keywords with at least two appearances in the reviewed publications, more than a half (53.3%) of these appearances are concentrated in six terms: polarization (20.4%), social networks (9.2%), echo chambers (6.6%), social media (5.9%), filter bubbles (5.9%), and Twitter (5.3%). The co-occurrence network of these 30 author keywords is shown in Fig. 5. Node sizes indicate the number of appearances of each keyword. “Polarization” is the most common keyword,

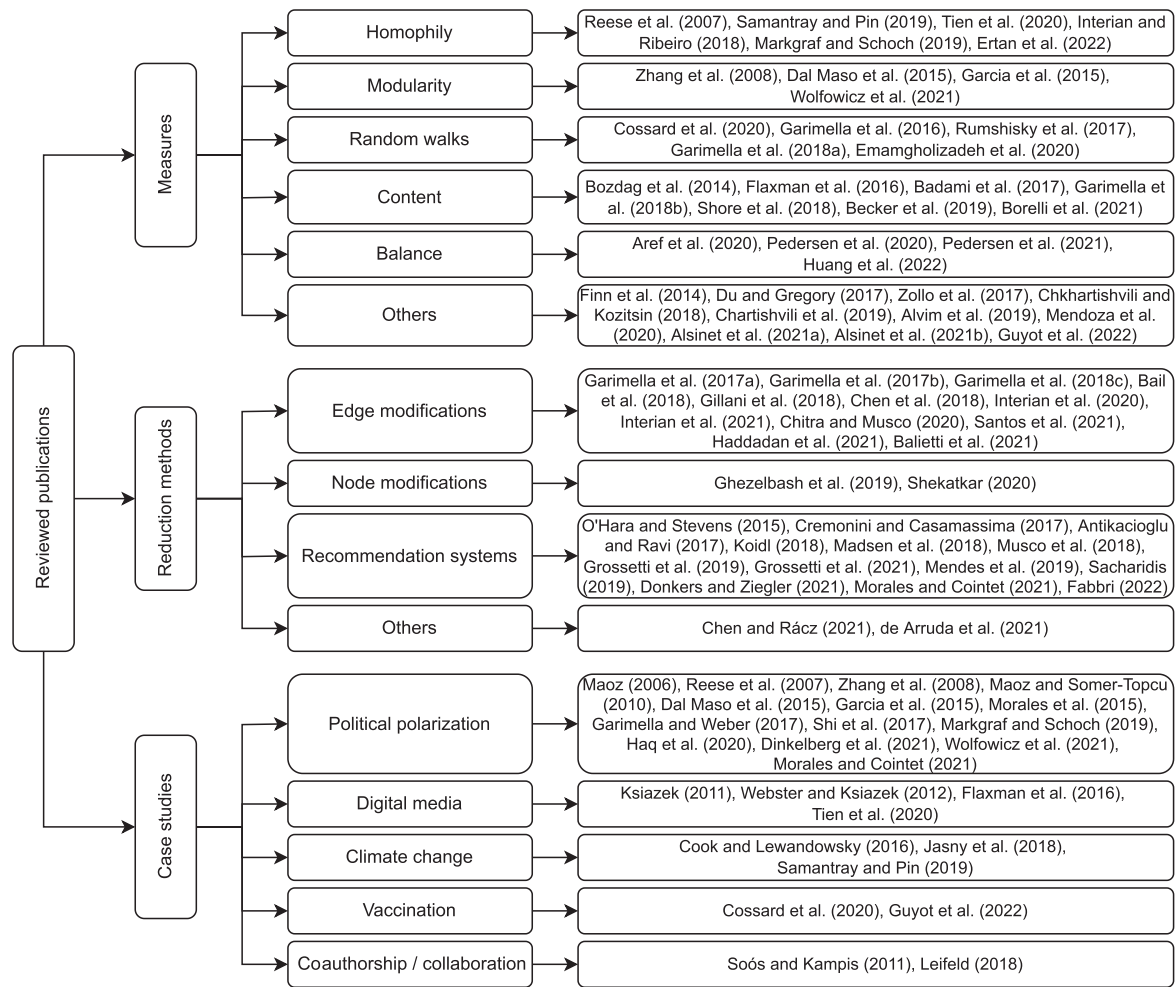


Fig. 3. Categorization of the 78 reviewed publications.

with the highest centrality in the co-occurrence network and connected to virtually all other keywords. The keywords “echo chambers” and “social networks” are also frequently found.

### 3. Polarization measures

This section presents the main approaches for measuring network polarization found in the reviewed publications.

The core of any polarization model lies in the concept of a group. Most of the time, group refers to any subset of network nodes that share some common characteristics. In some cases, groups are called communities. A strongly cohesive group that is loosely connected to other groups is commonly identified by the term echo chamber.



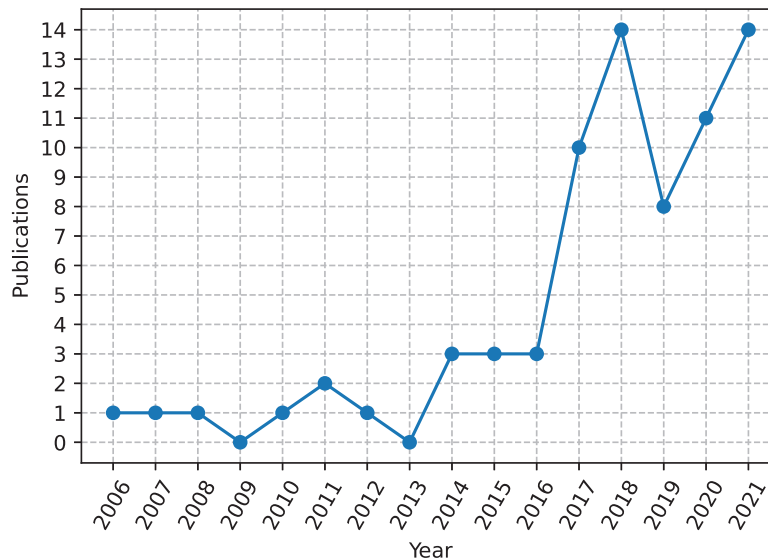


Fig. 4. Publications by year (total of 73 publications, five publications of 2022 omitted).

Let  $G = (V, E)$  be a graph or a network, where  $V = \{v_1, \dots, v_n\}$  is its node set and  $E \subseteq V \times V$  is its edge set. Let  $A = \{A_1, \dots, A_q\}$  be a set of node groups defined over  $V$ , that is, each  $A_i \subseteq V$  for any  $i = 1, \dots, q$ .

The group membership of some node  $v \in V$  can be modeled in several ways. In the typical and more frequent case, groups form a partition of the node set, that is, each node belongs to exactly one group (Garimella et al., 2016). This case also allows the representation of nodes that are isolated or do not belong to any specific group. In these situations, we denote by  $s(v)$  the only group to which node  $v$  belongs. However, other situations might exist. For example, groups could form a cover of the node set, that is, some nodes could belong to more than one group (Interian and Ribeiro, 2018). Groups could also be modeled by fuzzy sets (Zadeh, 1965), that is, with the node membership to a group  $A_i$  being given by a function  $f_{A_i} \in [0, 1]$ . The higher the value of  $f_{A_i}(v)$ , the higher the grade of membership of node  $v$  to  $A_i$  (Flaxman et al., 2016).

Perhaps the most interesting and less commonly used membership model is that of fuzzy membership. An example of fuzzy membership can be found in audience fragmentation models: each consumer can divide its attention into several, even ideologically diverse, media outlets. Other examples of models that use fuzzy membership are content qualification methods (Section 3.4), which use the content published or consumed by the users to measure their political leaning.

Understanding the polarization measures is crucial for applying the different approaches that model polarization. Understanding the measures is also important to assess the polarization reduction strategies that will be presented later in this review: any polarization reduction method must clearly establish the output that an intervention will reduce.

Several approaches are described in the literature for measuring the polarization of a network. Although some of these approaches seem to be more consolidated and have been more frequently used in case studies and applications, this is an emerging field and, naturally, different approaches

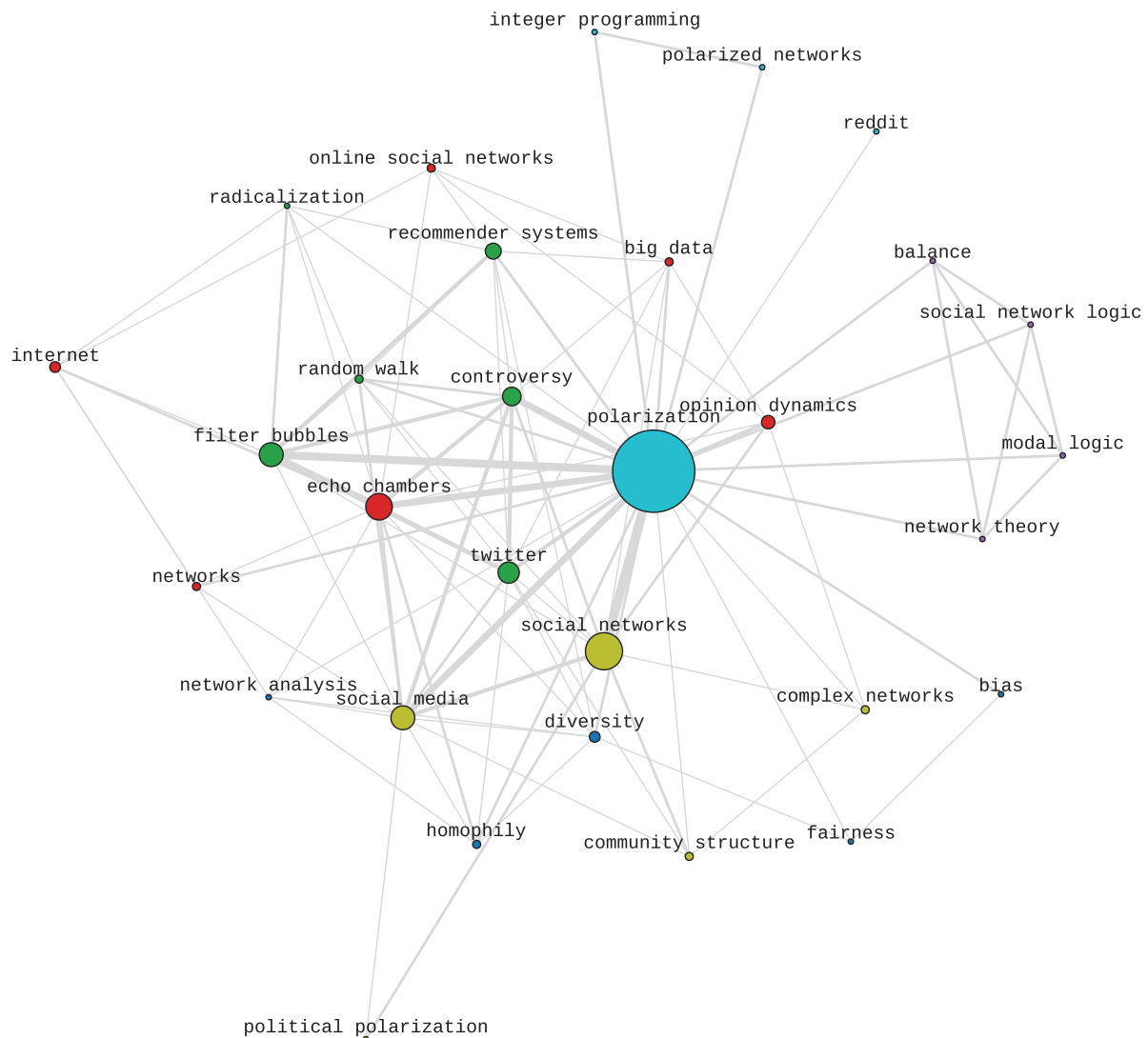


Fig. 5. Co-occurrence network of the 30 author keywords with at least two appearances (78 publications).

are considered. Therefore, Sections 3.1–3.5 present the most consolidated and used approaches in the literature, while other, more sparsely or less used approaches proposed for polarization evaluation are commented on in Section 3.6.

### 3.1. Homophily

*Homophily* (from Ancient Greek: *homo* = “self” and *philia* = “love,” love to oneself) is the tendency of individuals to associate with others that are similar to themselves. The strength of the homophily

is directly related to the strength of the network polarization. An important advantage of assessing the homophily is that it can be measured at the node, group, or network level. The term *assortativity* (Newman, 2003) is often used for defining a similar concept to that of homophily. The difference between them is that, in practice, assortativity is often used for measuring the preference for similar nodes in terms of their degrees.

The first use of the homophily for measuring the polarization of groups was made by Currarini et al. (2009) in the context of economic research. They used the term *segregation* to refer to the separation or isolation by some criterion, a very close concept to that of polarization. Later, this concept was also extended to evaluate the polarization of specific nodes by Interian and Ribeiro (2018).

Let the  $i$ -degree of a node  $v$  of a network be the number  $d_i(v)$  of neighbors of  $v$  that belong to the group  $A_i$ . The homophily (Interian and Ribeiro, 2018) of a node  $v$  with respect to the group  $A_i$  is defined as the ratio between its  $i$ -degree and the total number  $d(v)$  of neighbors of node  $v$ :

$$h_i(v) = \frac{d_i(v)}{d(v)}, \text{ for any } v \in A_i.$$

This definition only makes sense if  $d(v) > 0$ , and the homophily is only defined for nodes that fulfill this condition, that is, nonisolated nodes. The value of the homophily is a real number in the  $[0,1]$  interval, where 0 suggests *heterophily* (preference for the opposite), while 1 indicates extreme homophily.

The group-level homophily measure  $H_i$  defined by Currarini et al. (2009) denotes the average  $i$ -degree of all nodes in the group  $A_i$ , divided by their average degree:

$$H_i = \frac{\sum_{v \in A_i} d_i(v)}{\sum_{v \in A_i} d(v)}.$$

The network-level homophily  $H$  is similarly defined, considering all nodes in the network instead of those in a single group:

$$H = \frac{\sum_{v \in V} d_s(v)(v)}{\sum_{v \in V} d(v)}.$$

The homophily is a simple statistical measure that indicates the strength of the “preference for the similar,” a tendency present in many real-world networks. However, its simplicity comes with a set of drawbacks. For example, if the two groups  $A$  and  $B$  of a network divide nodes in the proportion of 90%:10%, then a group-level homophily measure of 0.5 can have a very different meaning for each group. For the small group, it seems to be very large. Contrarily, for the group representing the majority of the nodes, it appears to be insufficient to affirm that there is polarization. There are more refined polarization measures.

Reese et al. (2007) analyzed how polarized the major political blogs in the USA were, in terms of their homophily. They used a slight modification of the homophily measure, considering as neighbors nodes that are at a distance of one or two edges. Samantray and Pin (2019) studied a model of polarization of beliefs considering, in addition to the homophily, the degree of information credibility. They used a definition of group-level homophily (Currarini et al., 2009) and a polarization measure (Lelkes, 2016) that took into account the expressed opinion and the emotional content.

Tien et al. (2020) used principal component analysis (Pearson, 1901) to compute a left/right media score for each node. In this study, the assortativity measured the extent to which retweets occur between nodes with similar media preferences.

The publications below construct or use other polarization measures based on the definition of homophily.

Interian and Ribeiro (2018) analyzed the distribution of the homophily values over the nodes of a network as an indicator of the strength of polarization, using a probabilistic approach to define a new homophily-based polarization measure. It consists of the calculation, for each node, of the probability of observing a number of same-type successors that is greater than or equal to the actual number of same-type successors observed for this node. This measure was used to assess the statistical relevance of the homophily value. The authors also developed a probabilistic approach to compare the polarization of groups of nodes or entire networks, based on the computation of empirical cumulative distribution functions of the sampled data from the network. These empirical cumulative distributions provide a more insightful understanding of the status of the network. They may be used not only to compare the polarization of different groups of nodes or entire networks but also to estimate the impacts of external interventions in terms of polarization.

Markgraf and Schoch (2019) presented a framework for echo chamber research in online social networks. The first step lies in data collection. The second is community detection. The third is to assess the ideological views of the users. The fourth is to measure the degree of *echofication*, that is, to what degree a community qualifies as an echo chamber. A variant of the group-level homophily measure for multiparty networks was proposed, based on the cosine similarity between the users. The group-level homophily is calculated as the average similarity of the users to its neighbors in the network. A case study based on data from the 2017 German federal election to evaluate the framework is used. The authors argued that many researchers have reported the average homophily of the network's users but neglected separate groups that host particularly polarized users. According to them, this approach has led to underestimating the polarization of communities that host groups of political extremes.

Ertan et al. (2022) argued that there is well-established literature on measuring political polarization in two-party systems, but very limited for multiparty systems. They proposed measuring polarization for multiparty systems from survey studies. A cognitive political network (CPN) is generated for each respondent of the survey by asking them how they perceive the relationships between each possible pair among all  $n$  major political parties. From the CPN, measures of perceived party polarization were calculated. The  $E-I = \frac{ET-IT}{ET+IT}$  index (Krackhardt and Stern, 1988) proposed in the context of social psychology was used, where  $ET$  is the number of intergroup edges and  $IT$  is that of intragroup edges. It ranges between  $-1$  and  $1$ : values close to  $-1$  indicate the network is dominated by intragroup edges (homophily), and values close to  $1$  show the presence of heterophily, that is, extensiveness of intergroup edges. This measure corresponds to one minus twice the value of the homophily.

### 3.2. Modularity

Modularity optimization is a well-known method for community detection (Newman, 2006b). It treats community detection as an optimization problem by seeking an assignment of nodes to

communities that maximizes some objective function. However, the modularity has also been used for measuring the polarization inside the groups.

The *modularity* evaluates the number of intracommunity against intercommunity edges for a given set of node groups in a network. In short, the modularity is, up to a multiplicative constant, the number of intragroup edges in the network minus the expected number of intragroup edges in a network with the same nodes, communities, and node degrees, but with edges placed at random. The mathematical definition of the modularity was originally proposed by Newman (2006b):

$$Q = \frac{1}{4|E|} \sum_{u,v \in V: u \neq v} \left( a_{uv} - \frac{d(u)d(v)}{2|E|} \right) \cdot g(u, v),$$

where  $|E|$  is the number of edges in the network;  $d(v)$  is the degree of node  $v \in V$ ;  $a_{uv} = 1$  if there is an edge between nodes  $u$  and  $v$ , 0 otherwise; and  $g(u, v) = 1$  if  $u$  and  $v$  belong to the same group, 0 otherwise.

Some publications used the modularity in case studies that investigated political polarization. Zhang et al. (2008) used the modularity to quantify the increase of polarization in the U.S. Congress in the period 1979–2004. They identified communities of congressmen by employing a slight modification of the leading-eigenvector community detection method (Newman, 2006a). Dal Maso et al. (2015) used the modularity to evaluate the polarization between government and opposition in the Italian parliament. It was defined as the average modularity decrease after the group swap of two opposite-group nodes, calculated over all pairs of opposite-group nodes. The larger the decrease in the modularity, the larger is the polarization between the two groups. Garcia et al. (2015) presented an empirical analysis of politnetz.ch, a Swiss online platform focused on political activity. The approach focused on the construction of a multiplex network with politicians as nodes and three layers of directed links: one with support links, a second one with link weights as the number of comments a politician made to another politician, and the third one with weights counting the number of times a politician liked the posts of another. The polarization was studied in the three layers and measured as the modularity with respect to party labels. Wolfowicz et al. (2021) used the modularity as one of the polarization measures in a study about the interactive effects of filter bubbles and echo chambers on radicalization.

### 3.3. Random walk controversy

The *random walk controversy* (RWC) (Garimella et al., 2016, 2018b) is defined as follows. Given network  $G = (V, E)$ , let  $X, Y \subseteq V$ , with  $V = X \cup Y$ , define a partition of its node set into two subsets. Consider two random walks, one ending in  $X$  and the other in  $Y$ . The RWC measures the difference between two probabilities:

- $Pr[A]$ : probability that both random walks started from the same partition where they ended.
- $Pr[B]$ : probability that both random walks started from different partitions than they ended in.

Then, the

$$RWC = Pr[A] - Pr[B]$$

is close to one when the probability of crossing sides is low, that is, when the graph is polarized. On the other hand, it is close to zero when the probability of crossing sides is comparable to that of staying on the same side, meaning that there is no polarization.

There is a variant of the RWC measure (Garimella et al., 2018b) that can be computed more efficiently. It only considers random walks that end when reaching any of the  $k$  highest-degree nodes from either partition. In this variant, the high degree is used as a proxy for authoritativeness. The sets of the  $k$  highest-degree nodes of each group are denoted by  $X^+$  and  $Y^+$ .

The node-level RWC of a node  $v \in V$  with respect to the group  $X$  considers how often a random walk originating at  $v$  ends in nodes from  $X^+$  and  $Y^+$ . Formally, it is defined as the probability that a random walk started at node  $v$  given that it ended in  $X^+$ , divided by the sum of the probabilities that a random walk started at node  $v$  given that it ended in  $X^+$  or  $Y^+$ . The measure

$$\text{RWC}(v, X) = \frac{\Pr[\text{start} = v \mid \text{end} = X^+]}{\Pr[\text{start} = v \mid \text{end} = X^+] + \Pr[\text{start} = v \mid \text{end} = Y^+]}$$

is close to one if node  $v$  is located near the highest degree node  $X^+$  in the network, while it is close to zero when  $v$  is far from  $X^+$  and close to  $Y^+$ .

Cossard et al. (2020) analyzed the Italian vaccination debate on Twitter, using the RWC measure for quantifying the polarization.

Garimella et al. (2016, 2018b) presented and compared three new and two existing polarization measures. They argued that the random-walk-based measure outperforms other measures in capturing the intuitive notion of controversy, which is the concept used in this work to refer to polarization. To build graphs from raw data, the authors used “retweet” and “follow” relations between Twitter users. A three-stage pipeline that leads to quantifying controversy in any network is also proposed: build a conversation graph among the users who contribute to a topic, where an edge represents two users in agreement; partition the conversation graph to identify potential sides of the controversy; and measure the amount of controversy from characteristics of the graph. It is claimed that the RWC is able to separate controversial from noncontroversial topics and that this score can be used to generate recommendations that foster a healthier “news diet” on social media.

Rumshisky et al. (2017) looked at the RWC measure, text-based sentiment analysis, and the corresponding shift in word meaning and utilization by the opposing sides, using the 2014 Ukraine–Russian Maidan crisis as a case study. They analyzed the interplay of the division of network-based vs. language-based measures of conflict, using the RWC as a network measure and the standard deviation of sentiment and semantic drift as verbal measures. They observed that, as the conflict intensifies, the RWC and the standard deviation of the overall sentiment expressed by the opposing groups are positively correlated and increase in unison.

Emamgholizadeh et al. (2020) sought to determine to what extent an idea produced by some users is exposed to members with an opposite point of view. They introduced the biased random walk (BRW), a method for combining two sources of information: content or textual data and structural data. Content and structural network data are used by the biased random walker, which has some amount of initial energy (calculated considering the position of the node from where the walk starts) that is loosed in each step along the walk. The performance of BRW is compared with that of a pure random walk (Garimella et al., 2016, 2018b). They argued that, in some cases, using only structural data, it is not possible to evaluate the controversy level of the social network.



### 3.4. Content qualification

Content qualification methods use the content published or consumed by the users to measure their polarity. They do not consist of a single method: instead, each publication defines in its own way how the features of the published or consumed content will be transformed into the polarity of the nodes of the network. However, all of them employ user content (hashtags, web links, sentiment analysis) for identifying the group membership or the polarity of network nodes.

Bozdag et al. (2014) analyzed pluralism on Twitter, measuring information diversity in Netherlands and Turkey. They coined the term *cyberbalkanization* to refer to the Internet segregation into small groups with similar interests (i.e., polarization). Several metrics were used, among which we highlight three: source diversity, output diversity, and input–output correlation. Entropy is used to assess the diversity of information consumed or produced by each user. Source diversity is measured by the entropy of the tweets published by followed users from different groups. Output diversity is estimated from the entropy of the retweets and replies the user makes. The input–output correlation indicates whether the political position of the most common message category retweeted by a user is significantly skewed from the political position of the received messages. The results indicated a high source diversity, similar to Turkish and Dutch users. The output diversity is much lower than the source diversity. Considering the minority access, the content produced by minorities cannot reach a large fraction of the Turkish population.

Flaxman et al. (2016) examined the web-browsing patterns of 50,000 U.S.-located Internet users who regularly read online news, defining four channels individuals use to discover a news story: direct, aggregator (Google News), social (Facebook, Twitter), and search (web search queries on Google, Bing, Yahoo). They estimated the polarity of a news outlet by measuring how its popularity varies across counties as a function of their political compositions. A Bayesian model was used for estimating the polarity of each article and user. The polarity score of each article is inferred from the polarity of its publisher. Using the polarity scores of the articles read by some users, the model also estimates their polarity. The segregation (i.e., the ideological distance) between two individuals is defined as the expected value of the squared distance between their polarity scores. It was shown that articles found via social media or web search engines have higher ideological segregation than those users read by directly visiting news sites. However, it was also found that social media and web search engines are associated with greater exposure to opposing perspectives.

Badami et al. (2017) observed the importance of understanding how recommendation systems behave in polarized environments, studying polarization in the context of the users' interactions with a space of items. Their model works with ratings that capture the distribution of user opinions. In the absence of polarization, the distribution of opinions should be either J-shaped or bell shaped. As polarization emerges, the distribution becomes U-shaped, with two peaks emerging around the two confronted opinions at the extreme sides of the rating scale. The authors developed an approach to quantify polarization based on four stages: (1) building items' rating histograms from user-item rating data, (2) extracting a set of features from the histograms, (3) training a polarization classifier based on a sample of annotated cases, and (4) measuring the item-level polarization score. They performed comparisons of polarization measures on several benchmark datasets and showed that their framework can detect different degrees of polarization.

Garimella et al. (2018a) assessed the degree to which echo chambers are present in political discourse on Twitter, and how they are structured in terms of different user roles. Two node-level

measures related to how polarized is the content that each user consumes and produces are defined. The production polarity of a user is the average polarity of its tweets. The consumption polarity of an user is the average polarity of the tweets it receives from the followed users. A user is classified as partisan if it produces one-sided content, and bipartisan if it produces two-sided content. The authors also looked at gatekeeper users, who consume content of both leanings but produce single-sided content. The findings indicated a large correlation between the leaning of produced and consumed content. Partisan users enjoy a higher “appreciation” as measured by network and content features, indicating a “price of bipartisanship,” paid by users who try to bridge echo chambers. They pay the price in terms of network centrality and endorsements from other users, highlighting the existence of a latent phenomenon that effectively stifles mediation between the two sides.

Shore et al. (2018) sought evidence of echo chambers strengthening by analyzing the diversity of hyperlinks posted on Twitter. They used ordinary least squares models to test their hypotheses, combined with standard community detection algorithms to identify the groups. They found that the average account posts link to more politically moderate news sources than the ones they receive in their own feed. However, members of a tiny network core do exhibit cross-sectional evidence of polarization and are responsible for the majority of tweets received overall due to their popularity and activity, which could explain the widespread perception of polarization on social media. Evidence was also found that people in highly clustered positions (echo chambers) tweet more similarly to the people they follow.

Becker et al. (2019) analyzed partisan networks of Republicans and Democrats to test whether the wisdom of crowds is robust to partisan bias. They studied belief formation on controversial topics. Two web-based experiments were conducted, where each individual answered questions to elicit partisan bias before and after observing the estimates of peers (social information) in a politically homogeneous social network. Polarization was measured using two outcomes. First, the average distance (absolute value of the arithmetic difference) between the mean normalized belief for Republicans and Democrats at each experiment round. Second, the average distance between every possible two-person cross-party pairing, which reflects the expected distance between the beliefs of randomly selected Democrats and Republicans. The experimental results indicated that social information in politically homogeneous networks do not always amplify existing biases. Instead, in the studied networks, the information exchange increased belief accuracy and reduced polarization.

Borrelli et al. (2022) examined the relationship between online emotional reactions, affective polarization, and counternarratives, following an approach based on user-generated textual data. Affective polarization is the extent to which two opposing groups dislike one another. This occurs in online social networks as a result of a controversial event in the offline world. A content-based measure of affective polarization derived from user-generated content was proposed to evaluate the usage of a set of controversial words, reflecting how important a word is to a document in a collection. Also, a new method is proposed to assess the effectiveness of online counternarratives made by influential actors to counteract the rise of affective polarization. It was applied to five cases of controversial events that occurred in European soccer leagues using Twitter data, showing that there was a high polarization in online responses in most cases. Counteractive official communication from the clubs within 12 hours of the event often reduced the affective polarization.



### 3.5. Signed networks and balance theory

The notion of balance comes from the idea that, in a group of people, some logical rules are generally observed (e.g., people like their friends' friends, people hate their friends' enemies). If a social network always satisfies these rules, it is said to be balanced.

Cartwright and Harary (1956) and Heider (1946) studied the theory behind such relationships and attitudes. The network is modeled as a signed graph  $G(V, E^+, E^-)$ , which consists of a set  $V$  of vertices and two disjoint subsets  $E^+, E^-$  of positive and negative edges, respectively. Formally, balance is achieved whenever each triangle (or 3-cycle) has three positive edges (*my friend's friend is my friend*) or two negative and one positive edge (*my friend's enemy is my enemy*).

The structure theorem (Cartwright and Harary, 1956) shows that a signed graph is balanced if and only if its nodes can be separated into two disjoint subsets such that each positive edge joins two nodes of the same subset, while each negative edge joins nodes from different subsets. Balanced graphs may be used as a model of polarized networks. A graph with exactly two groups of nodes linked internally only by positive edges and with negative edges between the groups represents a perfectly polarized network. In the case of weak balance, the existence of triangles with three negative edges is also allowed. Weak balance (Pedersen et al., 2020, 2021) is characterized by the possibility of partitioning the nodes into any number of groups.

Harary (1959) defined some measures for evaluating how close a given graph is to balance. The degree of balance of a signed graph  $G$  is given by the ratio of the number of positive cycles to the total number of cycles, where the sign of a cycle is the product of the signs of its edges:

$$\beta(G) = \frac{c^+(G)}{c(G)}.$$

The line index of balance is the minimum number of edge modifications that must be made in order to achieve balance. The two most used modifications are edge removals and sign changes. The third measure is the point index, given by the smallest number of nodes whose deletion results in balance.

Aref et al. (2020) investigated the relationship between network structural configurations and tension in social systems by using balance theory and three levels of analysis for balance assessment: triads, groups, and the whole network, delivering empirical evidence for the argument that balance at different levels represents different network properties that should be evaluated independently. For triad-level balance, a new measure was developed by using semicycles that satisfy the condition of transitivity. For group-level balance, the measures of cohesiveness (intragroup solidarity) and divisiveness (intergroup antagonism) were proposed to capture balance within and among groups. For network-level balance measurement, the authors modified the line index of balance (Harary, 1959), introducing a normalized line index. Large values of this index represent high partial balance, and therefore a more balanced network. The investigation of different social networks showed that balance appeared differently across multiple levels of analysis. In most cases, relatively high values of balance were observed, corresponding to high triad, group, and network polarization.

Pedersen et al. (2020, 2021) proposed different ways of defining properties related to the concept of balance in signed social networks, where relations can be either positive or negative. To be able to formally reason about the social phenomenon of group polarization based on balance theory,

Table 1  
Comparison of the polarization measures in terms of their granularity

| Measure                       | Node level? | Triangle level? | Group level? | Network level? |
|-------------------------------|-------------|-----------------|--------------|----------------|
| Homophily                     | Yes         | -               | Yes          | Yes            |
| Modularity                    | -           | -               | -            | Yes            |
| Random walk controversy       | Yes         | -               | -            | Yes            |
| Content qualification methods | Yes         | -               | -            | -              |
| Balance-based measures        | -           | Yes             | Yes          | Yes            |

they used positive and negative relations logic (Xiong and Ågotnes, 2020). Positive and negative relations between nodes are interpreted as agreement or disagreement on a given issue. They studied a polarized network as a balanced graph of groups positively related within, but negatively related to the others. They differentiated strong and weak polarization. Strong polarization occurs when the network can be divided into two mutually opposed groups. Weak polarization is characterized by the division into many groups. The authors presented three measures: degree of imbalance, level of imbalance, and line index of imbalance, defined for both the strong and weak polarization cases. They discussed their strengths and weaknesses on examples of signed networks.

Huang et al. (2022) focused on predicting conflicts as negative links between users. According to the authors, negative links between polarized communities are too sparse to be predicted by state-of-the-art approaches. A polarization measure for signed graphs that incorporates social balance theory is proposed based on signed random walks. This measure guarantees polarized similarity consistency, satisfying two properties: (1) topologically close nodes are more similar than topologically distant ones and (2) positively related nodes are more similar than negatively related ones. Then, POLE (POLARized Embedding for signed networks), a signed embedding method for polarized graphs based on random-walk based measure is proposed. Through the experiments, the authors claimed that POLE outperforms state-of-the-art methods in hostile links prediction.

Table 1 shows the polarization measures found in the reviewed publications and discussed in Sections 3.1–3.5 that can be used for assessing the polarity of individual nodes, or the polarization of groups or entire networks. Most measures aim to evaluate the polarization strength at the node or network level.

### 3.6. Other approaches

There are other, less used methods for polarization measuring that do not fit in the more frequently used approaches exposed in the previous sections. There are many different terminologies, methodologies, and measures, some of them similar to others. This section discusses other measures that use alternative techniques and ideas for quantifying node, group, or network polarization.

Finn et al. (2014) introduced the co-retweeted network, which is the weighted graph that connects highly visible accounts retweeted by members of the audience during some real-time event. When applied to political conversations related to some event, the co-retweeted network enables the measurement of the political polarity of major players (including news outlets), based on the views of the audience on Twitter. Its first application is the measurement of the opinion polarization for

an issue or topic, that is, the computation of the polarity of the event itself. The second consists in measuring the media bias, as perceived by the Twitter users. The authors used their method to infer the polarity of all engaged accounts in the audience, in order to answer the question of whose supporters were more active and vocal during an event.

Du and Gregory (2017) investigated whether social media platforms increase the polarization of users, by checking if the community structure becomes stronger as time passes. Twitter networks that consider only reciprocated “follow” relations between users were used. The authors measured how often new edges appeared and whether edges tend to be removed (by “unfollowing”) inside or between communities. Two hypotheses were explored: (1) new edges are more likely to appear inside communities than between communities and (2) edges between communities are more likely to be removed than those inside them. These two hypotheses were contrasted with the null hypothesis when edges are added and deleted randomly. The authors showed that the number of intracommunity edges added in the real networks is always much greater than in the random case. They also showed that intercommunity edge deletion is more common than expected in the random case. Therefore, the polarization of the “follow” network becomes stronger. The authors argued that one possible explanation for this effect is the recommendation system of Twitter.

Zollo et al. (2017) examined the effectiveness of debunking on Facebook through a quantitative analysis of 54 million users from January 2010 to December 2014. Debunking posts strive to contrast misinformation spreading by providing fact-checked information on specific topics. The authors compared how Americans who consume proven (scientific) and unsubstantiated (conspiracy-like) information on Facebook interact with debunking posts. Their findings confirmed the existence of echo chambers where users interact primarily with either conspiracy-like or scientific pages. The user polarity is defined as the ratio of the difference in likes (or comments) on conspiracy and scientific posts. The probability density function of the polarity of all users is sharply bimodal, and most users may be divided into these two groups. The majority of likes and comments are made by users polarized towards science, while only a small minority is made by users polarized towards conspiracy. Only few users active in the conspiracy echo chamber interact with debunking information.

Chkhartishvili and Kozitsin (2018) proposed the binary separation index to quantify the echo chamber effect in social networks for a specific topic. It requires the ideological space to be binary. It does not require the information of all users’ opinions on the topic but only of a subset of accounts that disseminate information in the network and their political positions. For a given social network and a fixed topic, it generates a number between 0 and 1: the higher it is, the greater is the level of information separation between the groups. However, the authors have not considered all possible information spreaders, not examining group pages. They discussed the calculation of this index for the prevalent Russian social network VKontakte. Considering the attitude towards the Russian government as a topic, they obtained an index of 0.802 after data processing, which was evidence of a high level of information separation among VKontakte users.

Chartishvili et al. (2019) proposed an extension of the Esteban–Ray measure (Esteban and Ray, 1994) originally proposed for measuring economic characteristics of a population. It may be applied when opinions are evaluated by continuous scalar values representing personal attitudes towards a fixed topic. The proposed extension evaluates the level of polarization of the individuals’ opinions in a social network. An individual’s opinion is described by a scalar in the interval  $[0,1]$ , representing the degree to which the individual holds a particular position. The proposed polarization index is

proportional to the difference between the average opinions inside the groups and belongs to the interval  $[0,1]$ . It is sensitive to the cluster sizes and reaches its maximum value when the groups are equal in size. The authors showed how the measure works for real data, applying it to a time series of user opinions in the VKontakte social network that is devoted to a political topic, reporting an increase in the level of polarization.

Alvim et al. (2019) also developed an Esteban–Ray-based polarization measure (Esteban and Ray, 1994) and a social network model. The model includes information about each agent's quantitative strength of belief in a proposition and a representation of the strength of each agent's influence on every other agent. The authors considered how the model changes over time as agents interact and communicate. They included several different options for belief update, such as rational belief update and update taking into account irrational responses such as confirmation bias (groups may strengthen their beliefs by interpreting information in their favor) and the backfire effect (groups may strengthen their beliefs by strongly opposing an opinion if it contradicts their views). The authors considered the evolution of polarization over time under various scenarios, as well as the implications of these results for real-world social networks. Simulations were shown exploring how interaction graphs and cognitive biases may lead to polarization. Their experiments allowed us to identify that sometimes people with different opinions interacting more strongly may lead to more polarization.

Mendoza et al. (2020), introduced GENE (Graph Generation conditioned on Named Entities), a representation of user networks conditioned on the entities (personalities, brands, organizations) that users comment upon. The goal was the early detection of polarization and controversy in news events. GENE segments the user network, and the segmented network is used to study two controversy indices, the RWC (Garimella et al., 2016, 2018b) and the relative closeness controversy (RCC) proposed here. To evaluate the performance of GENE, the network of users of the online news site Emol (EMOL, 2022) in Chile was modeled. The results showed that over 60% of the user comments have a predictable polarity, allowing both controversy indices to detect the controversy successfully. The authors argued that the RCC index shows satisfactory performance in the early detection of controversies using the information collected during the first hours of the news event. A polarization dynamic can be anticipated, predicting the emergence of controversies before they occur.

Alsinet et al. (2021a, 2021b) introduced a quantitative model for measuring polarization in online discussions. They modeled the debates in Reddit using weighted graphs with labeled edges, where node weights represent the polarity of the users' opinions in the debate, and edge labels represent the sentiments between users' opinions. The proposed measure is based on the maximum polarization of a debate, considering all possible graph bipartitions. For each bipartition, the polarization is quantified by measuring the uniformity of the users' opinions within each partition and the negativity of the interactions between the partitions. The maximum polarization is computed by a greedy local search algorithm. The authors argued that their approach can be used for monitoring a discussion and generating a warning signal when the polarization of the debate reaches some threshold value. They performed empirical evaluations of different Reddit discussions. The quantitative model captured differences in the polarization of different discussions. Additionally, a graph neural network (Hamilton et al., 2017) was used to approximately compute the polarization measure of a Reddit debate.

Guyot et al. (2022) stated that users in the boundaries significantly contribute to network polarization, acting like gatekeepers of their communities. They used an approach that relies

on community boundaries to compute two measures: community antagonism and the porosity of boundaries. These measures assess the degree of opposition between communities and their aversion to external exposure, respectively. The authors evaluated their proposal using a case study obtained from Twitter and related to COVID-19 vaccination.

#### 4. Polarization reduction

The second research question targeted network polarization reduction methods suggested in the literature, which are now exposed in this section. All of them have in common some attempt of changing different features of the network: add or remove edges (Garimella et al., 2017b; Interian et al., 2021); introduce specific types of nodes (Shekatkar, 2019; informed agents, Ghezlbash et al., 2019) or the spread of random information (Cremonini and Casamassima, 2017). The publications may propose methods to compute optimal interventions or analyze the impact of such modifications in the network structure, evaluating their effect on the polarization of the entire network or on the polarity of specific nodes.

Surprisingly, no studies about removing (or adding) nodes for reducing the polarization were found in this review. However, this method is often used in practice for banning specific posts or accounts from social networks (The Conversation, 2020).

We observe that polarization reduction strategies seem to be more effective on users (nodes) that are new to the network (Madsen et al., 2018) or when a polarized discussion first emerges (Donkers and Ziegler, 2021), that is, when the polarization process is in formation, but not when it is already consolidated and the polarized groups are well established.

##### 4.1. Edge modifications

This approach, proposed in many publications, suggests that adding, removing, or changing the weight (or the strength) of some specific edges can reduce the polarization of the network. These interventions represent externally induced processes that promote edge appearances or deletions, such as marketing or fact-checking campaigns, regulatory actions, or direct manipulations that add or remove edges of the network.

Most studies suggest that adding edges that link different groups may decrease group polarization. These edges usually represent the exposure of individuals to different or opposing views. However, Bail et al. (2018) stated that this exposure to opposing views on social media may increase the polarization in some cases.

These are not necessarily conflicting hypotheses, since the exposure to opposing views may decrease polarization in the initial or intermediate phases of the process of polarization, but not when the polarization is already strong. The effectiveness of this approach may depend on several factors. Among other researchers that tested edge additions in real-world networks, Cossard et al. (2020) argued that exposure to contrarian content has been shown to be both effective (Horne et al., 2015; Garimella et al., 2017b) and counterproductive (Nyhan and Reifler, 2010; Bail et al., 2018) in reducing the polarization, depending on the specific network setting or the existing degree of polarization.

Garimella et al. (2017a) elaborated on a demo providing automated tools to help users explore and escape their echo chambers. A discussion topic is identified as the set of tweets that contain a specific hashtag. The topic is represented by an endorsement graph, where nodes represent users and edges represent endorsements. The two sides of a controversial topic are identified by employing a graph partitioning algorithm, dividing the graph into two subgraphs. Polarity scores are obtained for all users. The demo provides contrarian content recommendations, that is, content that expresses views from the opposing side of the controversy. However, not all recommendations are acceptable, especially if they do not conform to the users' beliefs. To reduce these effects, an acceptance probability is defined, quantifying the degree to which a user is likely to endorse the recommended content. The maximum reduction of the user-polarity score is achieved by putting the user in contact with an authoritative source from the opposing side. The authors claim that the contribution of their demo is twofold. First, as a tool to visualize retweet networks about controversial issues on Twitter. Second, as a solution proposal to reduce the polarization by exposing users to contrarian views.

Garimella et al. (2017b, 2018c) studied algorithms for bridging the echo chambers created on social media, and thus reducing controversy (i.e., polarization). They represented the discussion on a controversial issue with an endorsement graph, raising an edge-recommendation problem on this graph. The goal of the recommendation is to reduce the RWC score of the graph (Garimella et al., 2016, 2018b). The authors also took into account the acceptance probability of the recommended edge. The goal is to find edges that produce the largest expected reduction in the controversy score. They proposed an algorithm that considers only edges between high-degree nodes of each side of the controversy. For each edge, it computes the reduction in the RWC score obtained when that edge is added to the original graph, then selects the  $k$  edges that lead to the lowest scores when added to the graph individually. Experimental results showed that the algorithm is more time efficient than a simple greedy heuristic while producing comparable RWC score reduction.

Bail et al. (2018) surveyed a large sample of Democrats and Republicans who visit Twitter at least three times a week about a range of social policy issues. One week later, they randomly assigned respondents to a treatment condition. They were offered financial incentives to follow a Twitter bot for one month. This bot exposed them to messages from those with opposing political ideologies. Respondents were resurveyed at the end of the month to measure the effect of this treatment and at regular intervals throughout the study period to monitor treatment compliance. The authors found that Republicans who followed a liberal Twitter bot became substantially more conservative after treatment. Democrats exhibited small increases in liberal attitudes after following a conservative Twitter bot, although without statistical significance. The authors found no evidence that exposing Twitter users to opposing views reduces the political polarization. The study indicated that attempts to introduce people to a broad range of opposing political views on a social network such as Twitter might be not only ineffective but counterproductive.

Gillani et al. (2018) sought to mitigate political echo chambers by showing the participants a subset of their social networks and asking them to discover their level of social connectivity. The authors created a web application. Each participant answered questions regarding their engagement in political discourse on Twitter. Next, the application presented a visualization of the participant's network. The application then asked the participant to give their location in the network. After a guess was made, the tool revealed the true position of the participant. Sometimes the participant might also see a list of suggested accounts to follow that would increase their diversity score. Finally,

the participant was asked to complete a postsurvey. For each participant, the authors measured the difference in their answers to the survey questions and the political diversity of the accounts this participant followed on Twitter before and after treatment. Participants asked to find their accounts in their social network, with nodes colored by inferred political ideology, tended to increase their belief in how ideologically closed they really were, but the political diversity of who they chose to follow actually decreased several weeks after treatment. The diversity of the followers of participants recommended to follow Twitter accounts with opposing political views increased one week after treatment.

Chen et al. (2018) relied purely on the topology of Friedkin–Johnsen's model (Friedkin and Johnsen, 1990) of opinion formation to quantify the risk of conflict in a social network. A probabilistic model was proposed, assuming that the internal opinions of the  $n$  nodes of the network follow a uniform distribution over  $\{-1, 1\}^n$ . The average-case conflict risk is defined as the expected conflict with regard to the internal opinions. An alternative and more robust measure, the worst-case conflict risk, is defined as the maximum conflict over all possible internal opinion vectors. They showed how both risk measures can be minimized by locally editing the network for a number of preexisting measures of conflict and disagreement. Two algorithms were proposed to locally edit the network to reduce the worst-case and average-case risks for a number of measures of conflict. The authors focused on identifying a limited number of edges to add or remove in the network to reduce the risk of conflict. They showed the usefulness of these characterizations of conflict risk in a range of networks and claimed that their optimization minimized the actual risk on some random opinion assignments.

Interian et al. (2020, 2021) proposed the minimum intervention principle, which assumes that the smallest number of changes should be made in the original network by any polarization reduction method. The issue of the insufficient communication between the polarized groups is solved by edge additions. The minimum cardinality edge addition problem is proposed as a strategy for reducing the polarization in real-world networks. The problem was shown to be NP-hard. Preliminary results obtained by an iterated greedy heuristic were presented in Interian and Ribeiro (2019), while three integer programming formulations were compared in Interian et al. (2021) with computational results for both randomly generated and real-life instances. It was shown that the polarization could be reduced to the desired threshold by the addition of few edges, as established by the minimum intervention principle that guided the problem formulation. According to the authors, there is often a straightforward way of spreading polarization-breaking information, even in strongly polarized networks.

Chitra and Musco (2020) augmented the Friedkin–Johnsen opinion dynamics model (Friedkin and Johnsen, 1990) to include the filter bubble, the practice of connecting users with ideas they are already likely to agree with. A network administrator is introduced in the model as an external actor that dynamically adjusts the strength of specific edges of a social network. The network administrator seeks to minimize a standard measure of disagreement between interacting users in the social network since the authors considered that user engagement would increase by reducing users' disagreement. Individuals update their opinions according to the model's dynamics, and the administrator repeatedly adjusts the underlying network graph to achieve its own goal. The study showed that in Reddit and Twitter networks, after introducing the network administrator dynamics, even small changes to the edge weights may significantly increase the polarization. Finally, a simple modification in the network administrator's incentives that limit the filter bubble effect was

proposed for countering the increasing polarization. According to the authors, their solution increased user disagreement from 3% to only 5%, showing that this modification would minimally affect user engagement.

Santos et al. (2021) investigated the relationship between social networks and reputation-based cooperation in large populations, analyzing the impact of network topology on polarization. They proposed a game-theoretical evolutionary model and studied dynamics in networks with varying degrees of community structure. They showed that networks exhibiting modular structures hamper global cooperation. Strategies of cooperating exclusively with in-group members fixate, sustaining polarization and group bias. The model uses the stern-judging social norm: helping a good individual or refusing help to a bad one leads to a good reputation, whereas refusing help to a good individual or helping a bad one leads to a bad reputation. When communities are well-defined and reputations are attributed following stern-judging, polarization, and group bias emerge: cooperation thrives within communities, though not across communities. Global cooperation is recovered as long as intercommunity edges are added.

Haddadan et al. (2021) argued that structural bias may trap a user of the World Wide Web in a polarized bubble with no access to diversified opinions. They modeled user behavior by random walks and defined the polarized bubble radius (BR) of a node as a measure to quantify its polarity. It denotes the expected number of steps to go from this node to a page of a different opinion. A node is in a polarized bubble if the expected length of the random walk to a page of different opinions is large. The structural bias of the whole graph is the sum of the radii of its polarized bubbles. The authors studied the problem of decreasing the structural bias using edge insertions. As “healing” all nodes with a high polarized bubble radius is hard to approximate, they presented an algorithm for finding the best set with a fixed number of edges whose insertion maximally reduces the graph’s structural bias. The algorithm is able to return a constant-factor approximation using a greedy approach based on a specific variant of the random-walk closeness centrality (White and Smyth, 2003).

Baliatti et al. (2021) used informal political discussions with individuals sharing personal characteristics and social context to decrease polarization by exposing them to personal messages about a divisive political topic. According to the authors, friendship networks exhibit greater diversity of political views than is apparent to their members, and incidental conversations may expose interlocutors to diverse viewpoints. A large-scale experiment is performed by matching participants to peers having common interests and demographics and exposing them to a personal message about wealth redistribution. As a result, informal communication increases support for redistribution and reduces opinion polarization, suggesting that incidental conversations have the effect of increasing consensus on divisive and partisan topics. The authors showed that “feeling close to the match” is associated with an increase in the probability of assimilation of diverse political views.

#### 4.2. Node modifications

The approaches presented in this section are based on the hypothesis that introducing specific types of nodes (informed agents, Ghezelbash et al., 2019; zealots, Shekatkar, 2019) or changing the features of some nodes may decrease the network polarization.



Ghezelbash et al. (2019) investigated how a group of selected informed agents can lead society towards some desired opinion. Informed agents are individuals aware of the desired opinion and act as hidden advisers, leading the society to this opinion through interactions with other individuals. An optimization technique was developed to solve the informed agent selection problem. The opinion dynamics model uses a network modeled by a connected graph. The opinion of each agent on a certain issue can be represented by a real value in the interval  $[-1, 1]$ . Opinions at the two ends of the interval are extreme opinions. If the members of each group reach a consensus on some specific opinion, there is complete polarization. Fragmentation occurs when the divergence of opinions leads to more than two groups. Several goals such as polarization, fragmentation, and diversity of opinions were considered and formulated, showing that they are NP-hard optimization problems. For a specific sample graph, they were solved to optimality. Although the agents with more connections have more influence on the network, the authors showed that they are not necessarily the best candidates for being informed agents.

Shekatkar (2019) considered zealots, or “inflexible minorities,” as nodes in a social network that do not change their opinions under social pressure, investigating their effect on the polarization dynamics. The author proposed a quantifier called “correlated polarization” to measure the amount of network polarization. The correlated polarization is close to one if two extreme groups of comparable sizes and with opposite opinions are formed. Two types of zealots are studied: uniform and topology-based (when only high-degree vertices are zealots). The polarization dynamics are simulated by using a simple majority rule (Galam, 2002) model. Considering an undirected network, every model’s node could be in three different states:  $+1$ ,  $-1$ , or  $0$ , where  $+1$  and  $-1$  represent the opposite opinions and  $0$  corresponds to the neutral point of view. Each node in the network can be a zealot with some probability. If a zealot has an already definite opinion, it will never change its state. The results indicated that the presence of zealots in a social network does not have a fixed effect and can lead to either positive or negative changes in the polarization, depending on the initial conditions and other factors, such as the edge density.

#### 4.3. Network design or recommendation modifications

The publications in this section propose strategies for reducing polarization based on changing the very design of the social network.

A number of studies stand that social networks are in themselves causally sufficient to promote echo chambers due to their structure (Musco et al., 2018), size (Madsen et al., 2018), lack of trust (Koidl, 2018), or the embedded recommendation systems (Grossetti et al., 2020). Therefore, they argue that promoting features such as transparency (Mendes et al., 2019), trust (Koidl, 2018), or improved recommendation systems (Grossetti et al., 2020) can reduce polarization.

In particular, some studies indicate that content personalization produced by recommendation systems may increase the echo chamber effect and create filter bubbles (Grossetti et al., 2020). Therefore, some authors claim that modifying recommendation systems may reduce network polarization (Antikacioglu and Ravi, 2017; Grossetti et al., 2020; Morales and Cointet, 2021).

O’Hara and Stevens (2015) considered whether filtering and recommendation technology on the Internet could amplify groups’ viewpoints, leading to polarization of opinion across communities and increases in extremism. The echo chamber arguments of Sunstein (2007) were taken as

representatives of this point of view and examined in detail in the context of a range of research in political science and the sociology of religion. The question was not whether there are echo chambers on the Internet; for there is plenty of evidence that there are. The two key questions were therefore whether the Internet is complicit in the growth of echo chambers, and if so, whether it should be the target of a policy focus. The authors claimed that the support for the echo chamber thesis is not strong enough to justify regulation or intervention.

Cremonini and Casamassima (2017) studied control strategies for social networks based on the introduction of random information into some selected driver agents. The goal was to better distribute knowledge among the agents by reducing polarization and augmenting their average skill level. The authors defined two information diffusion metrics. The average knowledge is the average skill with respect to the topics actually known by the agents. The knowledge diffusion, instead, is the average skill of agents with respect to the total number of topics in the network. The network tends to polarize when the difference between the two metrics increases because the agents are gaining skills mostly on the same subset of topics without a corresponding increase in erudition. The authors studied how the control strategies affected the diffusion of topics and skills in the agent population. The two strategies they studied were based, first, on the selection of a few influencers and the manipulation of their behavior, and, second, on the selection of many ordinary users as the drivers of the network. They concluded that the strategic use of random information could represent a realistic approach to network controllability and that both strategies could achieve this control effect.

Antikacioglu and Ravi (2017) affirmed that collaborative filtering is a powerful framework for building recommendation systems. The propensity of such systems to favor popular products and create echo chambers has been observed. The authors addressed the problem of increasing diversity in recommendation systems based on collaborative filtering that use past ratings to predict a rating quality for potential recommendations. They formulated a recommendation system design as a subgraph selection problem from a candidate supergraph of potential recommendations where both diversity and rating quality are explicitly optimized. On the modeling side, they defined a new flexible notion of diversity that allows a system designer to prescribe the number of recommendations each item should receive and smoothly penalizes deviations from this distribution. On the algorithmic side, they showed that minimum-cost network flow methods yield fast algorithms for designing recommendation subgraphs that optimize this notion of diversity. They claimed the effectiveness of their model and method to increase diversity while maintaining high rating quality with empirical results in standard rating data sets from Netflix and MovieLens.

Koidl (2018) proposed a novel design of social media applications, whose main motivation is the creation of a social network architecture that follows a “trust by design” paradigm. The author argued that the concept of trust is the driving force behind all three most important challenges of existing social networks, that is, the existence of a filter bubble, the spreading of fake news, and the growing of echo chambers. For each user, a decentralized network is created. Due to the dynamics within the user environment, the author suggested a cloud-based storage solution. Each decentralized network follows a peer-to-peer architecture, in order to ensure a high level of privacy and control. The application’s main feature is to establish computational trust by enabling a numerical trust value towards each new element within a social network, empowering the user to assess what elements to trust. To compute trust, different approaches were proposed. The result is a trust graph that informs the social graph about the trustworthiness of each element in the network. The authors

claimed that this trust-based platform limits the ability to create fake or distorted representations of the individual and strongly relies on authenticity, excluding fake or unauthentic representations.

Madsen et al. (2018) employed an agent-based model that allows for relevant cognitive functions (Bayesian belief revision) and agent interaction (sharing their beliefs) to explore the emergent echo chambers. They showed that echo chambers can emerge among error-free Bayesian agents, and that larger networks encourage rather than ameliorate the growth of echo chambers. The authors tested interventions to reduce the formation of echo chambers, finding that system-wide truthful “educational” broadcasts ameliorate the effect but do not remove it entirely. Such interventions are shown to be more effective on agents newer to the network. The authors claimed that social networks are in themselves causally sufficient to promote echo chambers. This carries a critical implication for interventions aimed at reducing them: individual-based interventions may help reduce somewhat the harmful, or erroneous thinking that promotes the formation of echo chambers but are not sufficient to remove them. Instead, the authors argued that system-based interventions might be more effective, taking advantage of top-down system alterations for reducing echo chambers.

Musco et al. (2018) explored the trade-off between disagreement and polarization in online social networks. Their research question was given  $n$  agents, each with its own initial opinion on a topic, and an opinion dynamics model, what is the structure of a social network that minimizes disagreement and controversy simultaneously? This question is central to recommendation systems: Should a recommendation system prefer a link suggestion between two users with similar mindsets (to keep disagreement low) or between two users with different opinions to expose each other to a contrarian viewpoint (decreasing overall levels of polarization)? The authors provided a mathematical formulation for finding an optimal topology that minimizes the sum of polarization and disagreement under the Friedkin–Johnsen opinion formation model (Friedkin and Johnsen, 1990), which takes into account both consensus and disagreement in the opinion update process. They proved that there always exists a graph with  $O(n/\epsilon^2)$  edges that provides a  $(1 + \epsilon)$  approximation algorithm to the optimal solution. They performed an empirical study of their methods on synthetic and two real-world datasets (Twitter and Reddit), finding that there is space to reduce both controversy and disagreement in real-world networks.

Grossetti et al. (2020, 2021) studied communities on a large Twitter dataset to quantify how recommendation systems affect users’ behavior, and how content personalization can increase the echo chamber effect and create filter bubbles. A preliminary study was conducted to detect a filter bubble effect on users’ community profiles, proposing a community-aware model whose objective is to reduce the filter-bubble impact. This model can be deployed on top of any existing recommendation system, enhancing it with a new scoring function that permits re-ranking the recommendations. To determine the similarity between communities, the authors considered (1) topology, (2) semantic information, and (3) flows of information between these communities. The model works with a set of recommendations, selected from the recommendation list produced by the recommendation system, and a community score vector that matches as much as possible the user profile. The authors showed that their solution improved the quality of recommendations by matching more closely the users’ community profile and by reducing the filter bubble effect at a limited computation cost.

Mendes et al. (2019) presented a platform for crowdsourced social participation that increases engagement and counteracts the formation of opinion bubbles and the echo chamber effect of social networks. The authors argued that clearly separated opinion groups could not necessarily indicate polarization but might instead stem from rational disagreements. Polarized discussions

arise from distorted perceptions about the other group's motivations and points of view. The platform organizes topics of discussion around "conversations." Users can insert comments associated with a conversation or vote if they agree or disagree with other user's propositions. As the users interact with some conversation, it gradually becomes possible to recognize the different opinion profiles. As soon as the platform can classify users into distinct opinion groups, it displays results to all participants. Therefore, not only the owners of each conversation can extract useful metrics to guide policy and decision making. Each participant can access its own opinion profile or those from their network of peers and compare it with the whole. According to the authors, the proposed form of transparency towards information is an important factor to counteract the echo chamber effect.

Sacharidis (2019) studied the effect of social-based recommendation systems on the diversity and novelty of the recommendations they make, questioning whether they lead to the formation of echo chambers. Social-based recommendation systems seek to exploit the effects of homophily and influence observed in social networks in order to improve their accuracy. This is achieved by enforcing similar preferences among users that are socially connected. The Douban dataset used in the numerical evaluation concerns a popular Chinese social networking service that allows users to connect to each other and provide content and ratings for movies, books, music, and events. The results indicated that social-based recommendations can often increase the diversity and novelty of user recommendations when measured individually and when examined with respect to the social groups to which users belong. The author also claimed that the social-based recommendations resulted in more accurate recommendations, while not sacrificing diversity and novelty.

Donkers and Ziegler (2021) alleged that, while most scientific work has framed echo chambers due to imbalances between polarized communities, members of the echo chambers often actively discredit outside sources to maintain coherent world views. They argued that two different types of echo chambers occur in social media: epistemic echo chambers create information gaps mainly through their structure, whereas ideological ones systematically exclude counterattitudinal information. An agent-based modeling approach was applied to investigate the characteristics of this dual echo chamber view. To assess the depolarizing effects of diversified recommendations, the authors relied on knowledge graph embeddings (Dai et al., 2020). For community detection, they employed the Louvain algorithm (De Meo et al., 2011). To quantify the segregation between communities, modularity, and homophily measures were used. The results showed that counteracting the two different types of echo chambers may require fundamentally different diversification strategies. Moreover, interventions seem to be most effective when a discussion first emerges. This shows the importance of what people observe when they first engage with a new topic.

Morales and Cointet (2021) studied the impact of recommendation systems on polarization. They presented the analysis of friend recommendations using real-world networks, where nodes (users) have dynamical positions in an ideological space, and where dimensions are indicators of attitudes towards issues in the public debate. The network evolves following a recommendation system, and the users opinions coevolve following a DeGroot opinion model (DeGroot, 1974). The authors applied the Duclos–Esteban–Ray polarization measure (Duclos et al., 2004), which is an extension of the Esteban–Ray measure (Esteban and Ray, 1994) of the distribution of user attitudes in the ideological space. The authors showed that different well-known recommendation systems can modify the convergence or divergence of social systems, affecting the evolution of polarization. For evaluating the effects of different recommendation systems on polarization, the authors used

subgraphs of the Twitter network in the vicinity of French parliamentarians. The results indicated that some recommendation systems can decrease polarization, while others can increase it, leading the authors to argue that the use of a specific recommendation system can drive or mitigate the polarization appearing in real social networks.

Fabbri et al. (2022) studied the problem of mitigating radicalization pathways that occur when a user is exposed to polarized content, subsequently receiving increasingly radicalized recommendations. The authors model the set of recommendations of a “what-to-watch-next” recommendation system as a  $d$ -regular directed graph where nodes correspond to content items, links to recommendations, and paths to user sessions. They measured the polarity of a node as the expected length of a random walk from that node to any node representing nonradicalized content. High segregation scores are associated with larger chances to get users trapped in radicalization pathways. The problem of reducing the prevalence of radicalization pathways consists of selecting a small number of edges to “rewire” (following a similar idea to that in Interian et al., 2020, 2021) so the maximum segregation score among all radicalized nodes is minimized while maintaining the relevance of the recommendations. Finding the optimal set of recommendations to rewire is proved to be NP-hard, and a greedy algorithm is proposed for its solution.

#### 4.4. Other approaches

This section discusses other polarization mitigation models and strategies that do not fit in the approaches exposed in the previous sections.

Chen and Racz (2021) affirmed that while some actors spread misinformation to push a specific agenda, others aim to disrupt the network by increasing disagreement and polarization, thereby destabilizing society. Motivated by this phenomenon, the authors introduced a simple adversarial model of network disruption, where an adversary can take over a limited number of user profiles in a social network with the aim of maximizing disagreement, defined as a measure of how much neighboring nodes disagree in their opinions across the network. Polarization was defined as the variance of the opinions of all nodes, multiplied by the number of nodes. The authors investigated their model both theoretically and empirically, showing that the adversary will always change the opinion of a taken-over profile to an extreme to maximize disruption. They also showed that the adversary can increase polarization at most linearly in the number of user profiles it takes over. An empirical study of six adversarial heuristics on synthetic and real-world Reddit and Twitter networks was presented. The key conclusion was that the adversary can significantly disrupt the network (increasing polarization) using simple methods, such as targeting centrists.

Arruda et al. (2021) approached the appearance of echo chambers under the so-called underdog effect, emphasizing the influence of contrarian opinions in a multiopinion scenario. The underdog effect is the tendency to support the less popular option. A modified Sznajd opinion dynamics model (Benatti et al., 2020) is used. The authors considered an adaptation of the Sznajd model with the possibility of friendship rewiring, performed on several network topologies. They analyzed the relationship between topology and opinion dynamics by considering two measures: opinion diversity and modularity. Two strategies were tested: (1) the agents can reconnect only with others sharing the same opinion and (2) same as in the previous case, but with the agents reconnecting only within their limited neighborhood. The authors found that the underdog effect, if strong enough,

can increase the heterogeneity of opinions. This effect decreases the possibility of echo chamber formation. The number of opinions did not strongly affect the steady state of the network dynamics.

## 5. Conclusions

This review examined the most used network polarization measures and polarization reduction strategies. The use of measures based on homophily, modularity, random walks, content qualification, and balance theory has been proposed for measuring network polarization, also called in some studies by the terms controversy, disagreement, and conflict. Polarization reduction strategies included node or edge modifications (including edge insertions or deletions, and adjustments in edge weights), changes in network design, and changes in the recommendation systems embedded in the networks.

These polarization measures and reduction strategies have been used in many case studies and real-life applications in different fields, such as partisan and political polarization, polarization in digital media consumption, climate change discussions, vaccination debates, and scientific coauthorship and collaboration.

Some studies analyzed partisan polarization in parliaments, online participatory platforms, web, and blogging networks in the USA (Reese et al., 2007; Zhang et al., 2008; Garimella and Weber, 2017; Shi et al., 2017; Dinkelberg et al., 2021), Italy (Dal Maso et al., 2015), Switzerland (Garcia et al., 2015), Germany (Markgraf and Schoch, 2019), France (Morales and Cointet, 2021), Israel (Wolfowicz et al., 2021), Venezuela (Morales et al., 2015), India and Pakistan (Haq et al., 2020), and a group of 16 European countries (Maoz and Somer-Topcu, 2010). For instance, Markgraf and Schoch (2019) reported a case study based on data from the German Federal Election of 2017 for illustrating their echo chambers research framework. Morales and Cointet (2021) used subgraphs of the Twitter network in the vicinity of French parliamentarians, reporting that some recommendation systems can decrease polarization, while others can increase it. Maoz (2006), instead, analyzed the political polarization of alliances between different countries and its effect on conflicts among states.

Other authors studied online news consumption in mass media, newspapers, social networks, and blogs to identify factors that lead to polarization. Flaxman et al. (2016) examined the web-browsing patterns of 50,000 anonymized U.S.-located Internet users, showing that articles found via social media or web search engines are associated with higher ideological segregation than those found directly visiting news sites. Ksiazek (2011) and Webster and Ksiazek (2012) studied audience fragmentation, investigating how people allocate their attention across digital media. The authors found little evidence that audiences were composed of devoted loyalists, but the results suggested the presence of linguistic polarization. Tien et al. (2020) compared groups of Twitter users who participated in the conversation about the Charlottesville events (New York Times, 2017) in the USA, finding that the retweet network largely splits according to user media preferences.

Climate change emerged as one of the most polarizing discussion topics. Cook and Lewandowsky (2016) investigated belief polarization, showing that the worldview has a dominant influence on climate beliefs (the authors used free-market support as a proxy for participants' worldview). Jasny et al. (2018) investigated the information diffusion process among climate policy networks in the United States, finding an increase in the number of arcs that generate echo chambers from 2010

to 2016. Samantray and Pin (2019) studied a model of belief polarization in social networks that consider, in addition to the homophily (see Section 3.1), the information credibility. They concluded that tweets expressing antclimate change beliefs are largely not credible to the broader society.

The COVID-19 pandemic attracted great attention to the echo chambers of the anti-vaccine community, which led to a decline in vaccination rates. Horne et al. (2015) showed that it is possible to successfully counter people's anti-vaccination attitudes by making them appreciate the consequences of failing to vaccinate their children. Cossard et al. (2020) analyzed the Italian vaccination debate on Twitter. The two sides of the debate, one formed by vaccine advocates and the other formed by users skeptical about vaccination, tended to ignore each other's content, potentially leaving skeptics' concerns unanswered.

Some authors evaluated the polarization in coauthorship and collaboration networks. Soós and Kampis (2011) analyzed the leading Hungarian research organizations, comparing the diversity of their publication performance and the polarity of each researcher's profile. Leifeld (2018) analyzed two research traditions (the hermeneutic and the nomological) in the coauthorship network of researchers in Germany and Switzerland. A higher similarity between researchers leads to a greater probability of coauthorship, showing a homophilic behavior between hermeneutic and nomological researchers observed by philosophers of science.

The anonymous communication provided by social networks, to a great extent protected by the distance between those who participate in the dialogue, creates an incentive to expose more extreme opinions without the usual constraints that characterize the behavior observed in physical or face-to-face interactions. Individuals often soften their ideas during face-to-face interactions not to hurt other persons' sensibilities. The digital environment frees individuals to express their views more openly, without concerns about the possible reactions that these opinions may have on others, often triggering heated and polarizing attitudes. Attempts to encourage dialogue and improve intergroup communication might mitigate the extreme polarization in social networks.

## Acknowledgments

The authors are grateful to the suggestions and recommendations of two anonymous referees who contributed to improving the readability of this article. The work of Ruben Interian was supported by FAPERJ post-doctorate research grant E-26/204.200/2021. The work of Ruslán G. Marzo was supported by the FAPERJ doctorate scholarship E-26/200.330/2020. Isela Mendoza was supported by CNPq doctorate scholarship 143289/2021-7. Celso C. Ribeiro was supported by research grants CNPq 309869/2020-0 and FAPERJ E-26/200.926/2021. This work was also partially sponsored by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), under Finance Code 001.

## References

- Alsinet, T., Argelich, J., Béjar, R., Gibert, D., Planes, J., Torrent, N., 2021a. Argumentation reasoning with graph neural networks for reddit conversation analysis. *Frontiers in Artificial Intelligence and Applications* 339, 123–131.

- Alsinet, T., Argelich, J., Béjar, R., Martínez, S., 2021b. Measuring polarization in online debates. *Applied Sciences* 11, 11879.
- Alvim, M.S., Knight, S., Valencia, F., 2019. Toward a formal model for group polarization in social networks. In *The Art of Modelling Computational Systems: A Journey from Logic and Concurrency to Security and Privacy*, Lecture Notes in Computer Science. Vol. 11760. Springer, Cham, pp. 419–441.
- Antikacioglu, A., Ravi, R., 2017. Post processing recommender systems for diversity. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, New York, pp. 707–716.
- Aref, S., Dinh, L., Rezapour, R., Diesner, J., 2020. Multilevel structural evaluation of signed directed social networks based on balance theory. *Scientific Reports* 10, 15228.
- Arendt, H., 1968. *Between Past and Future*. Viking Press, New York.
- Arruda, H.F., Benatti, A., Silva, F.N., Comin, C.H., Costa, L.F., 2021. Contrarian effects and echo chamber formation in opinion dynamics. *Journal of Physics: Complexity* 2, 025010.
- Badami, M., Nasraoui, O., Sun, W., Shafto, P., 2017. Detecting polarization in ratings: an automated pipeline and a preliminary quantification on several benchmark data sets. In *Proceedings of the 2017 IEEE International Conference on Big Data*, IEEE, Piscataway, NJ, pp. 2682–2690.
- Bail, C.A., Argyle, L.P., Brown, T.W., Bumpus, J.P., Chen, H., Fallin Hunzaker, M.B., Lee, J., Mann, M., Merhout, F., Volfovsky, A., 2018. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences* 115, 9216–9221.
- Balietti, S., Getoor, L., Goldstein, D.G., Watts, D.J., 2021. Reducing opinion polarization: effects of exposure to similar people with differing political views. *Proceedings of the National Academy of Sciences* 118, 52, e2112552118.
- BBC News, 2022. Freedom convoy: Why Canadian truckers are protesting in Ottawa. Available at <https://www.bbc.com/news/world-us-canada-60164561> (accessed 22 May 2022).
- Becker, J., Porter, E., Centola, D., 2019. The wisdom of partisan crowds. *Proceedings of the National Academy of Sciences* 116, 10717–10722.
- Benatti, A., Arruda, H.F., Silva, F.N., Comin, C.H., Costa, L.F., 2020. Opinion diversity and social bubbles in adaptive Sznajd networks. *Journal of Statistical Mechanics: Theory and Experiment* 2020, 023407.
- Borrelli, D., Iandoli, L., Ramirez-Marquez, J.E., Lipizzi, C., 2022. A quantitative and content-based approach for evaluating the impact of counternarratives on affective polarization in online discussions. *IEEE Transactions on Computational Social Systems* 9, 3, 914–925.
- Bozdag, E., Gao, Q., Houben, G.J., Warnier, M., 2014. Does offline political segregation affect the filter bubble? An empirical analysis of information diversity for Dutch and Turkish Twitter users. *Computers in Human Behavior* 41, 405–415.
- Buskens, V., Corten, R., Weesie, J., 2008. Consent or conflict: coevolution of coordination and networks. *Journal of Peace Research* 45, 205–222.
- Cartwright, D., Harary, F., 1956. Structural balance: a generalization of Heider's theory. *Psychological Review* 63, 277–293.
- 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. In *2019 Twelfth International Conference Management of Large-Scale System Development*, IEEE, Moscow, pp. 1–5.
- Chen, M.F., Racz, M.Z., 2021. An adversarial model of network disruption: maximizing disagreement and polarization in social networks. *IEEE Transactions on Network Science and Engineering* 9, 728–739.
- Chen, X., Lijffijt, J., De Bie, T., 2018. Quantifying and minimizing risk of conflict in social networks. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, New York, pp. 1197–1205.
- Chitra, U., Musco, C., 2020. Analyzing the impact of filter bubbles on social network polarization. *Proceedings of the 13th International Conference on Web Search and Data Mining*. ACM, New York, pp. 115–123.
- Chkhartishvili, A.G., Kozitsin, I., 2018. Binary separation index for echo chamber effect measuring. *2018 Eleventh International Conference Management of Large-Scale System Development*, IEEE, Moscow, pp. 1–4.
- Cinelli, M., De Francisci Morales, G., Galeazzi, A., W, Q., Starnini, M., 2021. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences* 118, e2023301118.
- Conover, M.D., Gonçalves, B., Flammini, A., Menczer, F., 2012. Partisan asymmetries in online political activity. *EPJ Data Science* 1, 6.



- Cook, J., Lewandowsky, S., 2016. Rational irrationality: modeling climate change belief polarization using Bayesian networks. *Topics in Cognitive Science* 8, 160–179.
- Cossard, A., De Francisci Morales, G., Kalimeri, K., Mejova, Y., Paolotti, D., Starnini, M., 2020. Falling into the echo chamber: the Italian vaccination debate on Twitter. Proceedings of the 14th International AAAI Conference on Web and Social Media, AAAI, Menlo Park, CA, pp. 130–140.
- Cota, W., Ferreira, S.C., Pastor-Satorras, R., Starnini, M., 2019. Quantifying echo chamber effects in information spreading over political communication networks. *EPJ Data Science* 8, 35.
- Cremonini, M., Casamassima, F., 2017. Controllability of social networks and the strategic use of random information. *Computational Social Networks* 4, 10.
- Currarini, S., Jackson, M.O., Pin, P., 2009. An economic model of friendship: homophily, minorities, and segregation. *Econometrica* 77, 1003–1045.
- Dai, Y., Wang, S., Xiong, N.N., Guo, W., 2020. A survey on knowledge graph embedding: approaches, applications and benchmarks. *Electronics* 9, 750.
- Dal Maso, C., Pompa, G., Puliga, M., Riotta, G., Chessa, A., 2015. Voting behavior, coalitions and government strength through a complex network analysis. *PLoS ONE* 9, 1–13.
- De Meo, P., Ferrara, E., Fiumara, G., Provetti, A., 2011. Generalized Louvain method for community detection in large networks. 2011 11th International Conference on Intelligent Systems Design and Applications, pp. 88–93.
- DeGroot, M.H., 1974. Reaching a consensus. *Journal of the American Statistical Association* 69, 345, 118–121.
- Dinkelberg, A., O'Reilly, C., MacCarron, P., Maher, P.J., Quayle, M., 2021. Multidimensional polarization dynamics in us election data in the long term (2012–2020) and in the 2020 election cycle. *Analyses of Social Issues and Public Policy* 21, 284–311.
- Donkers, T., Ziegler, J., 2021. The dual echo chamber: modeling social media polarization for interventional recommending. Fifteenth ACM Conference on Recommender Systems, ACM, pp. 12–22.
- Du, S., Gregory, S., 2017. The echo chamber effect in Twitter: Does community polarization increase? In Cherifi, H., Gaito, S., Quattrociocchi, W. and Sala, A. (eds), *Complex Networks & Their Applications V*. Springer, Cham, pp. 373–378.
- Duclos, J.Y., Esteban, J., Ray, D., 2004. Polarization: concepts, measurement, estimation. *Econometrica* 72, 1737–1772.
- El-Moussaoui, M., Agouti, T., Tikniouine, A., El-Adnani, M., 2019. A comprehensive literature review on community detection: approaches and applications. *Procedia Computer Science* 151, 295–302.
- Elsevier, 2018. What is the difference between ScienceDirect and Scopus data? Available at <https://nonprod-devportal.elsevier.com/support.html> (accessed 22 May 2022).
- 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, 90.
- EMOL, 2022. Portal informativo EMOL. Available at <https://www.emol.com> (accessed on 19 March 2022).
- Encyclopedia Britannica, 2021. Confirmation bias. Available at <https://www.britannica.com/science/confirmation-bias> (accessed 22 May 2022).
- Ertan, G., Çarkoğlu, A., Erdem Aytac, S., 2022. Cognitive political networks: a structural approach to measure political polarization in multiparty systems. *Social Networks* 68, 118–126.
- Esteban, J.M., Ray, D., 1994. On the measurement of polarization. *Econometrica* 62, 819–851.
- Fabbri, F., Wang, Y., Bonchi, F., Castillo, C., Mathioudakis, M., 2022. Rewiring what-to-watch-next recommendations to reduce radicalization pathways. Proceedings of the ACM Web Conference 2022, ACM, New York, pp. 2719–2728.
- Finn, S., Mustafaraj, E., Metaxas, P.T., 2014. The co-retweeted network and its applications for measuring the perceived political polarization. Proceedings of the 10th International Conference on Web Information Systems and Technologies, ACM, New York, pp. 276–284.
- Flaxman, S., Goel, S., Rao, J.M., 2016. Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly* 80, 298–320.
- France 24, 2019. A year of insurgency: How Yellow Vests left 'indelible mark' on French politics. Available at <https://www.france24.com/en/20191116-a-year-of-insurgency-how-yellow-vests-left-indelible-mark-on-french-politics>, (accessed 22 May 2022).
- Friedkin, N.E., Johnsen, E.C., 1990. Social influence and opinions. *The Journal of Mathematical Sociology* 15, 193–206.

- Galam, S., 2002. Minority opinion spreading in random geometry. *The European Physical Journal B - Condensed Matter and Complex Systems* 25, 403–406.
- Garcia, D., Abisheva, A., Schweighofer, S., Serdült, U., Schweitzer, F., 2015. Ideological and temporal components of network polarization in online political participatory media. *Policy and Internet* 7, 46–79.
- Garimella, K., De Francisci Morales, G., Gionis, A., Mathioudakis, M., 2016. Quantifying controversy in social media. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, ACM, New York, pp. 33–42.
- Garimella, K., De Francisci Morales, G., Gionis, A., Mathioudakis, M., 2017a. Mary, Mary, quite contrary: exposing Twitter users to contrarian news. *Proceedings of the 26th International Conference on World Wide Web Companion, International World Wide Web Conferences Steering Committee*, Geneva, Switzerland, pp. 201–205.
- Garimella, K., De Francisci Morales, G., Gionis, A., Mathioudakis, M., 2017b. Reducing controversy by connecting opposing views. *Proceedings of the 10th ACM International Conference on Web Search and Data Mining*, ACM, New York, pp. 81–90.
- Garimella, K., De Francisci Morales, G., Gionis, A., Mathioudakis, M., 2018a. Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. *Proceedings of the 2018 World Wide Web Conference, International World Wide Web Conferences Steering Committee*, Geneva, Switzerland, pp. 913–922.
- Garimella, K., De Francisci Morales, G., Gionis, A., Mathioudakis, M., 2018b. Quantifying controversy on social media. *ACM Transactions on Social Computing* 1, 3, 1–27.
- Garimella, K., De Francisci Morales, G., Gionis, A., Mathioudakis, M., 2018c. Reducing controversy by connecting opposing views. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, Stockholm, Sweden, pp. 5249–5253.
- Garimella, K., Weber, I., 2017. A long-term analysis of polarization on Twitter. *Proceedings of the Eleventh International AAAI Conference on Web and Social Media*, pp. 528–531.
- Gass, R.H., 2015. Social influence, sociology of. In Wright, J.D. (ed.) *International Encyclopedia of the Social & Behavioral Sciences* (2nd edn). Elsevier, Oxford, pp. 348–354.
- Ghezelbash, E., Yazdanpanah, M.J., Asadpour, M., 2019. Polarization in cooperative networks through optimal placement of informed agents. *Physica A: Statistical Mechanics and its Applications* 536, 120936.
- Gillani, N., Yuan, A., Saveski, M., Vosoughi, S., Roy, D., 2018. Me, my echo chamber, and I: Introspection on social media polarization. *Proceedings of the 2018 World Wide Web Conference*, pp. 823–831.
- Google, 2019. Google Ngram viewer. Available at <https://books.google.com/ngrams> (accessed 22 May 2022).
- Grossetti, Q., du Mouza, C., Travers, N., 2020. Community-based recommendations on Twitter: Avoiding the filter bubble. In Cheng, R., Mamoulis, N., Sun, Y., Huang, X. (eds) *Web Information Systems Engineering – WISE 2019*. Lecture Notes in Computer Science, Vol. 11881. Springer, Cham, pp. 212–227.
- Grossetti, Q., du Mouza, C., Travers, N., Constantin, C., 2021. Reducing the filter bubble effect on Twitter by considering communities for recommendations. *International Journal of Web Information Systems* 17, 728–752.
- Guterres, A., 2018. Political, social polarization leading to rise in global insecurity, Secretary-General's report finds. Available at <https://www.un.org/press/en/2018/org1681.doc.htm> (accessed 22 May 2022).
- Guyot, A., Gillet, A., Leclercq, É., Cullot, N., 2022. ERIS: an approach based on community boundaries to assess polarization in online social networks. In Guizzardi, R., Ralyté, J., Franch, X. (eds) *Research Challenges in Information Science*. Springer, Berlin, pp. 88–104.
- Haddadan, S., Menghini, C., Riondato, M., Upfal, E., 2021. RePBubLik: Reducing polarized bubble radius with link insertions. *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, ACM, New York, pp. 139–147.
- Hamilton, W.L., Ying, R., Leskovec, J., 2017. Inductive representation learning on large graphs. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Curran Associates, Red Hook, pp. 1025–1035.
- Haq, E.u., Braud, T., Kwon, Y.D., Hui, P., 2020. Enemy at the gate: Evolution of Twitter user's polarization during national crisis. *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ACM, New York, pp. 212–216.
- Harary, F., 1959. On the measurement of structural balance. *Behavioral Science* 4, 316–323.
- Heider, F., 1946. Attitudes and cognitive organization. *The Journal of Psychology* 21, 107–112.
- Horne, Z., Powell, D., Hummel, J.E., Holyoak, K.J., 2015. Countering antivaccination attitudes. *Proceedings of the National Academy of Sciences* 112, 10321–10324.

- Huang, Z., Silva, A., Singh, A., 2022. Pole: polarized embedding for signed networks. In Proceedings, of the Fifteenth ACM International Conference on Web Search and Data Mining, Association for Computing Machinery, New York, pp. 390–400.
- Interian, R., Moreno, J.R., Ribeiro, C.C., 2020. Reducing network polarization by edge additions. Proceedings of the 2020 4th International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence, ACM, New York, pp. 87–92.
- Interian, R., Moreno, J.R., Ribeiro, C.C., 2021. Polarization reduction by minimum-cardinality edge additions: complexity and integer programming approaches. *International Transactions in Operational Research* 28, 1242–1264.
- Interian, R., Ribeiro, C.C., 2018. An empirical investigation of network polarization. *Applied Mathematics and Computation* 339, 651–662.
- Interian, R., Ribeiro, C.C., 2019. An iterated greedy heuristic for the minimum-cardinality balanced edge addition problem. Proceeding of the Metaheuristics International Conference 2019, Cartagena, pp. 195–198.
- Jasny, L., Dewey, A.M., Robertson, A.G., Yagatch, W., Dubin, A.H., Waggle, J.M., Fisher, D.R., 2018. Shifting echo chambers in US climate policy networks. *PLoS ONE* 13, e0203463.
- Kawada, Y., Nakamura, Y., Sunada, K., 2018. A characterization of the Esteban–Ray polarization measures. *Economics Letters* 169, 35–37.
- Koidl, K., 2018. Towards trust-based decentralized ad-hoc social networks. Companion Proceedings of the World Wide Web Conference 2018, Geneva, Switzerland, pp. 1545–1551.
- Krackhardt, D., Stern, R.N., 1988. Informal networks and organizational crises: an experimental simulation. *Social Psychology Quarterly* 51, 123–140.
- Ksiazek, T.B., 2011. A network analytic approach to understanding cross-platform audience behavior. *Journal of Media Economics* 24, 237–251.
- Leifeld, P., 2018. Polarization in the social sciences: assortative mixing in social science collaboration networks is resilient to interventions. *Physica A: Statistical Mechanics and its Applications* 507, 510–523.
- Lelkes, Y., 2016. Mass polarization: manifestations and measurements. *Public Opinion Quarterly* 80, 392–410.
- Liu, L., Wang, X., Chen, X., Tang, S., Zheng, Z., 2021. Modeling confirmation bias and peer pressure in opinion dynamics. *Frontiers in Physics* 9, 649852.
- Madsen, J.K., Bailey, R.M., Pilditch, T.D., 2018. Large networks of rational agents form persistent echo chambers. *Scientific Reports* 8, 12391.
- Maoz, Z., 2006. Network polarization, network interdependence, and international conflict, 1816–2002. *Journal of Peace Research* 43, 391–411.
- Maoz, Z., Somer-Topcu, Z., 2010. Political polarization and cabinet stability in multiparty systems: a social networks analysis of European parliaments, 1945–98. *British Journal of Political Science* 40, 805–833.
- Markgraf, M., Schoch, M., 2019. Quantification of echo chambers: a methodological framework considering multi-party systems. 27th European Conference on Information Systems, Stockholm, Sweden. Available at [https://aisel.aisnet.org/ecis2019\\_rp/91](https://aisel.aisnet.org/ecis2019_rp/91) (accessed 22 May 2022).
- Mendes, F.M., Poppi, R., Parra, H., Moreira, B., 2019. EJ: A free software platform for social participation. In Borgeleau, F., Sillitti, A., Meirelles, P., Lenarduzzi, V. (eds) *Open Source Systems*. IFIP Advances in Information and Communication Technology, Vol. 556. Springer, Cham, pp. 27–37.
- Mendoza, M., Parra, D., Soto, A., 2020. GENE: graph generation conditioned on named entities for polarity and controversy detection in social media. *Information Processing and Management* 57, 102366.
- Mill, J.S., 1859. *On Liberty*. J. W. Parker and Son, London.
- Morales, A.J., Borondo, J., Losada, J.C., Benito, R.M., 2015. Measuring political polarization: Twitter shows the two sides of Venezuela. *Chaos* 25, 033114.
- Morales, G.D.F., Monti, C., Starnini, M., 2021. No echo in the chambers of political interactions on Reddit. *Scientific Reports* 11, 2818.
- Morales, P.R., Cointet, J.P., 2021. Auditing the effect of social network recommendations on polarization in geometrical ideological spaces. Fifteenth ACM Conference on Recommender Systems, ACM, New York, pp. 627–632.
- Musco, C., Musco, C., Tsourakakis, C.E., 2018. Minimizing polarization and disagreement in social networks. Proceedings of the 2018 World Wide Web Conference, ACM, New York, pp. 369–378.

- New York Times, 2017. Far-right groups surge into national view in Charlottesville. Available at <https://www.nytimes.com/2017/08/13/us/far-right-groups-blaze-into-national-view-in-charlottesville.html> (accessed 22 May 2022).
- Newman, M.E.J., 2003. Mixing patterns in networks. *Physical Review E* 67, 026126.
- Newman, M.E.J., 2006a. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E* 74, 036104.
- Newman, M.E.J., 2006b. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences* 103, 8577–8582.
- Newman, M.E.J., Girvan, M., 2004. Finding and evaluating community structure in networks. *Physical Review E* 69, 026113.
- Nyhan, B., Reifler, J., 2010. When corrections fail: the persistence of political misperceptions. *Political Behavior* 32, 303–330.
- O'Hara, K., Stevens, D., 2015. Echo chambers and online radicalism: assessing the Internet's complicity in violent extremism. *Policy and Internet* 7, 401–422.
- Oxford, 2021. Polarization. Available at <https://www.lexico.com/definition/polarization> (accessed 22 May 2022).
- Pearson, K., 1901. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2, 559–572.
- Pedersen, M.Y., Smets, S., Ågotnes, T., 2020. Further steps towards a logic of polarization in social networks. In Dastani, M., Dong, H., van der Torre, L. (eds), *Logic and Argumentation*. Lecture Notes in Computer Science, Vol. 12061. Springer, Cham, pp. 324–345.
- Pedersen, M.Y., Smets, S., Ågotnes, T., 2021. Modal logics and group polarization. *Journal of Logic and Computation* 31, 2240–2269.
- Reese, S.D., Rutigliano, L., Hyun, K., Jeong, J., 2007. Mapping the blogosphere: Professional and citizen-based media in the global news arena. *Journalism* 8, 235–261.
- Rumshisky, A., Gronas, M., Potash, P., Dubov, M., Romanov, A., Kulshreshtha, S., Gribov, A., 2017. Combining network and language indicators for tracking conflict intensity. In Ciampaglia, G.L., Mashhadi, A., Yasseri, T. (eds) *Social Informatics*. Lecture Notes in Computer Science, Vol. 10540. Springer, Cham, pp. 391–404.
- Sacharidis, D., 2019. Diversity and novelty in social-based collaborative filtering. Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization, ACM, New York, pp. 139–143.
- Samantray, A., Pin, P., 2019. Credibility of climate change denial in social media. *Palgrave Communications* 5, 127.
- Santos, F.P., Santos, F.C., Pacheco, J.M., Levin, S.A., 2021. Social network interventions to prevent reciprocity-driven polarization. Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems, International Foundation for Autonomous Agents and Multiagent Systems, pp. 1643–1645.
- Shekatkar, S.M., 2019. Do zealots increase or decrease the polarization of social networks? *Journal of Complex Networks* 8, cnz036.
- Shi, Y., Mast, K., Weber, I., Kellum, A., Macy, M., 2017. Cultural fault lines and political polarization. Proceedings of the 2017 ACM Web Science Conference, ACM, New York, pp. 213–217.
- Shore, J., Baek, J., Dellarocas, C., 2018. Network structure and patterns of information diversity on Twitter. *MIS Quarterly: Management Information Systems* 42, 849–872.
- Soós, S., Kampis, G., 2011. Towards a typology of research performance diversity: the case of top Hungarian players. *Scientometrics* 87, 357–371.
- Spohr, D., 2017. Fake news and ideological polarization: filter bubbles and selective exposure on social media. *Business Information Review* 34, 150–160.
- Sunstein, C.R., 2003. *Why Societies Need Dissent*. Harvard University Press, Cambridge, MA.
- Sunstein, C.R., 2007. *Republic.com 2.0*. Princeton University Press, Princeton, NJ.
- The Conversation, 2020. Articles on social media banning. Available at <https://theconversation.com/us/topics/social-media-banning-71241> (accessed 22 May 2022).
- The New York Times, 2021. George Floyd protests: a timeline. Available at <https://www.nytimes.com/article/george-floyd-protests-timeline.html> (accessed 22 May 2022).
- The Washington Post, 2021. CONTAGION: Threats and disinformation spread across the country in the wake of the Capitol siege, shaking the underpinnings of American democracy. Available at <https://www.washingtonpost.com/politics/interactive/2021/fallout-jan-6-insurrection/> (accessed 22 May 2022).

- Tien, J.H., Eisenberg, M.C., Cherng, S.T., Porter, M.A., 2020. Online reactions to the 2017 ‘Unite the right’ rally in Charlottesville: measuring polarization in Twitter networks using media followership. *Applied Network Science* 5, 10.
- Turow, J., 1997. *Breaking Up America: Advertisers and the New Media World*. University of Chicago Press, Chicago, IL.
- Vicario, M.d., Scala, A., Caldarelli, G., Stanley, H.E., Quattrociocchi, W., 2017. Modeling confirmation bias and polarization. *Scientific Reports* 7, 40391.
- Washington, G., 1999. *George Washington’s Farewell Address: Little Books of Wisdom*. Applewood Books, Bedford.
- Webster, J.G., Ksiazek, T.B., 2012. The dynamics of audience fragmentation: Public attention in an age of digital media. *Journal of Communication* 62, 39–56.
- White, S., Smyth, P., 2003. Algorithms for estimating relative importance in networks. Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, pp. 266–275.
- Wohlin, C., Runeson, P., Höst, M., Ohlsson, M.C., Regnell, B., Wesslén, A., 2012. *Systematic Literature Reviews*. Springer, Berlin, pp. 45–54.
- Wolfowicz, M., Weisburd, D., B., H., 2021. Examining the interactive effects of the filter bubble and the echo chamber on radicalization. *Journal of Experimental Criminology*. <https://doi.org/10.1007/s11292-021-09471-0>
- Xiong, Z., Ågotnes, T., 2020. On the logic of balance in social networks. *Journal of Logic, Language and Information* 29, 53–75.
- Yang, Z., Algesheimer, R., Tessone, C.J., 2016. A comparative analysis of community detection algorithms on artificial networks. *Scientific Reports* 6, 30750.
- Zadeh, L.A., 1965. Fuzzy sets. *Information and Control* 8, 338–353.
- Zhang, Y., Friend, A.J., Traud, A.L., Porter, M.A., Fowler, J.H., Mucha, P.J., 2008. Community structure in congressional cosponsorship networks. *Physica A: Statistical Mechanics and its Applications* 387, 1705–1712.
- Zollo, F., Bessi, A., Vicario, M.D., Scala, A., Caldarelli, G., Shekhtman, L., Havlin, S., Quattrociocchi, W., 2017. Debunking in a world of tribes. *PLoS ONE* 12, e0181821.