Detection, Measurement, and Mitigation of Echo Chambers in Social Networks: A Survey

Xianju Zhu[1], Reynold Cheng[1], Laks V.S. Lakshmanan[2]
Xiaodong Li[1], Chenhao Ma[3], Mohammad Matin Najafi[1]
[1]The University of Hong Kong, Hong Kong SAR, China
[2]The University of British Columbia, Vancouver, B.C., Canada
[3]The Chinese University of Hong Kong, Shenzhen, China
zxj0302@connect.hku.hk, ckcheng@cs.hku.hk, laks@cs.ubc.ca,
xdli@cs.hku.hk, machenhao@cuhk.edu.cn, mohammad@cs.hku.hk

Abstract

Echo chambers in social networks are environments where like-minded users cluster together, often reinforcing shared beliefs while limiting exposure to opposing viewpoints. This phenomenon has profound implications for information diffusion, political discourse, and societal polarization. In this survey, we systematically review the landscape of echo chambers in online social networks, focusing on three key aspects: detection, measurement, and mitigation. We categorize existing methods for identifying echo chambers, analyze diverse metrics that quantify their effects, and examine intervention strategies aimed at alleviating their negative consequences. By synthesizing findings from recent literature and highlighting open challenges, our survey aims to provide a comprehensive reference for researchers and practitioners seeking to understand and address the echo chamber effect in contemporary digital platforms.

1 Introduction and Background

Online social networks, such as Twitter and Facebook, have fundamentally transformed the way individuals consume information, engage in discourse, and form social ties. While these platforms enable unprecedented connectivity and rapid information diffusion, they also give rise to complex social phenomena that challenge the health and diversity of public discourse. Among these, the *echo chamber* effect has emerged as a central concern for researchers across computer science, social science, and digital media studies.

Definition (Echo Chamber): An echo chamber is a phenomenon prevalent in online social networks, characterized by like-minded users predominantly interacting with each other. Within these echo chambers, users express and reinforce their beliefs on specific issues, thereby amplifying their shared viewpoints [14].

The terms *echo chamber* and *filter bubble* [82] carry distinct nuances but are frequently treated as synonyms in the literature [41]. While both limit exposure to diverse perspectives, their underlying mechanisms and effects differ. Echo chambers arise when users deliberately discredit and exclude dissenting views, fostering distrust of outsiders and resistance to counterevidence. In contrast, filter bubbles result from passive algorithmic curation that limits exposure to diverse perspectives, though users may remain open to opposing information when encountered. Recognizing this distinction is critical for designing interventions, as mitigating echo chambers requires addressing active exclusion and

psychological biases, whereas filter bubbles may be alleviated through algorithmic transparency and diversification [21].

The echo chamber effect has been linked to a range of negative consequences for both individuals and society at large. Five common attributes [21] associated with echo chambers are: the diffusion of misinformation [83, 84], the spread of conspiracy theories [85], the formation of social trends [86], increased political polarization [87], and the emotional contagion of users [88]. These dynamics can contribute to the amplification of extreme views, the hardening of group boundaries, and the erosion of trust in public institutions.

Despite growing attention, systematic understanding and management of echo chambers remain challenging. First, echo chambers are multifaceted: they involve both the structural arrangement of social ties (who interacts with whom) and the semantic alignment of opinions (what people believe and express). Second, the observable effects of echo chambers—such as polarization, misinformation, and network segregation—are not always easily separable from other social processes. As a result, existing research has produced a wide array of methods for detecting, quantifying, and mitigating echo chambers, yet a unified framework remains elusive.

While there have been a few recent surveys on the literature on echo chambers in specific contexts [14, 21, 41], they exclude one of the key aspects: detection, measurement, or mitigation, or omit many recent and relevant studies due to their publication date, or lack a coherent framework for classifying existing approaches. Motivated by these gaps, our survey provides a comprehensive and up-to-date synthesis of the literature on echo chambers, with a particular focus on social networks. We systematically compare and extend prior surveys by incorporating recent work, offering a clear and structured taxonomy of methods, and highlighting overlooked yet important perspectives. Our goal is to offer a one-stop reference for both scholars and practitioners seeking to understand the full landscape of echo chamber research. Throughout, we draw attention to the conceptual nuances that distinguish echo chambers from related ideas, and we identify open challenges and promising directions for future inquiry.

Scope of this survey: This paper provides a comprehensive review of echo chamber research in online social networks, focusing on three interrelated aspects: *detection*, *measurement*, and *mitigation*. We systematically categorize methods for identifying echo chambers, analyze the diverse metrics used to quantify their effects, and survey intervention strategies designed to alleviate their negative impact. Our discussion emphasizes the interplay between network topology and semantic content, and highlights both methodological advances and open challenges. While echo chamber-like phenomena have also been observed in domains such as e-commerce recommender systems [15], we restrict our focus to social media and communication networks, where the social, political, and informational stakes are particularly pronounced.

For further reading on the definitions, attributes, mechanisms, and modeling of echo chambers and related risks, readers are encouraged to consult recent surveys such as [14, 21]. The remainder of this paper is organized as follows: Section 2 surveys methods for echo chamber detection; Section 3 reviews metrics for measuring echo chamber effects; Section 4 discusses mitigation strategies; Section 5 outlines open challenges and opportunities; and Section 6 concludes with a summary and future outlook.

2 Echo Chamber Detection

In this section, we summarize existing methods for echo chamber detection, which can be broadly classified based on their use of semantic and/or topological information: (1) *Topology-Based Detection*, which treats echo chambers as communities characterized by dense internal interactions and sparse external connections. Methods in this category commonly apply community detection algorithms—such as the Louvain algorithm—emphasizing metrics related to inter-community edges; (2) *Content-Based*

Detection, which focuses on analyzing the semantic content generated by users (e.g., tweets or posts) to identify differences, similarities, or polarization within the network; and (3) *Hybrid Detection*, which integrates both topological and semantic information to detect communities with high internal alignment, pronounced polarization between groups, and limited cross-community interaction. Table 1 summarizes related works, detailing their classification, detection methods, and a brief description of their detection logic or graph construction process.

Table 1: Comparison of Echo Chamber Detection Methods

Category	**Detection Method**	**Reference**
Topology-Based	Community Detection (Fast Greedy [78], Infomap [80], Louvain [79], WalkTrap [81], etc)	[22–25, 31, 51, 53–56, 58]
	Graph Partition (METIS [59])	[22, 60]
Content-Based	Post-comment pair classification	[4]
	Engagement in opposing views	[1]
Hybrid	METIS [59] on semantic-enriched graph	[6]
	Maximizing the likelihood of observed cascades	[8]
	Polarized communities detection in signed networks	[97, 98]

2.1 Topology-Based Detection

Based on the definition and intuition of an echo chamber, users interact much more frequently within their own communities than with those outside. Therefore, topology-based detection methods focus on network structure and construct social networks using various user interactions, such as retweets, replies, and follows. They then identify communities by finding groups with dense internal connections and sparse interactions with users outside the group.

Cossard et al.[22] detect echo chambers in the Italian vaccination debate on Twitter by leveraging the structural properties of interaction networks, particularly the retweet network. They construct a weighted, directed retweet network where nodes represent users and edges represent repeated retweet interactions, filtering out low-weight edges to reduce noise. Using the graph partition algorithm, METIS [59], they repeatedly bi-partition the giant connected component of this network to assign each user a "leaning score" based on retweet behavior, effectively mapping users to one of two polarized communities: vaccine advocates and vaccine skeptics. Manual annotation of a random user sample confirmed high classification accuracy. Similarly, Amendola et al. [60] model the social network as a graph, with nodes representing users and edges denoting topological relationships such as mentions, retweets, or follower/followee links. To detect communities within this network, they apply the METIS algorithm, which partitions the graph into balanced communities based on user interaction patterns.

Community detection algorithms are widely used to identify echo chambers in social networks. For instance, the Fast Greedy algorithm [78] is utilized in [22, 24], while the Louvain method [79] has been applied in several studies [22, 23, 31, 51, 53–56, 58]. Similarly, Infomap [80] has been adopted in [22, 25]. These methods are typically applied to networks constructed from retweet, quote, mention, follow, comment, or reply interactions collected from Twitter or other platforms, as well as user-item

interactions in recommender systems. The resulting communities are then analyzed as potential echo chambers.

2.2 Content-Based Detection

Since content generated or endorsed by users often reflects their opinions, analyzing such content through stance detection or emotion classification enables the quantification of underlying semantic signals. Applying these techniques allows for the inference of a user's stance based on their interactions, facilitating the identification of users participating in echo chambers.

Calderon et al. [4] propose a novel content-based approach for detecting echo chambers on social media, specifically focusing on Facebook fan pages. Instead of relying on complete network structures, their method analyzes the content of posts and their associated comments to extract linguistic features indicative of echo chamber behavior. They design two main types of features: Target Stance, which captures whether a comment agrees or disagrees with the stance of the original post, and Emotion Intensity, which measures the type and strength of emotions expressed. These features are extracted using graph-based linguistic pattern mining and are then fed into a neural network model (ECHO model) that uses attention mechanisms to classify post-comment pairs as echoing or not. By aggregating these classifications at the fan page level, they compute an Echo Chamber Index that quantifies the degree of echo chamber behavior present. Their experiments demonstrate that this approach outperforms traditional models in both English and Mandarin datasets, highlighting the effectiveness of content-based echo chamber detection.

Del Vicario et al. [1] detect echo chambers on Facebook by systematically analyzing user interactions with public Facebook pages categorized as either science or conspiracy. The authors first classify pages into these two categories based on their content and self-description. To identify echo chambers, they examine each user's commenting behavior over time and assign users to a community if at least 95% of their total comments are made on posts within either the science or within the conspiracy category. This high threshold ensures that only users with a strong and consistent preference for one type of content are included in the respective echo chamber. By applying this method across their dataset, the authors empirically isolate two distinct, polarized communities—one centered on scientific information and the other on conspiracy theories. This quantitative approach allows for the objective detection and longitudinal tracking of echo chambers, rooted in users' selective engagement and reinforcement of like-minded content.

2.3 Hybrid Detection

Topology-based detection methods utilize structural information to identify segregation between communities, but cannot ensure that users within the same community are highly ideologically aligned based solely on topology. Content-based detection methods leverage information from user-generated or endorsed content to infer ideological leanings, but do not determine whether these users are structurally divided into opposing groups. To more accurately detect echo chambers, some approaches combining both topological and semantic information have been proposed.

In [8], Minici et al. introduce a probabilistic generative model that integrates content-based and semantic-based approaches to identify latent communities with echo chamber characteristics within social networks. By modeling communities based on their polarization and opinion polarity, the method employs a scalable adaptation of the Generalized Expectation Maximization algorithm to optimize the joint likelihood of social connections and information cascades. This dual approach captures the propagation of ideologically aligned content while distinguishing echo chambers from other communities, as validated through experiments on synthetic datasets and real-world cases, including the Brexit

referendum and COVID-19 vaccine discussions.

In their study on echo chamber detection during the early COVID-19 pandemic, Villa et al. [6] propose a method that integrates both topological and semantic aspects of online interactions. The process begins by modeling a Twitter conversation graph, where nodes represent users and edges are established through mention relationships, with edge weights reflecting the frequency of mentions. The authors enrich the graph with semantic information by adjusting edge weights according to users' sentiment similarity (using VADER sentiment analysis) and topic similarity (using LDA topic modeling), producing four distinct graph representations: topology-based, sentiment-based, topic-based, and a hybrid of sentiment and topic. To detect echo chambers, they apply the METIS community detection algorithm to partition each graph into two groups, corresponding to potential polarized communities. The presence of echo chambers is then assessed by quantifying controversy (using random walk-based and boundary connectivity measures) and community homogeneity (analyzing sentiment and topical coherence within each group). Their results show that incorporating content-based features, especially sentiment, enhances the detection of polarized, homogeneous groups—providing a robust framework for echo chamber identification in social media contexts.

Although graph partitioning techniques have been applied for community detection, they often produce overly balanced divisions that may fail to reflect the intricate structures found in real-world datasets. To address this limitation, [97] presents a novel generalized balanced subgraph model that allows for some degree of imbalance. They propose a region-based heuristic algorithm that effectively balances computational efficiency with solution quality. This approach is built on signed networks, which capture the consistency of opinions among users and, as a result, incorporate semantic information. Similarly, building on signed networks, [98] detect polarized communities characterized by mostly intra-community positive edges and inter-community negative edges, thus enabling fine-grained analysis of controversy in social networks.

3 Echo Chamber Effect Measurement

In this section, we collate and analyze the various metrics used to measure echo chamber effects. Typically, there is no single metric that directly quantifies whether a reported echo chamber is inherently "good" or "bad". Instead, echo chambers are typically assessed through their observable consequences, such as polarization, network segregation, and intra-group homogeneity. By evaluating these phenomena, researchers can gauge the extent to which echo chamber effects are present. In the literature, similar aspects of echo chamber effects are often quantified using different terms. For example, the degree to which communities within a social network are separated—and users predominantly interact with others within their own community—may be referred to as "segregation", "separateness", "controversy", or "polarization". Although these terms may emphasize slightly different nuances, they generally capture the same underlying intuition and tend to be used interchangeably.

Existing approaches to measuring echo chamber effects can be broadly categorized into three groups: (1) *Network Segregation-Based Metrics*, which analyze the structural properties of social networks to capture the characteristic that intra-community interactions are much more frequent than inter-community ones; (2) *Opinion Homogeneity-Based Metrics*, which assess the distribution of user opinions to determine the degree of opinion alignment within echo chambers, or to evaluate overall opinion diversity across the network; and (3) *Hybrid Metrics*, which integrate both network topology and user opinions to assess intra-community alignment alongside cross-community segregation. Additionally, some studies have explored echo chamber effects in non-social network contexts, such as e-commerce platforms, and have proposed corresponding metrics. In the following subsections, we classify and introduce these approaches in more detail.

3.1 Network Segregation-Based

Intuitively, a higher number of intra-community edges and fewer inter-community edges indicate a greater degree of segregation between groups of users, which structurally aligns with the concept of an echo chamber. Consequently, many approaches focus on the structural characteristics of the network by dividing it into communities, typically using community detection or graph partition methods that aim to minimize the number of edges between different communities. Based on the resulting community structure, various metrics are then employed to quantify the degree of segregation in the network.

Analyzing the challenges in accurately determining users' political preferences and the expected properties of informational separateness, Chkhartishvili et al. [2] introduced a metric to measure the echo chamber effect, called the **Binary Separation Index (BSI)**. This index requires only the identification of the set of accounts disseminating information within the network and their corresponding political positions. The BSI is calculated as follows:

$$\text{BSI} = 4 \times \alpha \times \beta$$

where α and β represent the proportions of users in each of the opposing sources I_1 and I_2, respectively, who exclusively connect with information from their own source and not the opposing one. Note that $\alpha, \beta \in [0, 1]$. The BSI attains its maximum value, 4, when the network is perfectly divided, with users exclusively engaging with only one source or the other, indicating strong informational separation and a pronounced echo chamber effect. Conversely, a lower BSI suggests greater cross-exposure between groups, reflecting a less segregated information environment.

With a focus on nodes, BSI utilizes the ratio of border users in each community. In contrast, edge-focused methods examine the connections between communities from a different perspective. Luo et al. [3] introduce a quantitative measure of segregation in social networks by examining the formation of edges both within and between communities. Specifically, they define **segregation** for a directed graph $G = (V, E)$, where the set of vertices $V = R \cup B$ is partitioned into red (R) and blue (B) communities. The segregation metric s is given by

$$s = 1 - \frac{|E_d|}{2|R||B|}$$

where E_d denotes the set of edges connecting members of different communities, and $2|R||B|$ represents the maximum possible number of inter-community edges. A segregation value of $s = 1$ corresponds to complete segregation, while lower values indicate greater integration between communities.

A related metric, the **E-I Index** [68], is used by Ertan et al. [32] to measure political polarization in network structures. After defining political blocs (e.g., based on formal electoral alliances), each respondent's network receives an E-I Index score, calculated as:

$$\text{E-I Index} = \frac{ET - IT}{ET + IT}$$

where ET is the number of external ties (cooperation across blocs) and IT is the number of internal ties (cooperation within blocs). The index ranges from -1 (homophily, mostly within-bloc ties) to $+1$ (heterophily, mostly cross-bloc ties). The key distinction between the E-I Index and the BSI lies in their denominators: the BSI normalizes by the maximum possible number of edges, whereas the E-I Index uses the actual number of observed edges.

As many works use community detection methods to find echo chambers, **modularity** proposed in [17] is widely used in works such as [6, 33, 34] to evaluate the quality of a discovered community structure. Modularity measures how well a network is partitioned into communities by comparing the density of links within communities to that expected in a random network. A higher modularity score

indicates a stronger community structure. Intuitively, modularity assesses whether more edges fall within communities than would be expected by chance. It does this by comparing the actual fraction of edges within each community to the expected fraction if the edges were distributed randomly, while preserving the network's degree distribution. If the observed number of intra-community connections is significantly higher than expected at random, the modularity score will be high. A value of modularity close to its theoretical maximum of 1 indicates a strong community structure, while a value near 0 suggests that the observed partition is no better than random.

However, Guerra et al. [7] note that "non-polarized networks still show positive modularity", limiting modularity's effectiveness for detecting polarization. To address this, they introduce **Boundary Connectivity (BC)**, P, a metric that measures antagonism by evaluating boundary nodes' connectivity preferences between two communities:

$$P = \frac{1}{|B|} \sum_{v \in B} \left[\frac{d_i(v)}{d_b(v) + d_i(v)} - 0.5 \right]$$

where $d_i(v)$ is the number of edges from node v to internal nodes, $d_b(v)$ is the number of edges to opposing boundary nodes, and B is the set of boundary nodes. A positive P indicates polarization, while negative or near-zero values suggest its absence. BC is applied in many studies [6, 16, 30, 35, 36, 60] to quantify controversy or polarization.

Guyot et al. [40] introduce **ERIS**, a metric akin to BC but incorporating edge weights, to assess polarization by examining community boundaries. For each pair of communities C_i and C_j, ERIS identifies the boundary area ($B_{i,j}$), defined as the set of users in C_i who interact both with their own community and with C_j. The first metric, Community Antagonism ($A_{i,j}$), quantifies the directed opposition from C_i to C_j based on the interaction patterns of boundary users. For an individual boundary user v, antagonism $A^v_{i,j}$ is computed as:

$$A^v_{i,j} = \frac{\sum_{e \in IE^v_{i,j}} w(e)}{\sum_{e \in IE^v_{i,j}} w(e) + \sum_{e \in EE^v_{i,j}} w(e)} - 0.5$$

where $IE^v_{i,j}$ and $EE^v_{i,j}$ denote the sets of v's internal and external edges, respectively, and edge weight $w(e)$ is the number of times u quoted v for edge $e = (u, v)$. The overall community antagonism $A_{i,j}$ is obtained by averaging $A^v_{i,j}$ over all users in $B_{i,j}$. The second metric, Boundary Porosity ($P_{i,j}$), measures the permeability of the boundary, defined as the proportion of boundary users who interact more with the external community:

$$P_{i,j} = \frac{\left| \{ v \in B_{i,j} \mid A^v_{i,j} < 0 \} \right|}{|B_{i,j}|} \times 100.$$

Together, these metrics capture both the degree of antagonism between communities and their susceptibility to external influence.

While many existing metrics primarily focus on edge counts, Garimella et al. [16] introduced a random walk-based metric, the **Random Walk Controversy (RWC)** score, to quantify the structural separation in partitioned conversation graphs. The underlying intuition is that, in highly controversial topics, opposing sides form distinct communities with minimal interaction between them. As a result, a random walk initiated within one community is more likely to remain there than traverse to the other. The RWC score formalizes this notion. Given a graph partitioned into two disjoint user sets, X and Y, the metric is defined as the difference between the probability that two random walks remain within their respective starting partitions and the probability that both cross over to the opposing partition.

Specifically, let P_{AB} denote the conditional probability that a random walk starting from a random node in partition A terminates in partition B. The RWC score is then calculated as:

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

Here, the term $P_{XX}P_{YY}$ represents the likelihood that both communities are internally cohesive and isolated, while $P_{XY}P_{YX}$ reflects the strength of cross-partition interaction. A score approaching 1 indicates strong separation between partitions and thus a high level of controversy. Conversely, a score near 0 suggests that the probability of crossing partitions is similar to that of remaining within them, implying a lack of clear division and therefore a non-controversial topic. For practical computation, the authors propose an efficient variant utilizing Random Walk with Restart (RWR) to estimate these probabilities in large, directed graphs. In [6], the authors employ the RWC score to measure controversy and further introduce variants such as Authoritative Random Walk Controversy (AWRC) and Displacement Random Walk Controversy (DRWC). The RWC score has also been applied in [22, 51, 52].

The aforementioned metrics evaluate segregation at the level of the entire network. To determine the degree of homogeneity within individual partitions, "**coverage**" is employed in [6], as originally introduced in [18]. Coverage quantifies the proportion of a graph's total edges that are intra-community, i.e., those connecting pairs of vertices within the same community. In the idealized case where communities are completely separated—forming disjoint subgraphs with no inter-community edges—the coverage reaches a maximum value of 1. Consequently, coverage offers a direct measure of partition cohesiveness by indicating how well the detected communities encompass the network's edge structure. Additionally, in [8, 13], "**conductance**" is utilized, which measures the fraction of the total edge volume that leaves the community. Intuitively, conductance reflects how well a community is separated from the rest of the network: it compares the number of edges that connect the community to the outside to the total number of edges associated with the community, both internal and external. To ensure that a community is sufficiently insular to be considered an echo chamber, conductance is constrained to be at most 0.5 in [8, 13], meaning that more than half of the total edges must remain within the community boundaries.

3.2 Opinion Homogeneity-Based

Echo chambers cluster like-minded users who predominantly interact with others sharing similar views. A straightforward approach to quantifying echo chambers is to measure how closely aligned users' leanings are within the same echo chamber, and how distinct the average leanings are between different echo chambers. Additionally, by examining the overall distribution of user leanings, one can determine whether there is significant polarization—for example, whether the distribution exhibits a "U"-shaped pattern indicative of extreme polarization, or is more uniform. From this perspective, many studies have investigated polarization and the echo chamber effect through the lens of user opinions.

A common approach for quantifying polarization is to focus on opinion homogeneity within a network. Several studies [26, 27, 64, 67] use the variance of expressed opinions as a direct indicator of **polarization**, which measures how much individual opinions deviate from the average opinion in the group. A higher variance indicates greater diversity or polarization in user opinions, while a lower variance suggests more consensus.

Chen et al. [43] define **controversy** as the summation of squared user opinions that captures the overall intensity of expressed opinions in a group by aggregating how strongly each opinion deviates from neutrality. The same formula is adopted in Musco et al. [50] under the term **polarization**, and Matakos et al. [63] propose a "**polarization index**" computed as the average squared opinion per user, i.e., dividing the sum above by the total number of users.

Moving beyond aggregate measures, Interian et al. [29] introduce a probabilistic framework to quantify network polarization by evaluating the statistical significance of node-level homophily. Instead

of relying solely on raw homophily counts, their method calculates a p-value for each node. This p-value represents the likelihood that a node's observed number of same-group connections (or more) could occur by chance in a "balanced" network, where connection probabilities are proportional to group sizes. The overall polarization for a network or group is then characterized by the distribution of these p-values: a distribution skewed towards zero indicates strong, statistically significant polarization. Empirical cumulative distribution functions of these p-values are used for robust comparison of polarization across groups or networks.

Cota et al. [9] take a different approach, quantifying the echo chamber effect by assigning each user a political position based on the average leaning of their tweets, with tweet classification performed manually as pro-impeachment, neutral, or anti-impeachment. Echo chambers are then measured by analyzing the correlation between a user's political position and the average leanings of both their nearest neighbors and the tweets they receive. Strong correlations indicate that users primarily interact with others holding similar political views, confirming the existence of echo chambers. This **joint distribution** approach is also utilized in [53, 54, 58, 61, 62]. Notably, this approach is primarily visualization-based and does not provide a single summary statistic. For more quantitative assessment, the correlation coefficient between the joint distribution and the $y = x$ line can be calculated to measure the degree of homophily between users' opinions and the average opinions of their neighbors.

Further exploring the dynamics of opinion formation, Sikder et al. [19] model agents in a social network tasked with determining the truth of a binary statement ($X = +1$ or $X = -1$). Opinion formation begins at $t = 0$ with each agent receiving a private signal. At each discrete time step, agents synchronously share all accumulated signals with their immediate neighbors, causing each agent's information set to grow recursively. Polarization is quantified in several steps: each agent i computes their "signal mix" $x_i(t)$ at time t, representing the proportion of positive signals ($N_i^+(t)$) among all signals received:

$$x_i(t) = \frac{N_i^+(t)}{N_i^+(t) + N_i^-(t)}.$$

Based on $x_i(t)$, the agent adopts a public orientation $y_i(t)$, set to $+1$ if $x_i(t) > 0.5$ and -1 otherwise. For a group C, **polarization** $z_C(t)$ is defined as the size of the minority camp:

$$z_C(t) = \min(y_C(t), 1 - y_C(t)),$$

where $y_C(t)$ is the fraction of positive orientations in C. Polarization reaches zero under full consensus and is maximized at 0.5 when the group is evenly split.

At the individual level, Al Atiqi et al. [10] introduce the **Individual Echo Chamber Coefficient (ECC)**, which measures the diversity of opinions among a user's neighbors. Instead of a formula, the ECC for a user quantifies how varied the opinions are within that user's network; if all neighbors share similar views, the ECC will be low, whereas a wide range of opinions among neighbors leads to a higher ECC. Values near zero indicate strong echo chambers. For network-wide assessment, they define the **Global Echo Chamber (GEC)** Indicator, which reflects the overall tendency of connected individuals to share the same or opposing opinions throughout the network. A lower GEC value suggests weaker clustering of like-minded individuals.

Botte et al. [11] propose two complementary metrics for echo chambers. The global **echo chamber size** measures the relative growth of fully homogeneous neighborhoods: instead of focusing on the exact calculation, this metric essentially compares how the number of individuals who are exclusively surrounded by others with the same opinion changes from the beginning to the end of the observation. They also assess **local polarization** by analyzing the distribution, the fraction of an individual's neighbors sharing their opinion. A bimodal distribution of these values, with peaks near 0 and 1, signals strong polarization into segregated groups.

Similarly, Madsen et al. [12] introduce the "**belief purity**" metric to quantify echo chamber formation. The intuition is that this metric reflects how similar the beliefs are among connected agents: as the average difference in belief between connected nodes decreases, the belief purity approaches its maximum value of 1. This indicates that, in highly purified networks, connected agents tend to share nearly identical beliefs.

The concept of "**purity**" is also employed in [8, 13]. In [8], purity is the ratio of users with the same ideological alignment, measured as the average leaning of reshared tweets. In [13], purity is defined as the product of the frequencies of the most common labels among nodes in a community. Combined with "conductance" (as discussed above), low conductance and high purity are identified as hallmarks of echo chambers.

To capture the distributional aspects of polarization, Lelkes [30] employs specific metrics. For ideological divergence between partisans, the **Overlap Coefficient (OC)** is used, which intuitively measures the extent to which the ideological distributions of two groups (such as Republicans and Democrats) coincide. An OC value of 1 indicates perfect overlap, while 0 indicates complete separation. For detecting polarization in the general public, the Bimodality Coefficient is utilized to formally test for the emergence of two peaks in the ideological distribution, thereby providing an alternative to mean comparisons.

To quantify "echofication" in multiparty networks, Markgraf and Schoch [31] present a two-pronged framework. First, they identify "Social Boundaries" (the chamber) via community detection and measure the insularity of these boundaries. Second, they assess "User Similarity" (the echo) by modeling user ideology as a multi-dimensional vector, based on the politicians each user follows. Cosine similarity between users provides a homophily score, serving as a direct proxy for the level of echo within each chamber.

3.3 Hybrid

Network topology-based methods can quantify the degree of segregation between different communities; however, they are unable to assess individuals' ideological alignment within each community or the ideological differences between communities, as analyzing interactions alone may not accurately reflect users' true opinions without analyzing their posts or other outputs. Although it is challenging to obtain users' genuine thoughts, opinion-based methods can evaluate whether the distribution of users' leanings as expressed in their posts aligns with the expected characteristics of echo chambers. However, these methods often overlook the assessment of users' interaction patterns (i.e., frequent intra-community interactions and infrequent inter-community interactions), particularly when the metrics quantify polarization only at the network level. For example, a variance score or leaning distribution calculated for the entire network does not contain information about the position of each node. Compared with two ideologically opposing groups that have high intra-group homogeneity, a network composed solely of users with extreme and polarized viewpoints, where these users frequently interact with both like-minded and opposing individuals, can still exhibit a high variance score or highly polarized leaning distribution. Therefore, in this section, we discuss measurement approaches that combine both the network's topological structure and users' opinions simultaneously. Such hybrid approaches are promising for providing a more comprehensive and precise quantification of the echo chamber effect.

To quantify the structural dimension of polarization beyond just opinion distribution, [20] proposed a metric that assesses the alignment between opinion similarity and the strength of connections within the network. Building upon the basic concept of edge homogeneity (the product of two connected users' opinions), they introduce the **average weighted mean edge homogeneity** ($\overline{\text{hom}}_w$). This measure incorporates the evolving connection strength between users, providing a more nuanced view of the network's structure. Instead of simply averaging opinion similarity across all connections, this measure

gives more weight to stronger connections, effectively capturing how strongly people with similar or different opinions are connected in the network. A high value of $\overline{\text{hom}}_w$ signifies a strongly polarized and segregated structure, characteristic of an echo chamber, where users with similar opinions are linked by strong connections and those with different views are linked by weak ones.

Several metrics explicitly integrate network topology with the distribution of opinions to quantify polarization and fragmentation. Shekatkar [48] proposes "**correlated polarization**" (ϕ), which is defined as the product of two components: (1) the balance or bimodality of opinions, $R = 1 - 2|n^- - 0.5|$, where n^- is the fraction of nodes holding one of the two opinions (maximized when opinions are evenly split), and (2) the **assortativity coefficient** r [49] with respect to node states, measuring the tendency of like-minded nodes to connect. Thus, $\phi = R \times r$ achieves high values only in networks exhibiting both a sharp division of opinions and pronounced structural segregation, capturing social fragmentation more effectively than methods based solely on opinion counts.

Morales et al. [52] introduce a **polarization index**, μ, to measure the degree to which an opinion distribution is divided into two distinct and opposing groups. Conceptually inspired by the electric dipole moment in physics, their metric defines perfect polarization as a state where a population is split into two factions of equal size holding maximally distant opinions. The index is calculated as the product of two key components: group size balance and ideological distance. The first component, $(1 - \Delta A)$, captures the population balance, where ΔA is the absolute difference between the relative sizes of the two opposing groups. This term equals 1 when the groups are perfectly balanced in size and 0 when one group comprises the entire population. The second component, d, represents the ideological distance, calculated as the normalized distance between the "gravity centers" (i.e., the mean opinions) of each group. This distance ranges from 0, for no ideological separation, to 1, for maximum opposition. The final polarization index, $\mu = (1 - \Delta A)d$, thus reaches its maximum value of 1 only when both conditions are met: the two groups are of equal size ($\Delta A = 0$) and their average opinions are maximally far apart ($d = 1$).

To address the limitations of approaches focused solely on network structure, Emamgholizadeh et al. [37] propose the **Biased Random Walk (BRW)** framework for quantifying controversy in attributed networks. Unlike methods that only consider topology, BRW integrates both structural and node-level attributes. The core of the framework is a random walk with a finite lifetime, simulated through a novel energy mechanism. The initial energy for a walk starting at a given node is determined by considering how close that node is to the center of its own community as well as how close it is to the center of the opposing community, reflecting both local and cross-community influence. As the walk traverses the network, it loses energy at each step, with the amount of energy lost depending on how far the current node is from the center of the opposing community. This means that the deeper a walk ventures into the opposing community, the faster it loses energy, making it increasingly difficult to reach the core of the opposition. Controversy is then measured by the "penetration depth"—the maximum level a walk can reach within a contradicting community before its energy is depleted. This approach models an idea's ability to be heard by an opposing audience, offering a more nuanced controversy score that captures the combined effects of network structure and user characteristics.

Alatawi et al. [51] propose the **Echo Chamber Score (ECS)**, a metric that quantifies the echo chamber effect by evaluating the geometric properties of user communities within a learned embedding space. Their approach first uses a self-supervised Graph Auto-Encoder, EchoGAE, to embed users into a low-dimensional space where distance corresponds to ideological similarity, leveraging both user interaction patterns and post content. The core idea behind ECS is to measure two key properties: cohesion, representing how similar users are within their own community, and separation, representing how distinct they are from users in other communities. For each user u in a community ω, cohesion is calculated as their average distance to all other users in ω. Separation is their average distance to users in the nearest neighboring community. These values are then aggregated into a score for the individual

community by assessing, for each member, the relative difference between their similarity to their own community and their dissimilarity to the closest external community. This aggregation is inspired by the silhouette score and normalized so that it ranges from 0 to 1, where higher values indicate stronger echo chamber effects. The overall ECS for the entire graph is then the average of these scores across all detected communities. A score approaching 1 signifies a strong echo chamber effect, with tightly clustered and well-separated communities, while a score near 0 indicates more ideologically integrated groups. A key advantage of this method is its unsupervised nature, as it requires neither pre-defined user labels nor a fixed number of communities.

Similarly, Amendola et al. [60] move beyond structural analysis to incorporate the semantic content of user interactions. Their approach is founded on the principle that true opinion alignment requires agreement on specific facets of a topic, not just a general sentiment. The method employs **Aspect-Based Sentiment Analysis (ABSA)** to capture nuanced opinions and **Group Decision-Making (GDM)** principles to measure consensus. The process begins by using ABSA to generate a sentiment vector for each user on every specific aspect of a topic. The core of the metric is built upon a multi-level aggregation of these opinions. First, the agreement between any two users on a specific aspect is calculated as a pairwise similarity, typically using cosine similarity. These pairwise similarities for a single aspect are then aggregated across all users to determine the **Aspect Consensus**, which intuitively reflects the average level of agreement within the community on that aspect, with greater weight given to stronger agreements. Finally, the consensus scores from all aspects are aggregated to produce a single **Community Consensus** value, representing the overall agreement within the community by summarizing the aspect-level agreements in a way that emphasizes more consistent patterns of alignment. The diagnostic power of this metric comes from comparing the within-community consensus (calculated among members of the same group) with the in-between-communities consensus (calculated between members of different groups). A high within-community consensus coupled with a low in-between-communities consensus provides a strong, content-driven signal of an echo chamber.

Hohmann et al. [65] introduce a holistic measure for ideological polarization that integrates three distinct factors into a single score: the extremity of opinions, the structural clustering of individuals into communities (homophily), and the mesoscale organization of these communities along an ideological spectrum. Their approach is based on the **generalized Euclidean (GE) distance**, which quantifies the "effort" required for influence to travel between opposing sides of a debate on a given network. For a graph G and an associated opinion vector o (a list of users' opinions, where opinions are normalized between -1 and $+1$), the polarization measure, denoted $\delta_{G,o}$, is formulated as:

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

In this equation, o^+ contains the positive opinions (and zeros elsewhere, i.e. $o_i^+ = \max(o_i, 0), \forall i$), o^- contains the absolute value of negative opinions, and L^+ is the Moore-Penrose pseudo-inverse of the graph's Laplacian matrix. The term L^+ embeds the network's topology, weighting the distance by the paths (or lack thereof) connecting individuals. Conceptually, the measure represents the network distance between the "centers of mass" of the opposing opinion groups. The authors demonstrate through synthetic and real-world data that this metric is uniquely sensitive to all three aspects of polarization, unlike measures that focus only on opinion distributions or local network assortativity. This measurement is further extended to multipolari polarization in [66].

Huang et al. [39] propose a partition-agnostic polarization measure based on the correlation between signed and unsigned random-walk dynamics. They define a signed random-walk transition matrix $\mathbf{M}(t)$, where matrix entries encode both topological distances and the signs of paths, and contrast this with its unsigned counterpart $|\mathbf{M}(t)|$. The polarization score for a node u is computed as the Pearson correlation between its signed and unsigned transition vectors: $\mathrm{Pol}(u;t) = \mathrm{corr}(|\mathbf{M}|_{:u}(t), \mathbf{M}_{:u}(t))$. Averaging these

node-level scores gives the overall graph polarization. This methodology captures polarization at multiple scales (controlled by the time parameter t), and crucially, it does not require pre-defined community partitions.

3.4 Others

Amelkin et al. introduce **Social Network Distance (SND)** [89] to quantify the evolution of user opinions in social networks, taking into account both the network structure and the dynamics of competing polar views. SND frames opinion change as a transportation problem, allowing it to capture not just individual shifts, but also the collective patterns of opinion propagation across the network. This approach is effective for detecting events that trigger or intensify polarization within society. For example, when applied to real-world X (formerly Twitter) data, SND is able to identify spikes in opinion divergence around major political events, such as elections or controversial policy debates. These spikes signal moments when public sentiment becomes sharply divided, enabling researchers to pinpoint the timing and nature of polarization-triggering events.

Bozdag et al. [38] present a comprehensive framework for empirically measuring the manifestation of offline political segregation in online environments. They operationalize information diversity using Shannon entropy, calculating separate scores for a user's information input and output. Specifically, "source diversity" quantifies the political heterogeneity of tweets a user receives, while "output diversity" measures the diversity of political content a user disseminates. The disparity between these scores, together with an "input-output correlation" metric, serves as an indicator of the filter bubble effect. Importantly, their framework goes beyond conventional diversity metrics by capturing structural exclusion of minority voices. They introduce the concept of "openness" measured by two indicators: "minority reach" which reflects the network-wide penetration of minority viewpoints, and "minority exposure" representing the proportion of minority content in an individual's feed. This approach demonstrates that even when overall source diversity appears sufficient, significant segregation can persist through the marginalization of minority actors.

As discussed in the echo chamber detection section, Calderón et al. [4] introduce the **Echo Chamber Index (ECI)**, which quantifies echo chamber behavior on a given fan page by averaging the echoing scores of comments under a post. This provides a straightforward yet effective numerical measure of echo chamber dynamics at the post level.

In [90], the measurement of the echo chamber effect is formalized within the influence maximization framework through the **Influence Maximization with Echo Chamber(IMEC) problem**. The authors model echo chamber influence as an additional probabilistic mechanism: a user's likelihood of adopting information is increased if many members of their group have already adopted it. This group-level effect is mathematically represented using an Ising-model-inspired function, where the probability of group-based activation depends on the number of activated users and a group closeness parameter. By comparing information diffusion outcomes with and without this global group influence, the paper directly quantifies the impact of echo chambers on the overall spread.

Although not directly focused on social networks, Ge et al. [15] develop a quantitative framework to measure interest reinforcement in the context of e-commerce echo chambers. Their approach segments users' interaction histories into temporal blocks, representing each user's interests within a block as an aggregate embedding of the items they engaged with. By comparing a "Following Group"—users who frequently interact with recommendations—to a control "Ignoring Group", they examine changes in the structure of user interest clusters over time. The reinforcement effect is assessed using cluster validity indices: a smaller decrease in the Calinski-Harabasz (CHK) score indicates that clusters remain more compact, while a higher Adjusted Rand Index (ARI) suggests less user drift between clusters. Their results demonstrate that this method effectively quantifies the reinforcement phenomenon central to

echo chambers.

Finally, there exist several works that design polarization or controversy metrics as optimization objectives for their mitigation strategies. These will be introduced in a subsequent section dedicated to mitigation approaches. It is important to note that there is no single universally accepted metric for measuring echo chambers; different metrics capture different aspects of echo chambers, such as topological structure, opinion distribution, or a combination of both and cater to different conditions. For example, if we can only obtain people's post content and no interaction information among users due to the privacy or terms of policies of the social media platform, we can embed people's generated content and infer their leanings with ML, then simply use the variance of leanings as an indicator of polarization or the echo chamber effect. On the other hand, if only people's interactions are available and we cannot extract their post content, topology-based metrics could be employed. When only internal and external edges are of concern or are available, the E-I index is one of the choices. If community size is also a concern (as opposing communities with similar sizes are generally considered more polarized compared to one very large and one very small community), BSI could take this into consideration.

Typically, these metrics do not quantify echo chambers directly; rather, they measure the effects associated with echo chambers. For example, they assess which network aligns more closely with the intuition or definition of an echo chamber, or better matches expected properties.

4 Echo Chamber Mitigation

Although this section focuses on echo chamber mitigation, we also include research on reducing related phenomena such as polarization, segregation, and controversy, as these efforts often aim to counter or mitigate echo chamber effects. For example, polarization is both a consequence of echo chambers and a factor that accelerates their formation [5]; thus, reducing polarization contributes to echo chamber mitigation. Since most studies do not explicitly distinguish between these terms, in this survey, we consider all such efforts to fall within the scope of echo chamber mitigation.

Existing social network-based approaches to mitigate echo chamber effects can be broadly categorized into three groups: (1) *Cross-group Promotion Approaches*, which involves adding edges between opposing groups or recommending posts from users with differing viewpoints to increase exposure to alternative perspectives; (2) *Opinion Dynamics-Based Approaches*, which simulate the evolution of user opinions through opinion dynamics models and seek to optimize a polarization metric that depends on the final opinions and/or the network topology structure, by modifying the topology, edge weights, or users' innate opinions (though the latter is less practical); and (3) *Agent Addition-Based Approaches*, which introduce artificial agents that strategically disseminate information with targeted leanings or ideologies to influence other users or affect recommender systems.

4.1 Cross-group Promotion

A common strategy for mitigating polarization is to bridge opposing groups and expose individuals to diverse perspectives. Many studies have explored methods such as adding cross-group connections or recommending contents with opposing viewpoints. By increasing the diversity of information people encounter, these approaches aim to prevent echo chambers and encourage more moderate ideologies.

Luo et al. [3] introduce a game-theoretic framework to mitigate segregation and echo chambers in social networks by incentivizing inter-community connections. They model user interactions as an edge formation game, where individuals trade off homophily (preference for same-group ties) against exogenous rewards for cross-community links. Their **Algorithmic Recommendation Mechanism (ARM)** leverages weak ties to encourage diverse connections, reshaping the Nash equilibrium from segregated networks to integrated ones. Simulations demonstrate ARM's efficacy in reducing segregation,

particularly during polarizing events, offering a scalable mechanism design solution to counteract echo chambers.

Several studies have approached the problem by identifying and algorithmically bridging structural divides in social networks. Garimella et al. [42] propose an algorithmic approach to reduce polarization by strategically bridging opposing communities in an online discussion. They begin by modeling a controversial topic as a directed "endorsement graph" (e.g., a retweet network), which is then partitioned into two disjoint communities representing the opposing sides. The authors' objective is to reduce a specific quantitative metric, the Random-Walk Controversy (RWC) score, which measures the isolation of these communities by calculating the probability that a random walker remains within its starting community. To achieve this, their method recommends a small set of new edges ("bridges") to be added between the two sides. Acknowledging that not all recommendations are equally likely to be accepted in reality, they introduce a model for "acceptance probability" based on user polarity, shifting the goal to minimizing the expected controversy score. They propose an efficient heuristic algorithm (**ROV-AP**) that identifies the most effective bridges by primarily considering connections between high-degree users on opposite sides, thus offering a practical method for algorithmically mitigating online echo chambers.

Interian, Moreno, and Ribeiro [44] solve the problem of reducing network polarization by formulating it as an Integer Linear Program (ILP). Their method, which addresses the **Minimum-Cardinality Balanced Edge Addition Problem (MinCBEAP)**, is designed to find the smallest set of new edges, E', to add to a graph $G = (V, E)$. The ILP model uses binary variables to represent the selection of new edges and to track the shortest path distances. The core of the solution method is an optimization that minimizes the number of added edges while enforcing a key structural constraint: ensuring that every vertex v in a target polarized group A can reach a vertex outside of its group ($V \setminus A$) within a specified distance threshold D. The ILP formulation directly solves the following problem:

$$\begin{aligned} \min \quad & |E'| \\ \text{subject to} \quad & d_{G'}(v, V \setminus A) \leq D, \quad \forall v \in A \\ \text{where} \quad & G' = (V, E \cup E') \end{aligned}$$

By solving this ILP model with standard optimization software, they compute the minimal set of edges required to structurally bridge the isolated groups in a network. Their subsequent work [45] further compares three ILP formulations and reports computational results on both simulated and real-world networks.

In [46], Haddadan et al. propose a method to reduce structural polarization in content networks by adding a budget of k links. They introduce the (Polarized) Bubble Radius (BR), which measures the expected random walk steps from a node v to a page with an opposing view. Nodes with a high BR are termed "parochial", and the network's overall structural bias is the sum of their BRs. The problem of minimizing the bias is framed as a submodular maximization problem, allowing for an efficient greedy algorithm. The **REPBUBLIK** algorithm iteratively adds links from source nodes that are selected based on a task-specific variant of Random-Walk Closeness Centrality (RWCC), thereby strategically creating shortcuts that have the broadest impact on reducing network-wide polarization.

Recent studies have also examined the application of the influence propagation paradigm within social networks to balance information exposure [91–94]. Garimella et al. [91] introduce an approach to mitigate filter bubbles and echo chambers in networks by focusing on balancing information exposure across opposing viewpoints. Their framework adopts a centralized perspective, seeking to maximize the number of users exposed to both sides of a controversial issue. The authors model the spread of information using the independent-cascade model and formalize the objective as maximizing the number of users who are either reached by both campaigns or by neither, thereby directly addressing the problem of information imbalance. They demonstrate that this balancing problem is NP-hard, and, crucially, that its objective

function is neither monotone nor submodular, making it difficult to find efficient approximation solutions. To address this, they propose and analyze several greedy algorithms with approximation guarantees, and validate their effectiveness through experiments on real-world Twitter data spanning political and social controversies. With a similar idea, Tu et al. [93] formally define **co-exposure maximization (COEM)** as the task of selecting initial user sets for each campaign in order to maximize the expected number of users who receive information from both campaigns, considering the probabilistic nature of information spread in the network. They propose a greedy approximation algorithm that uses a submodular lower bound for the co-exposure objective, offering theoretical guarantees. Additionally, they introduce a scalable estimation method based on generalized random reverse-reachable sets, which enables efficient computation of expected co-exposure in large networks. Matakos et al. [92] also address filter bubble mitigation by maximizing the diversity of information exposure in social networks. However, their approach models both user and content leanings, and strategically recommends news articles to select users so that, as articles propagate, users are exposed to a wider range of viewpoints. The problem is formulated as a submodular optimization under matroid constraints, and solved efficiently using a novel sampling technique called reverse co-exposure sets. Considering the ignorance of the competition between opposing opinions propagating in previous studies, Banerjee et al. [94] address filter bubble mitigation by modeling the realistic competition between opposing viewpoints as they spread in a social network. They propose the RIC-FB model, which distinguishes between awareness and adoption of opinions, and incorporates a competition parameter that makes it harder for users to adopt a second, opposing viewpoint after adopting the first. Their approach rewards co-adoption (users adopting both viewpoints), thus directly targeting filter bubble reduction rather than just exposure balancing. They formulate the mitigation task as an optimization problem and prove its computational hardness, then introduce specialized algorithms—including a reverse-influence-sampling-based heuristic—to effectively select seeds for the counter-campaign. Experiments on real networks show their competition-aware methods outperform existing baselines, especially when competition is strong.

Beyond structural interventions, Orbach et al. [28] address echo chambers by introducing a novel **Natural Language Understanding (NLU)** task for detecting countering speeches. Rather than merely identifying content with opposing stances, their approach retrieves texts that directly refute the specific arguments presented in an input document. By automatically surfacing targeted rebuttals, their method aims to expose users to diverse viewpoints and foster a more balanced and informed perspective.

Although many studies have explored methods for promoting cross-group edges and have found that inter-group contact can foster compromise and mutual understanding in some contexts, "confirmation bias" [73] often impedes effective discourse and connection—especially especially when users behave strategically for profit or other motives. Bail et al. [72] also point out that attempts to expose people to a broad range of opposing political views on social media platforms like Twitter may not only be ineffective, but even counterproductive. Other studies indicate that exposure to opposing political views may trigger "backfire effects"[69], which can intensify political polarization[69–71]. It has also been discussed in [41] that the effectiveness of this approach may depend on the specific network context and the existing degree of polarization. Exposure to opposing views may reduce polarization during the initial or intermediate phases of polarization, but it is less effective once polarization is already strong. Bail et al. suggest that "future attempts to reduce political polarization on social media will most likely require learning which types of messages, tactics, or issue positions are most likely to create backfire effects." Depolarizing users by exposing them to viewpoints only slightly less radical than their own may be more effective [21, 75].

4.2 Opinion Dynamics-Based

There is a class of works focusing on simulating opinion dynamics with existing or designed models and optimizing polarization or similar metrics, which are computed using users' final opinion values after the opinion dynamics process. These studies achieve their goals by changing the network topology (adding or deleting edges), modifying edge weights, adjusting parameters in the opinion dynamics model (e.g., innate opinions or weights in the Friedkin-Johnsen model [74] (FJ model)), or by employing other strategies.

A foundational direction in this area is to optimize the structure or weights of the network to mitigate polarization. Musco et al. [50] formalize the problem of finding a network structure that minimizes both polarization and disagreement. They propose a **Polarization-Disagreement Index** based on the FJ model, which sums two terms: polarization (the variance of final opinions) and disagreement (the opinion differences across edges). Crucially, the authors prove this index is a convex function of the network's edge weights, which allows the optimal graph topology to be computed efficiently. Their analysis also shows that this optimal network can be well-approximated by a sparse graph with only $O(n/\epsilon^2)$ edges.

Building on the FJ model, several works propose new metrics or optimization methods to further enhance network robustness. Chen et al. [43] present a structural approach to conflict mitigation that focuses on minimizing the risk of conflict rather than the conflict itself for a single, known issue. Departing from methods that require specific opinion data, they propose network-level metrics that are independent of any particular opinion distribution. They define the **Average-Case Conflict Risk (ACR)** and **Worst-Case Conflict Risk (WCR)**, which quantify a network's inherent propensity for disagreement over all possible opinion configurations. They then introduce optimization algorithms (e.g., **coordinate descent**) to minimize these risk measures by making a small number of targeted edge additions or deletions. This strategy aims to create a more robust and resilient network topology that is less susceptible to polarization, regardless of the specific topic of controversy. Their empirical results show that minimizing the WCR is particularly effective, as it tends to reduce the average-case risk as well, leading to a more generally conflict-resistant network structure.

The dynamics of filter bubbles and content recommendation are also a focus of recent research. In their analysis of filter bubbles, [27] demonstrate how a "network administrator"—modeling a social media platform's content-filtering algorithm—can dramatically increase polarization by minimizing user disagreement. They also propose a "simple remedy" to mitigate this effect. Their solution involves modifying the administrator's objective function by adding an L^2 regularization term. This "**Regularized Dynamics**" approach discourages the algorithm from making large, concentrated changes to a few social connections. Instead, it incentivizes smaller, more distributed adjustments across many edges in the network. The authors show this method to be highly effective: in their experiments, the regularized model limited the increase in polarization to just 4%, compared to an over 4000% increase in the non-regularized model. Crucially, this was achieved while only minimally impacting user disagreement (an increase of at most 5%), suggesting that platforms could control the formation of polarizing echo chambers without significantly harming their engagement-driven business objectives.

Other research investigates network modifications under practical constraints. In their work, [64] propose methods for a centralized planner to reduce sociopolitical polarization by perturbing a social network's structure under a fixed budget. Using the FJ model, where expressed opinions $\mathbf{z}$ are determined by innate opinions $\mathbf{s}$ and the graph Laplacian $\mathbf{L}$ via the relation $\mathbf{z} = (\mathbf{I} + \mathbf{L})^{-1}\mathbf{s}$, they define polarization $P(\mathbf{z})$ as the variance of the expressed opinions. The authors first consider a setting where the planner has full knowledge of the population's opinions. They derive the exact change in polarization from adding an edge and propose two greedy heuristics: a **Coordinate Descent (CD)** strategy that iteratively adds the edge yielding the largest marginal decrease in polarization, and a simpler **Disagreement-**

Seeking (DS) strategy that adds an edge between the two individuals with the highest expressed disagreement, $(z_i - z_j)^2$. In a second, more robust setting, they analyze the problem where opinions are chosen adversarially to maximize polarization. They demonstrate this minimax problem is equivalent to maximizing the spectral gap (λ_2) of the graph's Laplacian. This motivates their third strategy, the **Fiedler Difference (FD)** heuristic, which adds edges between vertices on opposite sides of the network partition induced by the Fiedler vector—the eigenvector corresponding to λ_2. This approach aims to make the network structure more robust against divisive opinion configurations by bridging its most prominent communities.

Extending these approaches to more realistic settings, Cinus et al. [67] address the problem of mitigating polarization and disagreement in social networks by proposing a method to rebalance a user's social feed. Their approach is set within a directed graph context, where edges represent follower-followee relationships, and it is based on the FJ model. The core idea is to re-weight the influence of existing connections rather than creating new ones, thereby preserving each user's total engagement. The authors formulate an optimization problem that seeks to find a new row-stochastic adjacency matrix A^* that minimizes the sum of network polarization and disagreement at the equilibrium state of opinions. For a vector of innate opinions s, their objective function for a directed graph is given by:

$$f(A, s) = s^T(2I - A)^{-T}s + s^T(2I - A)^{-T}\frac{D_{in} - I}{2}(2I - A)^{-1}s$$

where D_{in} is the diagonal matrix of in-degrees. A key contribution is the analysis of this problem's properties, where they demonstrate that while the feasible set of matrices is convex (maintaining original sparsity and row-stochasticity), the objective function is not matrix-convex. To solve this challenging non-convex problem scalably, they develop an algorithm named **Laplacian-Constrained Gradient Descent (LcGD)**, which is based on projected gradient descent. The algorithm efficiently computes the gradient without explicit matrix inversion by solving linear systems and then projects the solution back onto the feasible set to maintain the constraints. Their work is notable for being one of the first to tackle this problem in the more realistic setting of directed graphs, providing a proper generalization of previous work on undirected networks.

Apart from modifying network structure or edge weights, some works explore interventions at the level of user opinions. Matakos et al. [63] propose and formalize two distinct, NP-hard problems for polarization reduction by convincing a small set of k individuals to adopt a neutral stance. The first method, **MODERATEINTERNAL**, models interventions that change an individual's core beliefs, such as through education. The goal is to select a set of k nodes, T_s, and set their internal opinions to zero. If s is the original vector of internal opinions and s' is the modified vector where the opinions for nodes in T_s are zero, the objective is to find the set T_s that minimizes the polarization of the resulting expressed opinions: $\min_{T_s,|T_s|=k} \|(\mathbf{L} + \mathbf{I})^{-1}s'\|^2$. The authors find that this strategy is most effective when targeting "fringe" nodes with the most extreme expressed opinions. On the other hand, convincing fringe nodes with extreme opinions to adopt a neutral stance is not realistic. In contrast, the second method, **MODERATEEXPRESSED**, models interventions that incentivize individuals to moderate their public statements. This approach involves selecting a set of k nodes, T_z, and fixing their expressed opinions to zero, which directly alters the opinion dynamics as these nodes now propagate neutrality. The objective is to choose the set T_z that minimizes the polarization index. This strategy is shown to be most effective when targeting central and influential nodes, as their moderated expression has the greatest cascading effect throughout the network. While this intervention assumes that individuals can be perfectly incentivized to propagate neutral public stances, in practice, the number of individuals who can be targeted (k) is typically constrained by practical considerations such as resource availability and intervention costs. As k increases, the potential to reduce polarization grows, but so does the cost and complexity of implementation. Thus, there is an inherent trade-off between the scale of the intervention

and its feasibility. Nevertheless, the model offers valuable insight by providing an upper bound on the potential effectiveness of moderation-based interventions and can guide the design of more practical strategies.

Although opinion dynamics models allow researchers to simulate the evolution of opinions, design mitigation strategies, and evaluate their effectiveness based on final opinions, this class of methods has several limitations:

1. **Accuracy of Opinion Dynamics Models.** While many opinion dynamics models have been proposed to simulate opinion evolution (e.g., the FJ model, Sznajd model [76], and FJCB [77]—a Friedkin-Johnsen type model that incorporates confirmation bias to better capture echo chamber formation), their accuracy remains insufficiently analyzed and validated on real-world social networks. Moreover, parameter selection often lacks clear guidance or established rules, potentially limiting model performance. For instance, most methods rely on a simplified FJ model that assumes a uniform stubbornness value of 1 for all individuals. This assumption fails to reflect real-world scenarios, where individuals demonstrate varying degrees of stubbornness. Additionally, the model computes the final opinion as $z = (I+L)^{-1}s$, which implies that the sum or average of final opinions equals that of the innate opinions. This suggests that overall opinion remains unchanged after the opinion dynamics process, contradicting real-world observations where collective opinions can shift over time. Furthermore, the simplified FJ model assumes that interpersonal influence weights remain static throughout the process, whereas, in reality, these weights may evolve as individuals' opinions change due to factors such as exposure or personal leaning. Obtaining authentic innate opinions for users, which directly determine final opinions when the Laplacian matrix is given, is also nearly impossible. While the simplified model is frequently used for its analytical tractability and has been adopted in previous studies, these simplifications may significantly limit its ability to accurately capture the complexity of real-world opinion dynamics.

2. **Practicality of Proposed Strategies.** Although opinion dynamics can be simulated, the resulting mitigation strategies may be difficult to implement in practice. For example, in [63], the **MODERATEINTERNAL** intervention models changes to an individual's core beliefs through education or similar methods. However, attempting to educate users to alter their innate opinions is challenging and time-consuming, particularly in large-scale networks or when user identification is constrained by privacy concerns.

Despite these challenges, opinion dynamics-based methods offer a dynamic perspective for analyzing networks and user opinions. Some studies that balance polarization and disagreement (e.g., [50]) point to promising directions that consider both individual polarization and platform interests.

4.3 Agents Addition

Several studies have explored the strategy of introducing new agents into a network to disseminate ideas more effectively or to influence recommender systems, thereby promoting a more diverse flow of information among users.

Ghezelbash et al. [47] present an innovative analytical framework for strategically inducing polarization in social networks by selecting a minimal set of informed agents. Their approach models the social network as a linear dynamical system, where opinion formation is determined by an influence matrix A. The central insight is that steering the network toward a desired final opinion state, x_d, can be formulated as a control problem. By introducing the concept of "equilibratability", they show that the necessary set of informed agents corresponds exactly to the non-zero entries of the vector $(I - A)x_d$. Consequently, the challenge of selecting the fewest agents becomes a zero-norm minimization problem: identifying a target

opinion state x_d that satisfies specific constraints (e.g., ensuring two subgroups attain distinct average opinions) while simultaneously minimizing the sparsity of $(I - A)x_d$. As this problem is NP-hard, the authors reformulate it as a computationally tractable Integer Linear Programming (ILP) problem. A notable finding from their analysis of Zachary's Karate Club network is that the optimal agents for inducing polarization are not necessarily the most highly connected "hubs", but rather "bridge" agents whose connections span opposing factions. This work offers a principled, optimization-based method for targeted intervention, providing a formal alternative to heuristic or simulation-based agent selection strategies. Their method represents a synthesis of opinion dynamics and agent addition techniques.

Although not directly focused on social networks, Rastegarpanah et al. [26] propose a novel approach to reducing polarization in recommender systems by introducing "antidote data". Their method involves injecting a small set of new, artificial users whose ratings are strategically optimized to minimize a chosen polarization metric, specifically the variance of predicted ratings for an item across the user population. This additional data acts as a constraint during the system's training phase, compelling the underlying model (e.g., matrix factorization) to learn item representations that yield less divergent predictions for the original users. Notably, this technique can effectively mitigate polarization without necessitating any changes to the core recommendation algorithm or the original dataset.

4.4 Others

Alatawi et al.[21] summarize several "human-focused prevention" strategies, which empower users to curate their own information feeds and thereby reduce bias. For a more detailed analysis, readers are encouraged to consult [21]. Additionally, there are numerous field-studies and survey-based mitigation strategies, as well as approaches focusing on recommender systems; readers are referred to [41] for further information.

5 Challenges and Opportunities

In this section, we outline the key challenges and potential future research directions in the field of echo chamber studies.

1. **Echo Chamber Detection and Measurement:**
   Users' opinions are central to detecting and measuring echo chambers. Some approaches rely on manually annotated labels in user posts to infer leanings, which requires a deep understanding of the topic and users' intentions. Furthermore, these approaches are not scalable. Others depend on existing scores for certain websites or public figures, which are not always generalizable to other datasets or topics. Automated methods—such as regression, classification models, and LLM-based approaches—have been applied to stance detection in text. Recent advances in LLMs have substantially boosted the accuracy of traditional methods, with F1 scores now exceeding 80%. Their application is still in an early, exploratory stage, and developing effective techniques to adapt and utilize these models for specific stance detection tasks remains an open research challenge. Additionally, the roles, background knowledge, and potential biases of LLMs in stance detection contexts require further investigation [95]. To enable more accurate and efficient opinion inference and support in-depth analysis, further improvements in stance detection—especially those tailored to social networks [96] and controversial topics—might be needed.

2. **Modeling Opinion Evolution:**
   Numerous opinion dynamics models have been proposed to study the evolution of users' opinions. Among these are models addressing polarization and echo chambers, which incorporate crucial

factors such as confirmation bias and social influence. However, as discussed previously, some models are overly simplistic and fail to capture the complex dynamics of real-world social networks. Moreover, there is a lack of analysis regarding the accuracy of these models on large-scale, long-term datasets where echo chambers may form and evolve. Developing opinion dynamics models that can accurately reflect and predict opinion evolution would greatly benefit echo chamber mitigation strategies and help minimize their negative side effects. In addition to improving theoretical models and evaluating them on large-scale datasets, conducting field studies to validate these models in real-world social networks is crucial. Such empirical validation can ensure that the models not only fit observed data but also reliably predict opinion dynamics and the formation of echo chambers. Additionally, field studies can also be conducted to identify factors contributing to the formation of echo chambers, which may help in designing more realistic opinion evolution models.

3. **Quantification Metrics:**
   There is currently no universal, widely accepted metric for quantifying echo chambers in research. Comparative studies of existing metrics are also lacking, making it difficult to assess whether these metrics reliably capture any specific characteristics of echo chambers and provide trustworthy scores. The GE method proposed in [65], which incorporates both network topology and opinion information, offers a promising direction. This method has been compared with other metrics like RWC on certain networks, demonstrating strong performance and sensitivity to both structural and opinion distribution factors in polarization quantification. Given the large number of metrics proposed for echo chamber and polarization measurement, deeper analysis and comparison based on large scale field studies are needed to uncover their strengths and limitations as effective measurement tools.

4. **Mitigation Strategies:**
   Practical mitigation strategies should balance the interests of platform companies, legal requirements, individual rights, ethical considerations, content diversity, and user engagement to foster healthier social network environments. As highlighted in [41], "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." While demoting or banning radical and influential users is often applied in practice (e.g., banning malicious accounts [41]), the effectiveness and consequences of such actions can be complex and not easily predictable. On one hand, user/post demotion may not necessarily trigger the backfire effects sometimes observed with cross-group promotion. On the other hand, high-profile bans, such as the removal of US President Trump's Facebook account, have resulted in significant controversy and political backlash, illustrating that such interventions can have side effects whose costs are not easily measurable. Beyond actions available to network operators, third parties—such as dedicated users, civil society organizations, or automated agents—may also help mitigate echo chambers and filter bubbles by disseminating diverse or corrective content, engaging in counter-speech, or promoting fact-checking initiatives. Furthermore, gamification strategies, such as incentivizing users to engage with diverse viewpoints or rewarding civil discourse, may also be promising approaches to encourage healthier interactions and reduce polarization.

5. **Distinguishing Echo Chambers from Polarization:**
   As noted in [14], "Another issue with the existing definition is the equating of polarization in the network with echo chambers, and subsequently, many of these studies attempt to propose an approach to address polarization". Consequently, much of the related work focuses on polarization mitigation. However, by definition, echo chambers should not be equated solely with polarization, as the "reinforcement" effect is a distinct characteristic, separate from polarity. Thus, an important

future research direction is to develop models and measurement techniques that accurately capture the reinforcement mechanisms unique to echo chambers. Detecting echo chambers by focusing specifically on these reinforcement dynamics—particularly during their formation and evolution—could provide deeper insights into their structure and impact. Furthermore, research into effective mitigation strategies should aim to disrupt or weaken the reinforcement processes that sustain echo chambers, rather than merely reducing overall polarization.

6. **Data Availability and Privacy:**
   There is a lack of large-scale, long-term datasets for researching echo chamber formation and evolution. Collecting such datasets, for example from platforms like Twitter, can be costly and challenging. Furthermore, due to privacy concerns, specific content from these platforms cannot be readily shared. Researchers using datasets from prior studies often need to re-collect content through tweet IDs or links. Therefore, better platform policies are needed—ones that both protect user privacy and facilitate social network research.

6 Conclusion

Echo chambers represent a critical challenge in the landscape of online social networks, with far-reaching consequences for public discourse, information diversity, and societal cohesion. In this survey, we have provided a comprehensive overview of echo chamber research, systematically reviewing approaches for detection, measurement, and mitigation. Our analysis highlights the diversity of methods available, spanning topological, semantic, and hybrid techniques, as well as the multitude of metrics employed in the literature to capture the nuanced effects of echo chambers.

Despite significant progress, key challenges remain. The field lacks universally accepted definitions and robust, generalizable metrics for quantifying echo chamber effects. Many mitigation strategies, while promising in simulation or small-scale studies, face practical and ethical constraints when deployed in real-world systems. Furthermore, the interplay between algorithmic design, user behavior, and societal context complicates the task of designing interventions that not only are effective, but also balance users' rights to diverse information, commercial interests of social media platforms, and practical feasibility.

Looking ahead, future research may address these challenges by developing more accurate models of opinion dynamics, improving the reliability of detection and measurement techniques, and designing mitigation strategies that balance competing interests. Interdisciplinary collaboration—spanning computer science, social science, psychology, and ethics—will be essential for advancing our understanding and management of echo chambers. By fostering a more nuanced and evidence-based approach to echo chambers, we can work towards healthier, more inclusive online environments that promote informed and diverse public discourse.

Acknowledgments

During the literature review and synthesis of related work, we used large language models to assist in generating initial draft summaries of surveyed papers, primarily using methodological descriptions and formulas from the original works. LLMs were also used to help polish and refine the written text throughout the manuscript. All LLM-generated content was subsequently reviewed, classified, and fact-checked by the authors, with connections and relationships between them established manually. The identification and analysis of drawbacks, challenges, and opportunities were carried out by the authors, with LLMs providing assistance for language clarity and editing where needed.

References

[1] Del Vicario, M., Vivaldo, G., Bessi, A., Zollo, F., Scala, A., Caldarelli, G. & Quattrociocchi, W. Echo chambers: Emotional contagion and group polarization on facebook. *Scientific Reports*. **6**, 37825 (2016)

[2] Chkhartishvili, A. & Kozitsin, I. Binary separation index for echo chamber effect measuring. *2018 Eleventh International Conference" Management Of Large-scale System Development"(MLSD*. pp. 1-4 (2018)

[3] Luo, R., Nettasinghe, B. & Krishnamurthy, V. Controlling segregation in social network dynamics as an edge formation game. *IEEE Transactions On Network Science And Engineering*. **9**, 2317-2329 (2022)

[4] Calderón, F., Cheng, L., Lin, M., Huang, Y. & Chen, Y. Content-based echo chamber detection on social media platforms. *Proceedings Of The 2019 IEEE/ACM International Conference On Advances In Social Networks Analysis And Mining*. pp. 597-600 (2019)

[5] Al Atiqi, M., Chang, S. & Deguchi, H. Agent-based approach to resolve the conflicting observations of online echo chamber. *2020 Joint 11th International Conference On Soft Computing And Intelligent Systems And 21st International Symposium On Advanced Intelligent Systems (SCIS-ISIS)*. pp. 1-6 (2020)

[6] Villa, G., Pasi, G. & Viviani, M. Echo chamber detection and analysis: A topology-and content-based approach in the COVID-19 scenario. *Social Network Analysis And Mining*. **11**, 78 (2021)

[7] Guerra, P., Meira Jr, W., Cardie, C. & Kleinberg, R. A measure of polarization on social media networks based on community boundaries. *Proceedings Of The International AAAI Conference On Web And Social Media*. **7**, 215-224 (2013)

[8] Minici, M., Cinus, F., Monti, C., Bonchi, F. & Manco, G. Cascade-based echo chamber detection. *Proceedings Of The 31st ACM International Conference On Information & Knowledge Management*. pp. 1511-1520 (2022)

[9] Cota, W., Ferreira, S., Pastor-Satorras, R. & Starnini, M. Quantifying echo chamber effects in information spreading over political communication networks. *EPJ Data Science*. **8**, 1-13 (2019)

[10] Al Atiqi, M., Chang, S. & Hiroshi, D. Agent-based approach to echo chamber reduction strategy in social media. *2018 Joint 10th International Conference On Soft Computing And Intelligent Systems (SCIS) And 19th International Symposium On Advanced Intelligent Systems (ISIS)*. pp. 1301-1306 (2018)

[11] Botte, N., Ryckebusch, J. & Rocha, L. Clustering and stubbornness regulate the formation of echo chambers in personalised opinion dynamics. *Physica A: Statistical Mechanics And Its Applications*. **599** pp. 127423 (2022)

[12] Madsen, J., Bailey, R. & Pilditch, T. Large networks of rational agents form persistent echo chambers. *Scientific Reports*. **8**, 12391 (2018)

[13] Morini, V., Pollacci, L. & Rossetti, G. Toward a standard approach for echo chamber detection: Reddit case study. *Applied Sciences*. **11**, 5390 (2021)

[14] Mahmoudi, A., Jemielniak, D. & Ciechanowski, L. Echo chambers in online social networks: A systematic literature review. *IEEE Access*. **12** pp. 9594-9620 (2024)

[15] Ge, Y., Zhao, S., Zhou, H., Pei, C., Sun, F., Ou, W. & Zhang, Y. Understanding echo chambers in e-commerce recommender systems. *Proceedings Of The 43rd International ACM SIGIR Conference On Research And Development In Information Retrieval*. pp. 2261-2270 (2020)

[16] Garimella, K., Morales, G., Gionis, A. & Mathioudakis, M. Quantifying controversy on social media. *ACM Transactions On Social Computing*. **1**, 1-27 (2018)

[17] Newman, M. & Girvan, M. Finding and evaluating community structure in networks. *Physical Review E*. **69**, 026113 (2004)

[18] Fortunato, S. Community detection in graphs. *Physics Reports*. **486**, 75-174 (2010)

[19] Sikder, O., Smith, R., Vivo, P. & Livan, G. A minimalistic model of bias, polarization and misinformation in social networks. *Scientific Reports*. **10**, 5493 (2020)

[20] Prasetya, H. & Murata, T. A model of opinion and propagation structure polarization in social media. *Computational Social Networks*. **7** pp. 1-35 (2020)

[21] Alatawi, F., Cheng, L., Tahir, A., Karami, M., Jiang, B., Black, T. & Liu, H. A survey on echo chambers on social media: Description, detection and mitigation. *ArXiv Preprint ArXiv:2112.05084*. (2021)

[22] Cossard, A., Morales, G., Kalimeri, K., Mejova, Y., Paolotti, D. & Starnini, M. Falling into the echo chamber: the Italian vaccination debate on Twitter. *Proceedings Of The International AAAI Conference On Web And*

Social Media. **14** pp. 130-140 (2020)

[23] Nourbakhsh, A., Liu, X., Li, Q. & Shah, S. Mapping the Echo-chamber: Detecting and Characterizing Partisan Networks on Twitter. *Proceedings Of The 2017 International Conference On Social Computing, Behavioral-Cultural Modeling, & Prediction And Behavior Representation In Modeling And Simulation*. pp. 5-8 (2017)

[24] Del Vicario, M., Zollo, F., Caldarelli, G., Scala, A. & Quattrociocchi, W. Mapping social dynamics on Facebook: The Brexit debate. *Social Networks*. **50** pp. 6-16 (2017)

[25] Du, S. & Gregory, S. The echo chamber effect in twitter: Does community polarization increase?. *Complex Networks & Their Applications V: Proceedings Of The 5th International Workshop On Complex Networks And Their Applications (COMPLEX NETWORKS 2016)*. pp. 373-378 (2017)

[26] Rastegarpanah, B., Gummadi, K. & Crovella, M. Fighting fire with fire: Using antidote data to improve polarization and fairness of recommender systems. *Proceedings Of The Twelfth ACM International Conference On Web Search And Data Mining*. pp. 231-239 (2019)

[27] Chitra, U. & Musco, C. Analyzing the impact of filter bubbles on social network polarization. *Proceedings Of The 13th International Conference On Web Search And Data Mining*. pp. 115-123 (2020)

[28] Orbach, M., Bilu, Y., Toledo, A., Lahav, D., Jacovi, M., Aharonov, R. & Slonim, N. Out of the Echo Chamber: Detecting Countering Debate Speeches. arXiv 2020. *ArXiv Preprint ArXiv:2005.01157*.

[29] Interian, R. & Ribeiro, C. An empirical investigation of network polarization. *Applied Mathematics And Computation*. **339** pp. 651-662 (2018)

[30] Lelkes, Y. Mass polarization: Manifestations and measurements. *Public Opinion Quarterly*. **80**, 392-410 (2016)

[31] Markgraf, M. & Schoch, M. Quantification of echo chambers: a methodological framework considering multi-party systems. (2019)

[32] Ertan, G., Çarkoğlu, A. & Aytaç, S. Cognitive political networks: A structural approach to measure political polarization in multiparty systems. *Social Networks*. **68** pp. 118-126 (2022)

[33] Zhang, Y., Friend, A., Traud, A., Porter, M., Fowler, J. & Mucha, P. Community structure in Congressional cosponsorship networks. *Physica A: Statistical Mechanics And Its Applications*. **387**, 1705-1712 (2008)

[34] Dal Maso, C., Pompa, G., Puliga, M., Riotta, G. & Chessa, A. Voting behavior, coalitions and government strength through a complex network analysis. *PloS One*. **9**, e116046 (2014)

[35] Garcia, D., Abisheva, A., Schweighofer, S., Serdült, U. & Schweitzer, F. Ideological and temporal components of network polarization in online political participatory media. *Policy & Internet*. **7**, 46-79 (2015)

[36] Wolfowicz, M., Weisburd, D. & Hasisi, B. Examining the interactive effects of the filter bubble and the echo chamber on radicalization. *Journal Of Experimental Criminology*. **19**, 119-141 (2023)

[37] Emamgholizadeh, H., Nourizade, M., Tajbakhsh, M., Hashminezhad, M. & Esfahani, F. A framework for quantifying controversy of social network debates using attributed networks: biased random walk (BRW). *Social Network Analysis And Mining*. **10**, 90 (2020)

[38] Bozdag, E., Gao, Q., Houben, G. & Warnier, M. 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** pp. 405-415 (2014)

[39] Huang, Z., Silva, A. & Singh, A. Pole: Polarized embedding for signed networks. *Proceedings Of The Fifteenth ACM International Conference On Web Search And Data Mining*. pp. 390-400 (2022)

[40] Guyot, A., Gillet, A., Leclercq, É. & Cullot, N. ERIS: an approach based on community boundaries to assess polarization in online social networks. *International Conference On Research Challenges In Information Science*. pp. 88-104 (2022)

[41] Interian, R., G. Marzo, R., Mendoza, I. & Ribeiro, C. Network polarization, filter bubbles, and echo chambers: an annotated review of measures and reduction methods. *International Transactions In Operational Research*. **30**, 3122-3158 (2023)

[42] Garimella, K., De Francisci Morales, G., Gionis, A. & Mathioudakis, M. Reducing controversy by connecting opposing views. *Proceedings Of The Tenth ACM International Conference On Web Search And Data Mining*. pp. 81-90 (2017)

[43] Chen, X., Lijffijt, J. & De Bie, T. Quantifying and minimizing risk of conflict in social networks. *Proceedings Of The 24th ACM SIGKDD International Conference On Knowledge Discovery & Data Mining*. pp. 1197-1205

(2018)
[44] Interian, R., Moreno, J. & Ribeiro, C. Reducing network polarization by edge additions. *Proceedings Of The 2020 4th International Conference On Intelligent Systems, Metaheuristics & Swarm Intelligence*. pp. 87-92 (2020)
[45] Interian, R., Moreno, J. & Ribeiro, C. Polarization reduction by minimum-cardinality edge additions: Complexity and integer programming approaches. *International Transactions In Operational Research*. **28**, 1242-1264 (2021)
[46] Haddadan, S., Menghini, C., Riondato, M. & Upfal, E. Repbublik: Reducing polarized bubble radius with link insertions. *Proceedings Of The 14th ACM International Conference On Web Search And Data Mining*. pp. 139-147 (2021)
[47] Ghezelbash, E., Yazdanpanah, M. & Asadpour, M. Polarization in cooperative networks through optimal placement of informed agents. *Physica A: Statistical Mechanics And Its Applications*. **536** pp. 120936 (2019)
[48] Shekatkar, S. Do zealots increase or decrease the polarization of social networks?. *Journal Of Complex Networks*. **8**, cnz036 (2020)
[49] Newman, M. & Girvan, M. Mixing patterns and community structure in networks. *Statistical Mechanics Of Complex Networks*. pp. 66-87 (2003)
[50] Musco, C., Musco, C. & Tsourakakis, C. Minimizing polarization and disagreement in social networks. *Proceedings Of The 2018 World Wide Web Conference*. pp. 369-378 (2018)
[51] Alatawi, F., Sheth, P. & Liu, H. Quantifying the echo chamber effect: an embedding distance-based approach. *Proceedings Of The International Conference On Advances In Social Networks Analysis And Mining*. pp. 38-45 (2023)
[52] Morales, A., Borondo, J., Losada, J. & Benito, R. Measuring political polarization: Twitter shows the two sides of Venezuela. *Chaos: An Interdisciplinary Journal Of Nonlinear Science*. **25** (2015)
[53] Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W. & Starnini, M. The echo chamber effect on social media. *Proceedings Of The National Academy Of Sciences*. **118**, e2023301118 (2021)
[54] Cinelli, M., Morales, G., Galeazzi, A., Quattrociocchi, W. & Starnini, M. Echo chambers on social media: A comparative analysis. *ArXiv Preprint ArXiv:2004.09603*. (2020)
[55] Lai, M., Patti, V., Ruffo, G. & Rosso, P. Stance evolution and twitter interactions in an italian political debate. *Natural Language Processing And Information Systems: 23rd International Conference On Applications Of Natural Language To Information Systems, NLDB 2018, Paris, France, June 13-15, 2018, Proceedings 23*. pp. 15-27 (2018)
[56] Li, Z., Dong, Y., Gao, C., Zhao, Y., Li, D., Hao, J., Zhang, K., Li, Y. & Wang, Z. Breaking filter bubble: A reinforcement learning framework of controllable recommender system. *Proceedings Of The ACM Web Conference 2023*. pp. 4041-4049 (2023)
[57] Citraro, S. & Rossetti, G. Eva: Attribute-aware network segmentation. *International Conference On Complex Networks And Their Applications*. pp. 141-151 (2019)
[58] Zhang, P., Haq, E., Zhu, Y., Hui, P. & Tyson, G. Echo chambers within the russo-ukrainian war: The role of bipartisan users. *Proceedings Of The International Conference On Advances In Social Networks Analysis And Mining*. pp. 154-158 (2023)
[59] Karypis, G. & Kumar, V. METIS: A software package for partitioning unstructured graphs, partitioning meshes, and computing fill-reducing orderings of sparse matrices. (1997)
[60] Amendola, M., Cavaliere, D., De Maio, C., Fenza, G. & Loia, V. Towards echo chamber assessment by employing aspect-based sentiment analysis and gdm consensus metrics. *Online Social Networks And Media*. **39** pp. 100276 (2024)
[61] Barberá, P., Jost, J., Nagler, J., Tucker, J. & Bonneau, R. Tweeting from left to right: Is online political communication more than an echo chamber?. *Psychological Science*. **26**, 1531-1542 (2015)
[62] Garimella, K., De Francisci Morales, G., Gionis, A. & Mathioudakis, M. Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. *Proceedings Of The 2018 World Wide Web Conference*. pp. 913-922 (2018)
[63] Matakos, A., Terzi, E. & Tsaparas, P. Measuring and moderating opinion polarization in social networks. *Data Mining And Knowledge Discovery*. **31** pp. 1480-1505 (2017)
[64] Rácz, M. & Rigobon, D. Towards consensus: Reducing polarization by perturbing social networks. *IEEE*

Transactions On Network Science And Engineering. **10**, 3450-3464 (2023)

[65] Hohmann, M., Devriendt, K. & Coscia, M. Quantifying ideological polarization on a network using generalized Euclidean distance. *Science Advances*. **9**, eabq2044 (2023)

[66] Weidemann, C. Quantifying Multipolar Polarization. *ArXiv Preprint ArXiv:2405.16352*. (2024)

[67] Cinus, F., Gionis, A. & Bonchi, F. Rebalancing social feed to minimize polarization and disagreement. *Proceedings Of The 32nd ACM International Conference On Information And Knowledge Management*. pp. 369-378 (2023)

[68] Krackhardt, D. & Stern, R. Informal networks and organizational crises: An experimental simulation. *Social Psychology Quarterly*. pp. 123-140 (1988)

[69] Nyhan, B. & Reifler, J. When corrections fail: The persistence of political misperceptions. *Political Behavior*. **32**, 303-330 (2010)

[70] Lord, C., Ross, L. & Lepper, M. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence.. *Journal Of Personality And Social Psychology*. **37**, 2098 (1979)

[71] Taber, C. & Lodge, M. Motivated skepticism in the evaluation of political beliefs. *American Journal Of Political Science*. **50**, 755-769 (2006)

[72] Bail, C., Argyle, L., Brown, T., Bumpus, J., Chen, H., Hunzaker, M., Lee, J., Mann, M., Merhout, F. & Volfovsky, A. Exposure to opposing views on social media can increase political polarization. *Proceedings Of The National Academy Of Sciences*. **115**, 9216-9221 (2018)

[73] Nickerson, R. Confirmation bias: A ubiquitous phenomenon in many guises. *Review Of General Psychology*. **2**, 175-220 (1998)

[74] Friedkin, N. & Johnsen, E. Social influence and opinions. *Journal Of Mathematical Sociology*. **15**, 193-206 (1990)

[75] Bail, C. Breaking the social media prism: How to make our platforms less polarizing. (Princeton University Press,2022)

[76] Sznajd-Weron, K., Sznajd, J. & Weron, T. A review on the Sznajd model—20 years after. *Physica A: Statistical Mechanics And Its Applications*. **565** pp. 125537 (2021)

[77] Chen, T., Wang, X. & Tsourakakis, C. Polarizing opinion dynamics with confirmation bias. *International Conference On Social Informatics*. pp. 144-158 (2022)

[78] Newman, M. Fast algorithm for detecting community structure in networks. *Physical Review E—Statistical, Nonlinear, And Soft Matter Physics*. **69**, 066133 (2004)

[79] Blondel, V., Guillaume, J., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *Journal Of Statistical Mechanics: Theory And Experiment*. **2008**, P10008 (2008)

[80] Rosvall, M., Axelsson, D. & Bergstrom, C. The map equation. *The European Physical Journal Special Topics*. **178**, 13-23 (2009)

[81] Pons, P. & Latapy, M. Computing communities in large networks using random walks. *Computer And Information Sciences-ISCIS 2005: 20th International Symposium, Istanbul, Turkey, October 26-28, 2005. Proceedings 20*. pp. 284-293 (2005)

[82] Spohr, D. Fake news and ideological polarization: Filter bubbles and selective exposure on social media. *Business Information Review*. **34**, 150-160 (2017)

[83] Törnberg, P. Echo chambers and viral misinformation: Modeling fake news as complex contagion. *PLoS One*. **13**, e0203958 (2018)

[84] Choi, D., Chun, S., Oh, H., Han, J. & Kwon, T. Rumor propagation is amplified by echo chambers in social media. *Scientific Reports*. **10**, 310 (2020)

[85] Metaxas, P. & Finn, S. The infamous# Pizzagate conspiracy theory: Insight from a TwitterTrails investigation. *Wellesley College Faculty Research And Scholarship*. **188** pp. 1-5 (2017)

[86] Asur, S., Huberman, B., Szabo, G. & Wang, C. Trends in social media: Persistence and decay. *Proceedings Of The International AAAI Conference On Web And Social Media*. **5**, 434-437 (2011)

[87] Colleoni, E., Rozza, A. & Arvidsson, A. Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. *Journal Of Communication*. **64**, 317-332 (2014)

[88] Wollebæk, D., Karlsen, R., Steen-Johnsen, K. & Enjolras, B. Anger, fear, and echo chambers: The emotional basis for online behavior. *Social Media+ Society*. **5**, 2056305119829859 (2019)

[89] Amelkin, V., Bogdanov, P. & Singh, A. A distance measure for the analysis of polar opinion dynamics in

social networks. *ACM Transactions On Knowledge Discovery From Data (TKDD)*. **13**, 1-34 (2019)

[90] Zhu, J., Ni, P., Tong, G., Wang, G. & Huang, J. Influence maximization problem with echo chamber effect in social network. *IEEE Transactions On Computational Social Systems*. **8**, 1163-1171 (2021)

[91] Garimella, K., Gionis, A., Parotsidis, N. & Tatti, N. Balancing information exposure in social networks. *Advances In Neural Information Processing Systems*. **30** (2017)

[92] Aslay, C., Matakos, A., Galbrun, E. & Gionis, A. Maximizing the diversity of exposure in a social network. *2018 IEEE International Conference On Data Mining (ICDM)*. pp. 863-868 (2018)

[93] Tu, S., Aslay, C. & Gionis, A. Co-exposure maximization in online social networks. *Advances In Neural Information Processing Systems*. **33** pp. 3232-3243 (2020)

[94] Banerjee, P., Chen, W. & Lakshmanan, L. Mitigating filter bubbles under a competitive diffusion model. *Proceedings Of The ACM On Management Of Data*. **1**, 1-26 (2023)

[95] Zhang, B., Dai, G., Niu, F., Yin, N., Fan, X., Wang, S., Cao, X. & Huang, H. A survey of stance detection on social media: New directions and perspectives. *ArXiv Preprint ArXiv:2409.15690*. (2024)

[96] Zhang, C., Zhou, Z., Peng, X. & Xu, K. Doubleh: Twitter user stance detection via bipartite graph neural networks. *Proceedings Of The International AAAI Conference On Web And Social Media*. **18** pp. 1766-1778 (2024)

[97] Chen, J., Mang, Q., Zhou, H., Peng, R., Gao, Y. & Ma, C. Scalable Algorithm for Finding Balanced Subgraphs with Tolerance in Signed Networks. *Proceedings Of The 30th ACM SIGKDD Conference On Knowledge Discovery And Data Mining*. pp. 278-287 (2024)

[98] Bonchi, F., Galimberti, E., Gionis, A., Ordozgoiti, B. & Ruffo, G. Discovering polarized communities in signed networks. *Proceedings Of The 28th Acm International Conference On Information And Knowledge Management*. pp. 961-970 (2019)