RESEARCH-ARTICLE

Comparing Echo Chamber Detection Metrics: A Cross-modeling and Cross-platform Analysis of Twitter and Reddit

PAOLA IMPICCICHÈ, University of Milano-Bicocca, Milan, MI, Italy

MARCO VIVIANI, University of Milano-Bicocca, Milan, MI, Italy

Published: 16 October 2025
Online AM: 10 December 2024
Accepted: 24 October 2024
Revised: 11 September 2024
Received: 31 January 2024

Comparing Echo Chamber Detection Metrics: A Cross-modeling and Cross-platform Analysis of Twitter and Reddit

PAOLA IMPICCICHÈ, Dipartimento di Informatica, Sistemistica e Comunicazione, Università degli Studi di Milano-Bicocca, Milan, Italy
MARCO VIVIANI, Dipartimento di Informatica, Sistemistica e Comunicazione, Università degli Studi di Milano-Bicocca, Milan, Italy

Social media platforms have become central arenas for public discourse, enabling the exchange of ideas and information among diverse user groups. However, the rise of echo chambers, where individuals reinforce their existing beliefs through repeated interactions with like-minded users, poses significant challenges to the democratic exchange of ideas and the potential for polarization and information disorder. This article presents a comparative analysis of the main metrics that have been proposed in the literature for echo chamber detection, with a focus on their application in a cross-platform scenario constituted by the two major social media platforms, i.e., Twitter (now renamed 𝕏) and Reddit. The echo chamber detection metrics considered encompass network analysis, content analysis, and hybrid solutions. The findings of this work shed light on the unique dynamics of echo chambers present on the two social media platforms, while also highlighting the strengths and limitations of various metrics employed to identify them, and their transversality to the different social graph modeling and domains considered.

CCS Concepts: • **Information systems** → **Social networks**; • **Human-centered computing** → **Social media**; **Social content sharing**; **Social network analysis;**

Additional Key Words and Phrases: Echo chambers, filter bubbles, polarization, social media, Social Network Analysis (SNA), sentiment analysis, User-Generated Content (UGC), misinformation

ACM Reference Format:
Paola Impicciché and Marco Viviani. 2025. Comparing Echo Chamber Detection Metrics: A Cross-modeling and Cross-platform Analysis of Twitter and Reddit. *ACM Trans. Web* 19, 4, Article 43 (October 2025), 23 pages. https://doi.org/10.1145/3707701

1 Introduction

The widespread use of social media has transformed the way information is accessed and opinions are shaped [43]. It has created new opportunities for sharing information while also raising concerns about its influence on social dynamics, perception of reality, and public opinion [44]. This is related to the numerous factors that affect the dynamics of *information spreading* on social media.

Authors' Contact Information: Paola Impicciché, Dipartimento di Informatica, Sistemistica e Comunicazione, Università degli Studi di Milano-Bicocca, Milan, Italy; e-mail: paola.impicciche@gmail.com; Marco Viviani (Corresponding author), Dipartimento di Informatica, Sistemistica e Comunicazione, Università degli Studi di Milano-Bicocca, Milan, Italy; e-mail: marco.viviani@unimib.it.

ACM 1559-1131/2025/10-ART43
https://doi.org/10.1145/3707701

On the one hand, we have factors characterized by a *technological* nature. Personalized search engines and recommender systems, for instance, can limit the variety of content users are exposed to by suggesting primarily content that aligns with their past consumption patterns, leading to the generation of *filter bubbles*. This concept, introduced by Eli Pariser in 2011 [37], refers to personalized information bubbles created by algorithms, where individuals are mainly exposed to information that aligns with their preferences, beliefs, and online behavior, which are usually collected in a *user profile*. On the other hand, also *psychosocial* factors can affect information spreading, such as *selective exposure* [17] (i.e., the inclination to seek information that reinforces existing beliefs), *confirmation bias* [36] (i.e., the tendency to interpret information in a way that confirms pre-existing opinions), and *homophily* [31] (i.e., the inclination to connect with people who share similar characteristics). All these factors together, exacerbated by the filter bubble phenomenon, can contribute to the emergence of polarized environments commonly referred to as *echo chambers*. They represent closed groups of people where specific opinions and beliefs on a topic are amplified and reinforced through repeated interactions with like-minded users or sources [7]. Hence, the formation of echo chambers is inherently tied to the concept of *social interaction*, which is absent in filter bubbles. In fact, while in a filter bubble, a person operates on an individual level, i.e., exists in isolation within their own bubble [37], echo chambers involve a collective dimension. Individuals have a certain level of agency in deciding whether to be part of an echo chamber, as they can choose whom to connect with or form relationships. In this sense, a more formal definition of an echo chamber has been recently provided in the literature by Morini et al. [35]:

> "*Given a network describing users' interactions centered on a controversial topic, an echo chamber is a subset of the network nodes (users) who share the same ideology and tend to have dense connections primarily within the same group.*"

Given these characteristics, echo chambers also affect the way in which information disseminates over a network and *influence* people [12, 48], also with the risk of increasing people's reliance on online *misinformation* [40] and other forms of *information disorder* [6, 46]. Indeed, the *feedback loop* established within an echo chamber, whereby sharing content within a group increases its credibility for members, causes misinformation to be reinforced and considered increasingly true and reliable. Another reason is the *lack of contradiction*: echo chambers are basically closed systems that hinder users' exposure to different perspectives and critical opinions that could correct or challenge the erroneous information circulating in a group.

At this point, analyzing the echo chamber phenomenon by highlighting, in particular, how to identify dense connections and the sharing of the same ideology over controversial topics, becomes of paramount importance as it is linked to the increase in polarization in online environments and the fragmentation of society. However, recently, the possibility of identifying echo chambers (and, hence, their actual existence) has been questioned due to the scarcity of *comparative studies* on the solutions available for echo chamber detection [15]. In the literature, some approaches concentrate solely on the network's structural aspects by analyzing the connections among users; others draw upon distinct semantic features extracted from user-generated texts; finally, a subset of approaches combines both network topology and content semantics in hybrid methodologies. Each of these approaches relies on the use of specific and distinct *metrics* for the detection of echo chambers that, based on the considered approach, exploit different modeling of the social graph. In addition to this, a predominant number of studies assess these metrics through case studies confined to individual platforms or specific contexts, predominantly political, casting uncertainties on their actual generalizability across different domains and social platforms [15].

1.1 Contribution

In order to tackle the above-mentioned open issues, this article proposes a study of the effectiveness, generalizability, and limitations of the main echo chamber detection metrics defined or considered by the different approaches proposed in the literature, taking into account data from different domains involving controversial topics. Furthermore, since different social media imply different interaction modes for users, we focus our analysis on two platforms: Twitter (now rebranded as 𝕏),[1] a microblogging platform where users share short messages known as "tweets" (now 𝕏's posts) and use hashtags to engage in conversations on specific topics, and Reddit,[2] organized as a forum where news is aggregated into dedicated "subreddits" covering a multitude of topics and communities.

The obtained results demonstrated that network-based metrics, particularly those derived from random walks, effectively identify echo chambers on Twitter, specifically when focusing on the "retweet" interaction. These findings highlighted the critical role of the underlying network structure shaped by this specific type of user relationship. This result corroborates findings from previous studies, including Garimella et al. [19], which had highlighted similar conclusions on Twitter. However, earlier research also noted the limited effectiveness of these measures for other types of interaction networks. Our results also confirmed this behavior for Twitter on the "mention" interaction network, and observed it on Reddit given the diverse social dynamics associated with different user relationships (more content-oriented) characterizing this platform. This consistency with previous findings on Twitter, along with their validation on Reddit, establishes a solid foundation for our cross-platform analysis. Furthermore, our results demonstrate that integrating alternative models incorporating semantic elements enhances the effectiveness of echo chamber detection metrics, particularly sentiment-based and hybrid metrics on Reddit. This suggests that accounting for content is essential for certain types of interaction networks, even across different platforms.

1.2 Organization

The rest of the article is organized as follows: Section 2 provides an overview of the literature related to the problem of echo chamber detection. Section 3 describes in detail data, graph modeling, and metrics used to verify the presence of echo chambers on social media platforms. Section 4 concerns the experimental evaluation of the results obtained in the comparison among distinct metrics. Section 5 discusses the conclusions and future work.

2 Related Work

A whole range of approaches to identifying online echo chambers have been and are being proposed in the literature. Some take more into account the network's *topology*, while others, especially recently, also focus on the *content* being disseminated. In this regard, the concepts of *controversy* and *homogeneity* play a key role with respect to focusing on the structure of the network or the content therein to identify echo chambers. When related to topics, the former concept, according to Beelen et al. [4], can be defined as a public debate that divides society, bringing out ideological divisions and oppositions between different value systems. It can therefore be directly linked to a topological separation, persistent over time, between groups representing divergent views within a social network. The concept of homogeneity when discussing topics, well explored

[1] https://x.com/. Although the platform has changed its name from Twitter to 𝕏, we use the former name in this article because the data used refer to the period when it was still called Twitter.
[2] https://www.reddit.com/

by Quattrociocchi et al. [38], can be understood as the closeness of ideas or similarity between individuals who, according to the principle of homophily, are more likely to establish connections.

At a more refined level of detail, recently Morini et al. [35] proposed to account for differences in the *scale* of the detected echo chambers, going so far as to distinguish: (*i*) *macro-scale* approaches: these identify echo chambers by looking at the *interaction network* at a high level, often forcing the subdivision of the network into *two partitions* in which polarization is expected to be found. Peculiarities that might arise in some specific areas of the network are not taken into account in these methods; (*ii*) *micro-scale* approaches: these focus on analyzing the behavior of *individual users* by specifically analyzing the *content* they disseminate. A disadvantage of such approaches is the loss of the aggregate dimension of the interaction network; (*iii*) *meso-scale* approaches: involve approaches that identify *subsets of nodes* in the network that have the characteristics of an echo chamber. These methods are based on the assumption that neither all users are part of a single polarized community, nor that a controversial topic leads to the formation of only two opposing factions in the social network. They therefore aspire to the identification of multiple echo chambers representative of a multifaceted debate in which the effects of controversy are not limited to a bipartisan view.

In this article, among the echo chamber detection literature solutions, we focus particularly on those that have proposed or studied suitable metrics, both relying on controversy or homogeneity, for the task under consideration. In the following, these approaches are roughly categorized into three groups [16, 35]: *network-based*, *content-based*, and *hybrid*, and with respect to each of these categories, reference will be made to whether they allow the identification of macro-, micro-, or meso-scale echo chambers.

2.1 Network-based Approaches

Network-based approaches rely on the assumption that echo chambers are detectable by looking only at network topology. Madsen et al. [29], for example, discussed how the structure of a network alone can significantly influence the formation of persistent echo chambers. In general, when it comes to a controversial topic, there are distinguishable clusters in the interaction network with a high level of segregation that indicate the presence of a polarized environment. Hence, network-based approaches purposely neglect the content produced by users in the network, as it is deeply dependent on the context [16], in particular, the language used and the platform that shapes the way users produce and consume content. The typical steps that constitute a *network-based pipeline* for echo chamber detection [19], include *graph construction*, *graph partitioning*, and *measuring controversy* employing some suitable *controversy metrics*.

Among these approaches, Guerra et al. [21] propose a *polarization metric* that focuses on the boundary between potentially polarized communities, aiming at better capturing the notions of antagonism and polarization. In particular, the research demonstrates that, while modularity is commonly used to measure the level of segmentation within a network, it is erroneously applied as a direct measure of polarization. Indeed, non-polarized networks may also exhibit high modularity, highlighting the limitations of using such a metric as a direct indicator of antagonism between groups. Morales et al. [34] introduce a methodology for examining and quantifying political polarization in social interactions. Their work employs a model where a minority of influential individuals shape opinions within a network, resulting in a probability density function for opinions. A *polarization index* is then used to gauge the level of political polarization in the distribution. By applying this methodology to a Twitter conversation about Hugo Chavez, the research shows that polarization varies depending on the network's structure, revealing a strong connection between influential individuals and emerging polarization in social media discussions. Garimella et al. [19] present a methodology to detect and measure controversy in social media

discussions, with a focus on identifying controversial topics in various domains. The authors show that their newly proposed *Random-Walk-based metric* outperforms existing methods in identifying controversial topics, making it a valuable tool for real-world applications. Conover et al. [10] employed the *Label Propagation algorithm* [20] to unveil the political structure of specified retweet and mention networks. The findings reveal a highly segregated retweet network with limited connectivity between left- and right-leaning users, while the "mention" network shows more interaction between ideologically opposed individuals. Furthermore, Cossard et al. [11] employ the *Infomap algorithm* [39] to search for strongly segregated user communities within the Italian vaccination debate. All these approaches involve the identification of *macro-scale* echo chambers. The work by Coletto et al. [9] proposes an alternative approach for controversy identification that makes no assumptions about graph partitioning, but is based on the study of local patterns in the interaction network. Starting from Twitter data belonging to *topic* of different domains (news, politics, gossip, celebrity, entertainment, etc.), this approach models social networks through three different structures, i.e., a user graph, a reply tree, and a reply graph. According to the authors, the three structures are characterized by *motif*, or recurring patterns of user interactions, which can be used to discriminate between controversial and non-controversial content. Considering the importance of patterns in addition to the global structural characteristics of the network, the approach allows for more granular analysis of certain areas of the graph and thus results in a network-based technique that allows for the identification of meso-scale echo chambers.

2.2 Content-based Approaches

Content-based approaches focus on detecting polarized environments by examining a user's shared or consumed content independently of their interactions with others. They analyze, in most cases, *User-Generated Content* (UGC) using NLP techniques and in particular *Sentiment Analysis* [32], regardless of the structural characteristics of the social network. These approaches, considering that *homogeneity* is one of the primary drivers for the dissemination of content [13], first infer the polarized opinion (if any) of individual users based on the content produced and/or consumed, and then the presence of distinct homogeneous (and hence polarized) communities.

Despite Garimella et al. [19] showing that simple forms of textual content representation (e.g., bag-of-words) are not effective in measuring the level of controversy, Mejova et al. [33] highlight statistically significant differences between controversial and non-controversial topics in terms of vocabulary usage, sentiment of words, and the occurrence of negative terms (more prevalent in controversial topics). These aspects are therefore considered most relevant for evaluating the level of conflict within a debate. For example, in the work of Sriteja et al. [40], the authors focus on journalistic articles related to the topics discussed by users on the Facebook platform. In the case of a polarized debate, the assumption is that there is a high probability of encountering semantically similar words such as "disagreement", "controversy", and "criticism". The occurrence of these controversy words in sentences is then used as an indicator of controversy. Other features considered include user opinions and interest in the debated topic, extracted through Sentiment Analysis, vocabulary analysis, and user interactions. Similarly, Beelen et al. [4] extract from comments to news articles three categories of content-related features: structural, linguistic, and emotional. The former concern the total number of comments as well as the percentage and speed of responses. Linguistic features capture variations in debate in terms of *Part-of-Speech* (PoS) tags and vocabulary (occurrence of words expressing agreement or disagreement). Emotional features identify the sentiment orientation of individual comments to then determine overall how varied the reaction to an article is.

The works just described illustrate, by way of example, the possibility of analyzing the content and extracting features for the potential identification of controversy within topics. When such

features are applied to the context of echo chamber identification, they are done so through approaches that are mostly micro-scale. For instance, An et al. [3] explore the American political discourse on Facebook and Twitter, aiming at identifying partisan users who predominantly share news articles aligned with their political ideologies. For each social media user, a *net partisan skew* is calculated as the difference between the number of shared contents with a conservative stance and the number of contents with a liberal stance. The metric defined in this way allows quantifying the political orientation of each individual based on how balanced or unbalanced the sharing of content is. The bimodal distribution of the net partisan skew determines the two echo chambers within the social network, with a peak representing conservative users and an opposite peak for liberal users. Another study that leverages the opinions expressed by users on a social network is the work by Matakos et al. [30], which proposes a *polarization index* capable of quantifying the extent to which opinions circulating in a network tend to concentrate in communities, leading to the formation of echo chambers. The authors, in particular, employ the *opinion formation model* introduced by Friedkin and Johnsen in [18]. Del Vicario et al. [13] deduces user polarization directly from their content-related actions on the social media platform. The study analyzes the dissemination of scientific and conspiratorial content on Facebook, calculating for each individual the *edge homogeneity*, indicating the fraction of an individual's *likes* given to conspiratorial content. The study demonstrates that information transmission occurs within clusters where all connections are homogeneous, or in mixed areas of the network where there is still a majority of homogeneous connections. Similarly, Quattrociocchi et al. [38] determines user polarization on Facebook based on scientific or conspiratorial content and calculates the average *edge homogeneity* of post-sharing chains to highlight how content tends to be confined within echo chambers. The two communities identified using this approach, besides being characterized by similar patterns of information consumption (statistics related to *likes*, posts, and comments), also exhibit a similar network structure. Finally, Al-Ayyoub et al. [1] conduct Sentiment Analysis to assess whether a topic is controversial or not, introducing *sentiment-based metrics* that are valuable for detecting and quantifying controversial topics and echo chambers.

Pretty recently, also another NLP technique, i.e., *Stance Detection* [2], has been employed in the framework of echo chamber detection. In particular, Calderón et al. [5] explored content-based echo chamber detection on social media platforms using a combination of Stance Detection and Sentiment Analysis to generate an *echo chamber index*. They proposed a model that evaluates user posts to determine their stance on various topics, effectively identifying echo chambers by clustering users with similar stances. This approach provides insights into the underlying topics driving the polarization. Lo et al. [28] extended the work by developing methods to escape from echo chambers. They proposed a model that identifies users' stances and recommends diverse content that challenges their existing views. By promoting exposure to different perspectives, their approach aims at reducing the reinforcement of echo chambers and encourage more balanced information consumption.

2.3 Hybrid Approaches

Hybrid approaches share characteristics from both network and content-based approaches. They utilize the network structure to measure the degree of segregation between communities while also analyzing semantic information extracted from the content to identify patterns of homophily.

The work of Morini et al. [35] presents a general framework for identifying echo chambers that involves four steps: the identification of a controversial issue, the inference of users' ideology on the controversy, the construction of users' debate network, and the detection of homogeneous

meso-scale communities. Zarate et al. [47] employ community detection techniques to partition the interaction network, and subsequently utilize sentence embedding techniques to associate a vector with each user that encapsulates the syntactic and semantic properties of the text posted online. Vector distances are used to identify users who are semantically related or have similarities in their text content. Villa et al. [45] focus on detecting and analyzing echo chambers on Twitter related to the spread of COVID-19. It introduces different graph representations based on topology and content aspects and evaluates echo chambers using various controversy metrics. Cinelli et al. [7] combines one of the typical steps of content-based methods, namely the inference of the so-called *individual leaning* for each user (i.e., the user's attitude over a given topic), with the community detection step typical of network-based strategies. This work is the only one, among those described in this Related Work section, specifically focusing on the patterns of homophily found in the interaction networks of various platforms such as Facebook, Twitter, Gab, and Reddit. It also considers different controversial topics (e.g., abortion, vaccinations, gun control).

3 A Comparison of Echo Chamber Detection Metrics

The presence of numerous diverse solutions in the literature addressing the problem has led to sometimes conflicting results, casting doubt on the actual identifiability/existence of echo chambers and their effects [7]. Dubois and Blank [15] have particularly highlighted two important criticisms: (*i*) the fact that many studies focus on a single platform, significantly limiting the generalizability of the proposed approaches; (*ii*) the detection of exposure to contrasting ideas does not take into account the various ways individuals gather information on social media. In addition to these considerations, there is another criticism: (*iii*) the case studies in the presented works mostly revolve around often single-domain topics, casting doubt on the effectiveness of the methods in different domains. To tackle the above-mentioned issues, in this article we aim at providing a comparative evaluation from the perspective of the effectiveness, generalizability, and limits of the echo chamber detection metrics from some of the most promising literature approaches outlined in the previous Related Work section, also considering different topics and social media platforms, in our case Twitter and Reddit. Therefore, in this section, we provide details on the cross-platform data used (Section 3.1), the methods employed for modeling the social graphs related to distinct topics and their partitioning (Section 3.2), and the metrics considered (Section 3.3).

3.1 Cross-platform Data

In total, we considered 10 datasets, i.e., 6 datasets from Twitter and 4 from Reddit, constituted by data related to *topics* that have ignited extensive discussions, both online and offline. One of the Twitter datasets revolves around the theme of *vaccinations* and is accessible through Kaggle.[3] The remaining Twitter datasets are sourced from the work by Garimella et al. [19] and constitute our ground truth w.r.t. controversy, being three datasets already labeled as controversial (resolving around the *#beefban, #gunsense, #ukraine* hashtags) and two as non-controversial (*#ultralive, #nationalkissingday*). The Reddit datasets were collected by considering *subreddits* related to sociopolitical issues, i.e., *gun control*,[4] *minority discrimination*,[5] *political sphere*,[6] and *vaccinations*.[7] The first three thematic areas are the same as those selected in the study by Morini et al. [35]. For each

[3] https://www.kaggle.com/keplaxo/twitter-vaccination-dataset
[4] https://www.reddit.com/r/guncontrol/, https://www.reddit.com/r/gunpolitics/, and https://www.reddit.com/r/antiwar/
[5] https://www.reddit.com/r/racism/ and https://www.reddit.com/r/againsthatesubreddits/
[6] https://www.reddit.com/r/republican/, https://www.reddit.com/r/marchagainsttrump/, and https://www.reddit.com/r/democrats/
[7] https://www.reddit.com/r/debatevaccines/

selected subreddit, we retrieved the top 1,000 posts, along with their respective comments,[8] using the Reddit API.[9]

3.2 Graph Building and Partitioning

For each dataset (resolving around a given topic), we built an *undirected*, *weighted graph* G, in which each *vertex* of the network represents a *user*, and where the *edges* are generated based on the different *interaction modes* provided by the two platforms, i.e., *retweet*, *mention*, and *comment*. For the calculation of the different *metrics* that will be detailed below (Section 3.3), it was necessary to construct different graph modeling, i.e., a *structural modeling* (Section 3.2.1) and a *content-enriched modeling* (Section 3.2.2), as follows. Details are then provided on the choices related to graph partitioning (Section 3.2.3).

3.2.1 Structural Modeling. The *weight* on the edges corresponds to the total number of interactions observed between two users with respect to that interaction mode. In particular, depending on the type of available data and the available or characteristic interaction mode, we constructed one of the following graphs for each dataset:

- *Retweet graph*: an edge e exists between users u and v if u retweeted at least one post generated by v or vice versa. We follow this modeling for the graphs built over *#beefban, #gunsense, #ukraine, #ultralive,* and *#nationalkissingday*;[10]
- *Mention graph*: for the *vaccinations* graph related to Twitter, two users u and v are connected by en edge e if u mentioned v in a post or vice versa;[11]
- *Comment graph*: an edge e exists between users u and v if u replied to a comment or post generated by v or vice versa. We used this modeling for the graphs built over the Reddit data.[12]

3.2.2 Content-enriched Modeling. For datasets including textual content (i.e., *vaccinations* from Twitter, and *gun control*, *minority discrimination*, *political sphere*, and *vaccinations* from Reddit), we built additional graphs based on a modeling that integrates semantic aspects of the content posted by users. Specifically, by following the hybrid echo chamber detection approach proposed by Villa et al. in [45], edge weights on the graph are modified based on the following three content-based scores:[13]

- *Sentiment similarity score*: it increases the weight in the structural modeling for those edges that connect users with a similar sentiment. Formally, $\omega_{ss}(u, v) = 1 + ss(u, v)$, where u and v representing two generic users and ω_{ss} is the new edge weight;
- *Topic similarity score*: it increases the weight in the structural modeling for those edges that connect users discussing the same topics. Formally, $\omega_{ts}(u, v) = 1 + ts(u, v)$, where u and v representing two generic users and ω_{ts} is the new edge weight;
- *Hybrid score*: it employs both sentiment similarity and topic similarity for defining new weights. Formally, $\omega_{hs}(u, v) = 1 + ss(u, v) + ts(u, v)$, where u and v representing two generic users and ω_{hs} is the new edge weight.

[8]Original data from Morini et al. [35] primarily consists of an edge list with associated node labels, without any textual content. Our purpose in this article is to consider user interactions involving such kind of content. Therefore, the performed data collection did not replicate the exact user set from Morini et al.'s study, but aimed to capture a broad range of interactions within the selected subreddits, providing a comprehensive view of discussions during the data collection period.

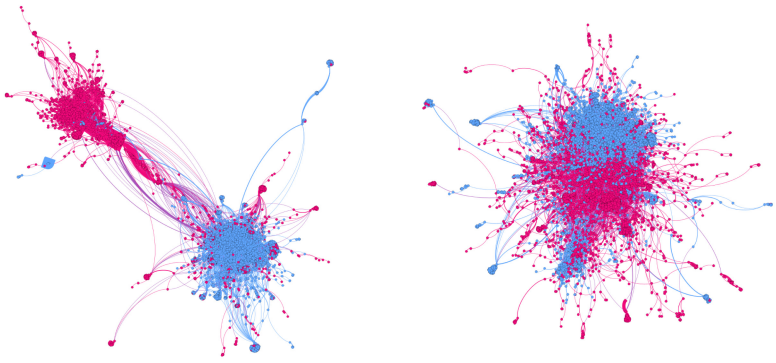
[9]https://www.reddit.com/dev/api/

[10]Only the *retweet* interaction mode was considered in the original dataset.

[11]Only the *mention* interaction mode was considered in the original dataset.

[12]The *commment* interaction mode is the typical one in Reddit.

[13]Further details on the computation of the *sentiment similarity score* and *topic similarity score* can be found in [45].



(a) *#gunsense* (b) *#ultralive*

Fig. 1. Visualization of the communities in a controversial (a) and a non-controversial (b) graph.

To perform Sentiment Analysis, we employed the *Valence Aware Dictionary and sEntiment Reasoner* (VADER) [22], a domain-free model that uses a lexicon-based approach meant for social media content to assign a semantic orientation to words in a text. This tool generalizes very well, requires no training, and has defined rules for handling emojis, emoticons, and slang, making it particularly suitable for the social media context. To perform topic modeling, we used ProLDA [41] implemented in the OCTIS framework [42], a tool that enables preprocessing, training, and evaluation of various state-of-the-art topic models.

3.2.3 Graph Partitioning. While recent efforts have aimed at moving away from treating echo chambers as a binary partition between two strongly opposed communities, we acknowledge that many existing solutions in the literature still operate under this assumption. In our examination, we have specifically concentrated on the METIS graph partitioning algorithm [25] to compute echo chamber detection metrics on these solutions. This algorithm employs a particularly effective multilevel approach to graph partitioning with respect to other literature solutions [45]. It starts by simplifying the input graph into a smaller-sized graph, on which it performs balanced partitioning to achieve a minimal-weight edge cut. Subsequently, it projects the partitioning back to the original graph. The algorithm further refines the partitions by transferring nodes between communities to enhance balance and minimize the edge-cut weight. In this phase, it is possible to observe the difference between controversial graphs where two distinct partitions are well separated and non-controversial graphs where the two communities tend to overlap. For example, Figure 1 illustrates the visualizations of the *#gunsense* and *#ultralive* graphs (based on the structural modeling), created using the *ForceAtlas2 algorithm* [23] in *Gephi*.[14]

3.3 Echo Chamber Detection Metrics

In order to perform a comparison of echo chamber detection metrics on data belonging to different domains and from more than one social platform, those metrics were selected from the literature works that meet the following requirements:[15]

[14] https://gephi.org/

[15] Based on the established criteria, we provide here a brief overview of selected literature works illustrated in Section 2 but excluded from our analysis due to their dependence on platform-specific interactions, and/or topic-dependence, and/or reliance on external data sources. For example, the method proposed by Del Vicario et al. [14] determines user polarization based on a fraction of Facebook "like" interactions, which are not directly comparable with Twitter and Reddit interactions.

— *Independence from the topic of debate* on which the social network is centered;
— Use of *features common to multiple social platforms* that, with due consideration, can be considered comparable even if they belong to two different platforms;
— *Independence from external sources* in determining polarization. That is, to determine the orientation of a user given a certain topic is exploited solely on the content posted on the platform.

This choice led to the selection of 12 metrics divided into: *network-based metrics* (Section 3.3.1), *sentiment-based metrics* (Section 3.3.2), and *hybrid metrics* (Section 3.3.3). The first category includes metrics introduced within network-based and hybrid approaches for echo chamber detection, following the network-based pipeline proposed by Garimella et al. in [19]. In the second category are those metrics from content-based approaches that leverage Sentiment Analysis to find a state of controversy. Hybrid metrics combine both structural and semantic information to compute a final controversy score.

3.3.1 Network-based Metrics. These are the most widely used metrics in the literature for quantifying controversy. They are: *Boundary Connectivity* (BC), proposed by Guerra et al. [21], the *Dipole Moment* (DM) polarization index, proposed by Morales et al. [34], *Random Walk Controversy* (RWC), *Betweennees Centrality Controversy* (BCC), and *Embedding Controversy* (EC), proposed by Garimella et al. [19], *Displacement Random Walk Controversy* (DRWC) and *Authorative Random Walk Controversy* (ARWC), proposed by Villa et al. [45]. These metrics aim at quantifying controversy by assessing the level of segregation of clusters identified by the partitioning task. Below, we provide a more detailed description of each of these metrics.

Boundary Connectivity. This metric analyzes the structural patterns at the boundary between potentially polarized communities in a network. It focuses on assessing the degree of interaction and antagonism between these communities. The metric involves identifying *boundary nodes* (nodes connecting different communities) and *internal nodes* within the graph. It then computes a *polarization index*, which measures the extent to which boundary nodes prefer to connect with other internal nodes within the same community rather than nodes from the opposing community. Formally:

$$\mathrm{BC} = \frac{1}{|B|} \sum_{v \in B} \left[\frac{d_i(v)}{d_i(v) + d_B(v)} - 0.5 \right].$$

where $d_i(v)$ represent the count of edges connecting vertex u to internal vertices in a set I, and $d_B(v)$ is the number of edges linking vertex u to boundary vertices in a set B. The values of this metric range in the $[-0.5, 0.5]$ interval. A positive value indicates a preference for internal connections, suggesting potential antagonism between groups.

Dipole Moment. This polarization index draws inspiration from the physical concept of an *electric dipole*. In the context of a graph divided into two clusters representing opposing ideologies,

Similarly, Conover et al. [10] examine just political polarization using Twitter hashtags, violating the criterion of topic independence. The approach by Sriteja et al. [40] incorporates external sources, such as journalistic articles, introducing dependencies beyond the platform's native content. An et al. [3] evaluate partisan sharing on Facebook using a proposed metric that is inherently topic-specific, rendering it unsuitable as a generalizable measure. Concerning the approaches based on Stance Detection [5, 28], while this could be a valuable method to measure public and possibly polarized opinion on social media, it often depends on datasets specifically labeled for this purpose. In our study, however, we focus on comparing scenarios that do not require external resources at any level to evaluate the existence of echo chambers.

the polarization depends on the separation between these contrasting viewpoints. The polarization value, denoted as $R(u)$ ranges in the $[-1, 1]$ interval, and is assigned to each vertex $u \in V$. For the top 5% highest-degree vertices in each partition, R is set to ± 1, while *label propagation* determines R values for the remaining vertices. The index calculates the *normalized absolute difference*, namely ΔA, between the counts of vertices with positive and negative polarization (n^+ and n^-). It then computes the average polarization values, namely gc^+ and gc^+, among these vertices and defines d as half their absolute difference. Formally:

$$\text{DM} = (1 - \Delta A)d.$$

The DM index ranges in the $[0, 1]$ interval. The higher the value, the greater the separation between partitions, with larger differences in partition sizes contributing to lower index values.

Random Walk Controversy. It employs the concept of *random walks* on graphs to measure the likelihood of a node within a community being exposed to content from an *authoritative user* in the opposing community. In a partitioned graph $G = (V, E)$, with clusters X and Y, high-degree nodes representing authoritative users are selected from each group. Two random walks are initiated, one in each partition, terminating at authoritative nodes. The RWC score quantifies the difference in probabilities: (i) both walks remain in their starting partition (P_{XX}, P_{YY}), and (ii) both walks switch to the opposing partition (P_{XY}, P_{YX}). Hence:

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

RWC values range in the $[-1, 1]$ interval. The proximity to -1 indicates the probability for a random walk to end up in the opposite partition, signaling the absence of controversy, while the proximity to 1 indicates a high probability of remaining in the starting partition, thus indicating the presence of controversy. Since the RWC has proven to be particularly effective for echo chamber detection in [19], many recent approaches used it as a starting point to define new measures, including [16, 45].

Betweennees Centrality Controversy. This metric uses the concept of *edge betweenness* in a graph cut formed by two partitions, X and Y. In particular, to quantify the controversy, the *Kullback-Leibler* (KL) *divergence* [27] is computed between the distributions of edge betweenness in the *cut* and the rest of the graph. The hypothesis is that well-separated partitions result in a cut containing edges bridging structural holes, leading to higher betweenness values. Conversely, if partitions are not well-separated, the cut consists of strong ties, resulting in betweenness values similar to the rest of the graph. Formally:

$$\text{BCC} = 1 - e^{-d_{kl}}.$$

BCC values range in the $[0, 1]$ interval. The value is close to 0 in the case of low divergence (absence of controversy) and approaches 1 in the case of high divergence (presence of controversy).

Embedding Controversy. It is based on a *low-dimensional embedding* of a graph generated by the *ForceAtlas2* algorithm. Given the partitions X and Y, the *average embedded distances within each partition*, namely d_x and d_y, and the *average embedded distance across the partitions*, namely d_{xy}, are computed. Formally:

$$\text{EC} = 1 - \frac{d_x + d_y}{2d_{xy}}.$$

EC values range in the $[0, 1]$ interval. Values tend to approach 1 for controversial graphs, reflecting well-separated clusters and heightened controversy, while values nearing 0 suggest non-controversial graphs.

Displacement Random Walk Controversy. This metric assesses the level of controversy between communities by considering the ratio of steps leading to a change of community during a fixed-length random walk to the total walk length. If a node rarely changes its community during the walk, it suggests strong connections within its own community, signifying controversy between communities. Conversely, frequent transitions across communities imply lower controversy. Formally:

$$\text{DRWC} = \frac{\sum_{\forall v \in N} \left[1 - \left(\frac{n(v)_{cc}}{l_{rw}}\right)\right]}{|N|},$$

where $n(v)_{cc}$ represents the number of steps in the random walk of vertex v where the node changes community, N is the set of randomly selected vertices to be considered in computing the measure, and l_{rw} is the total length of the random walk. DRWC values range in the $[0, 1]$ interval. Higher values of this measure correspond to higher controversy between communities and vice versa.

Authoritative Random Walk Controversy. This metric is similar to the *Random Walk Controversy* (RWC) previously defined, with the key distinction that in ARWC, the random walks start exclusively from authoritative nodes.

3.3.2 Sentiment-based Metrics. Among the metrics taken by content-based approaches to echo chamber identification, those that meet the characteristics we set out to achieve are described in this section. In particular, they are based on the use of Sentiment Analysis to calculate some kind of ratio between positively and negatively polarized content in the social network.[16] Those considered in this article are the metrics from the study by Al-Ayyoub et al. [1], where their effectiveness was considered comparable to that of the *Dipole Moment* described in the previous section. Since they have been originally designed to be used on data from Twitter, these metrics refer to counts of tweets that are considered positive or negative, however, they can be generalized by matching a tweet to any textual unit. In the case of Reddit's data, the textual unit can represent a post or a reply comment. This simple repurposing allows the application of such metrics to data from any social platform. Specifically, the considered metrics are the *ratio of positive to negative sentiment scores counts* (PN), the *ratio between positive and negative sentiment scores count* (RPN), and the *ratio between positive to negative sentiment scores count of the ratio of positive and negative text unit count to the total number of text units* (PNPNT).

Ratio of positive to negative sentiment counts. This metric is based on the assumption that a controversial topic has an opposite side in terms of language tone used in the conversation and measures the ratio of positive sentiment count (i.e., the number of textual units, i.e., in our case either a tweet or a comment, with a positive sentiment) to negative count. Formally:

$$\text{PN} = \frac{(\textit{Positive sentiment count})}{(\textit{Negative sentiment count})}.$$

PN values range in the $[0, 1]$ interval. The larger the value, the higher the degree of controversy, and vice-versa.

Ratio between positive and negative sentiment counts. This metric is an extension of the PN metric. It calculates the ratio between the smaller sentiment counts (either positive or negative) and the

[16] As previously detailed, the sentiment of the textual units in this work was assigned using VADER [22], a domain-free model particularly well-suited for applications in the social context.

larger one. Formally:

$$\text{RPN} = \frac{\min(\textit{Positive sentiment count}, \textit{Negative sentiment count})}{\max(\textit{Positive sentiment count}, \textit{Negative sentiment count})}.$$

RPN values range in the $[0, 1]$ interval. Also in this case, larger values indicate a higher degree of controversy and vice-versa.

Ratio between positive to negative sentiment counts of the ratio of positive and negative sentiment counts to the total number of text units. In addition to considering PN, in the second part of the formula this metric is based on the assumption that a controversial topic has a large number of opposing tone textual units when compared with neutral textual units. Hence, it measures the ratio of the opposing number of textual units to the total number of textual units. Formally:

$$\text{PNPNT} = \text{PN} \cdot \frac{\textit{Positive sentiment count} + \textit{Negative sentiment count}}{\textit{Total text unit count}}.$$

Also PNPNT values range in the $[0, 1]$ interval, with the same semantics as above.

3.3.3 Hybrid Metrics. The hybrid metrics considered in this article are those proposed by Zarate et al. in [47] and Morini et al. in [35]. Specifically, we denote them as the *Zarate polarization index* (Z) and *echo chamber risk* (ECR).

Zarate polarization index. This metric utilizes *sentence-embedding techniques* to associate a vector with each user, encapsulating both syntactic and semantic properties of the published posts. In this work, we used *FastText* [24] for the text embedding task (both because of the speed of the training and because in [47] was found to be more effective than other models). Subsequently, we identified *central users* in the graph using the HITS algorithm [26] (corresponding to 30% of users with the highest *hub score* and 30% of users with the highest *authoritative score*), and utilized their embeddings to calculate the centroids of the network. With D_i as the sum of distances between the embeddings of central users and the centroids of their respective groups, and D_{glob} as the sum of distances from the global centroid, we finally computed the Z score as follows:

$$Z = \frac{D_1 + D_2}{D_{glob}},$$

and derived an inverse $\tilde{Z}$ score as

$$\tilde{Z} = 1 - Z.$$

$\tilde{Z}$ values range in the $[0, 1]$ interval, so that, similarly to the other metrics, higher values are associated with the presence of controversy and lower values with the absence of controversy.

Eco Chamber Risk. This metric does not refer to the global network, but it is associated with each identified community in the network and expresses the risk that it is identifiable as an echo chamber. To identify these communities, it is necessary to carry out an *inference step* about each user's orientation to controversy, which can be done through different approaches, based on supervised, unsupervised, or deep learning. In this work, we used the unsupervised approach (more details are described directly in Morini et al. [35]). Then, the risk of a community being an echo chamber is computed by considering both *purity* and *conductance*. Formally:

$$\text{ECR} = \textit{purity} - \textit{conductance}.$$

Purity measures how well users in a community belong to the same class or category. In contrast, *conductance* measures how well-defined a community is concerning its neighbors and the rest of the graph. In this work, we employed k-means to perform clustering on the posted texts, assigning each user a label based on the resulting groups. Subsequently, we partition the graph using the

Table 1. Results of Network-based Metrics

Social Graph	BC	DM	BCC	EC	RWC	ARW	DRWC
#gunsense	0.1777	0.5973	0.0485	0.4280	0.8524	0.8137	0.9851
#beefban	0.1852	0.6143	0.1025	0.4474	0.8810	0.4558	0.9879
#ukraine	0.1738	0.4617	0.0347	0.4241	0.8171	0.7151	0.9799
#ultralive	0.119	0.4273	0.0249	0.1833	0.6310	0.4512	0.9577
#nationalkissingday	0.0196	0.2618	0.0035	0.2055	0.3810	0.7329	0.7081
gun control	0.0663	0.0017	0.1778	0.1054	0.1200	0.1909	0.4646
minority discrimination	0.0015	0.0133	0.1016	0.0802	0.2128	0.2308	0.5778
political sphere	0.0597	0.0049	0.1410	0.0722	0.1448	0.2018	0.5115
vaccinations–Reddit	0.0723	0.2522	0.2068	0.1224	0.1093	0.2491	0.4487
vaccinations–Twitter	0.1653	0.0068	0.0285	0.1123	0.6494	0.6642	0.9430

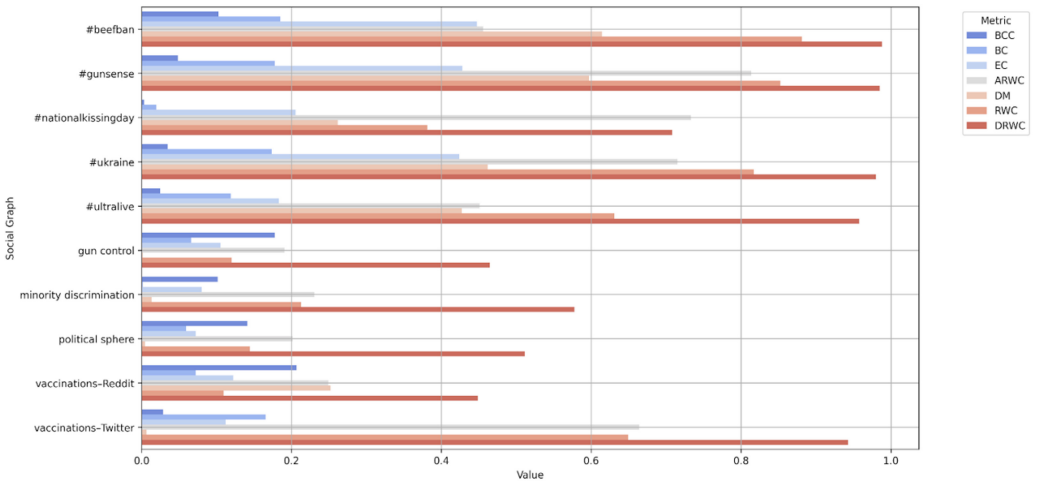


Fig. 2. Graphical representation of the results of network-based metrics.

Labeled Community Detection algorithm EVA [8], which considers both the structural aspects of the network and the label associated with each node. ECR values range in the [0, 1] interval. The higher the value obtained from the metric, the greater the probability that a cluster represents an echo chamber.

4 Comparative Evaluation

In this section, we present the results obtained by applying different echo chamber detection metrics to the Twitter and Reddit datasets illustrated in Section 3.1 (and, hence, their related social graphs). First, in Section 4.1, we assess the effectiveness of the application of network-based metrics on social graphs modeled considering only user interactions. In the same section, we further evaluate the same metrics on graphs that incorporate also semantic aspects directly into the graph modeling. Moving on to Section 4.2, we showcase the results of sentiment-based metrics. Lastly, in Section 4.3, we evaluate hybrid metrics that integrate both network structure and content information.

4.1 Network-based Metrics

The evaluation results obtained by each metric for echo chamber detection with respect to the considered social graphs are illustrated both numerically in Table 1 and visually in Figure 2.

Fig. 3. Correlation between network-based metrics.

The *retweet graphs* constructed from the data of Garimella et al. [19], namely *#gunsense, #beefban, #ukraine, #ultralive*, and *#nationalkissingday*, constitute our *ground truth* for evaluating the discriminatory power of each metric. The first three are indeed associated with controversy, while the last two are expected to be free of controversy. On these graphs, almost all metrics tend to have high values when controversy is present, and lower values for graphs without controversy. However, there are some anomalous cases; one of these is represented by the ARW metric, which indicates *#nationalkissingday* as more controversial than *#beefban*. This result suggests that, in the former network, users are more exposed to authoritative nodes of their partition, whereas in the latter case, users are exposed to authoritative nodes of the opposite partition. The DRWC metric takes values close to the upper extreme of its range on all *retweet* graphs, except for *#nationalkissingday*, which, despite having a lower value, is still considered a potentially controversial graph according to this metric. This result indicates that, in the graphs considered as *ground truth*, a change of partition in a random walk is rarely observed. On the other hand, the BC metric correctly assumes a very low value for the *#nationalkissingday* graph but considers *#ultralive* controversial almost to the same extent as *#ukraine*. For the remaining cases, the discriminative power of structural metrics on graphs belonging to the *ground truth* is confirmed, with most metrics identifying *#beefban* as the most controversial graph.

Analyzing the values obtained on graphs constructed using a different interaction mode of the social platform, notably the *comment graphs* of *Reddit* (*gun control, minority discrimination, political sphere,* and *vaccinations–Reddit*), very different results can be observed. Indeed, in the Reddit case, no metric seems capable of recognizing social graphs as polarized. The values of the metrics are, in some cases, even lower than those obtained on non-controversial graphs from the *ground truth*. Such a result suggests that structural metrics may not be suitable for identifying echo chambers on *Reddit*. The way debates are conducted on forums allows individuals with opposing ideologies to easily engage in the same thread. Therefore, modeling the network based on *comments* results in a network with many connections between people supporting opposing views. Structural metrics are not effective in such a context as they attempt to identify controversy based on the lack of links between the obtained communities. The only metric attributing higher values to *Reddit* graphs compared to *Twitter* graphs is BCC; this metric is negatively correlated with the others, as highlighted in Figure 3. The heatmap shows that the strongest correlations are between RWC and DRWC and between RWC and DM.

Table 2. Network-based Metric Results on Graphs with Content-enriched Modeling Based on *Sentiment Analysis*

Social Graph	BC	DM	BCC	EC	RWC	ARW	DRWC
gun control	0.0750	0.0091	0.2231	0.1117	0.2056	0.7181	0.5316
minority discrimination	0.0801	0.0732	0.0997	0.0859	0.3043	0.4865	0.6573
political sphere	0.1248	0.0334	0.0195	0.3352	0.5535	0.8609	0.8552
vaccinations–Reddit	0.0657	0.0086	0.2373	0.1174	0.1299	0.5000	0.4829
vaccinations–Twitter	0.1762	0.2317	0.0337	0.1015	0.6893	0.7000	0.9543

Table 3. Network-based Metric Results on Graphs with Content-enriched Modeling Based on *Topic Modeling*

Social Graph	BC	DM	BCC	EC	RWC	ARW	DRWC
gun control	0.0682	0.0023	0.1978	0.1074	0.1370	0.4200	0.4793
minority discrimination	0.0135	0.0301	0.0999	0.0810	0.2678	0.2740	0.5805
political sphere	0.0971	0.0126	0.1510	0.0901	0.5270	0.6812	0.5701
vaccinations–Reddit	0.0661	0.0187	0.2228	0.1213	0.1856	0.3760	0.4567
vaccinations–Twitter	0.1691	0.1074	0.0290	0.10912	0.6684	0.6867	0.9501

Table 4. Network-based Metric Results on Graphs with *Hybrid* Content-enriched Modeling

Social Graph	BC	DM	BCC	EC	RWC	ARW	DRWC
gun control	0.0852	0.0102	0.2345	0.1129	0.2178	0.7795	0.5576
minority discrimination	0.0861	0.0856	0.1078	0.0972	0.3875	0.5682	0.7842
political sphere	0.1463	0.0587	0.0206	0.3754	0.5967	0.8790	0.8854
vaccinations–Reddit	0.0745	0.0091	0.2742	0.1284	0.1307	0.5187	0.5140
vaccinations–Twitter	0.1954	0.2671	0.0791	0.1275	0.6914	0.7185	0.9671

For the *mention graph vaccinations–Twitter*, values are also lower compared to the graphs of the *ground truth*. The *mention* interaction mode is not directly linked to an endorsement as in the case of *retweets*. However, some metrics yield values comparable to those obtained from Garimella et al.'s data [19], such as BC, ARW, and DRWC.

Assessing the Impact of the Content-enriched Modeling on Network-based Metrics. Network-based metrics were also applied to social graphs modeled according to the *content-enriched modeling* (described in Section 3.2.2), which is based on the work by Villa et al. [45]. For obvious reasons, these results pertain only to graphs built on data equipped with textual content, namely *gun control*, *minority discrimination*, *political sphere*, *vaccinations–Reddit*, and *vaccinations–Twitter*. Tables 2-4 show the values assumed by the metrics on the content-enriched modeling based on *Sentiment Analysis*, *topic modeling*, and hybrid modeling that considers both aspects together. Corresponding graphical representations of the results are also provided in Figures 4–6.

Immediately noticeable is how content-enriched modeling leads to an increase in the values of structural metrics across all graphs, especially those constructed from *Reddit* data. In fact, the *vaccinations–Twitter* graph maintains approximately the same level of controversy as the baseline modeling for all considered metrics, while more significant changes are observed in the remaining cases. This aspect is illustrated by the *parallel plots* in Figures 7 and 8, which allow for an appreciation of metric differences across various modeling approaches for each graph.

It is also observed that the metrics primarily affected by the increase are those based on *random walk*, namely RWC, ARW, and DRWC, while the modeling associated with higher controversy values is the hybrid one, simultaneously considering both the discussed *topics* and *sentiment*.

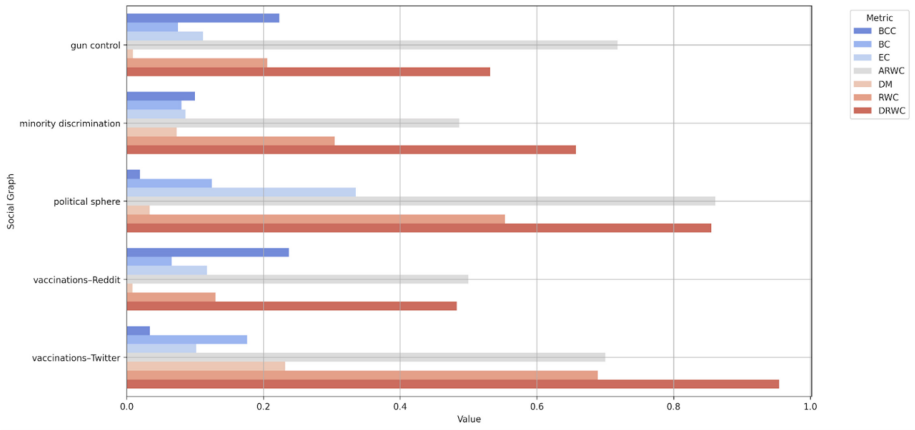


Fig. 4. Graphical representation of network-based metric results on graphs with content-enriched modeling based on *Sentiment Analysis*.

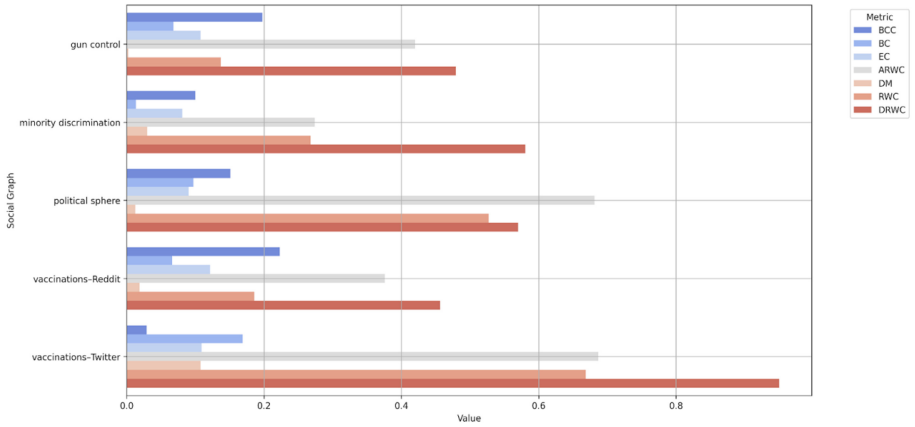


Fig. 5. Graphical representation of network-based metric results on graphs with content-enriched modeling based on *topic modeling*.

4.2 Sentiment-based Metrics

In Table 5, the results obtained through the metrics *ratio of positive to negative sentiment scores* (PN), *ratio between positive and negative sentiment scores count* (RPN), and *ratio between sentiment scores count of the ratio of positive and negative text unit counts to total number of text units* (PNPNT) are reported, considering that all of them are based on the same range of values in $[0, 1]$.

The first observation concerns the equality of values for the PN and RPN metrics. This happens because, for all considered datasets, the number of texts categorized with negative sentiment is always greater than those categorized as positive, leading to the equivalence of the two metrics.

Secondly, it is observed that the values obtained for the PNPNT metric are much lower than those of the two previous metrics. This result can be attributed to the fact that the metric considers, in addition to the number of positive and negative texts, also the neutral ones (based on an analysis of the considered datasets, the neutral sentiment constitutes the predominant category in all data). For this reason, it is believed that this metric is more effective than the other two in assessing

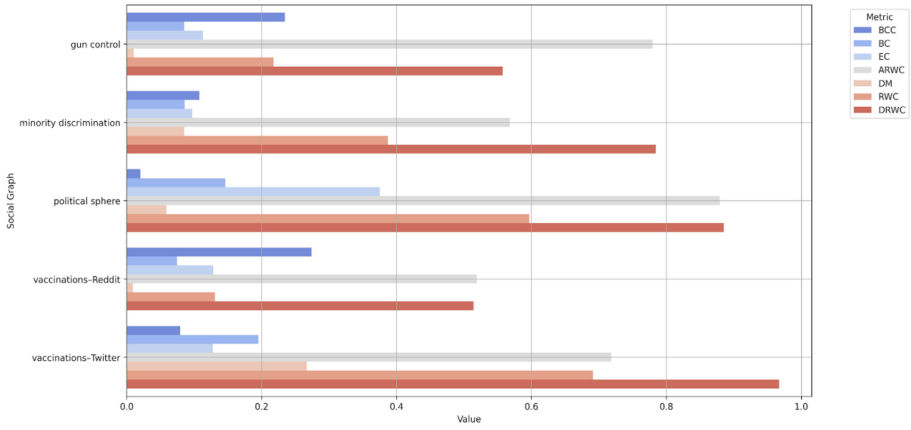


Fig. 6. Graphical representation of network-based metric results on graphs with *hybrid* content-enriched modeling.

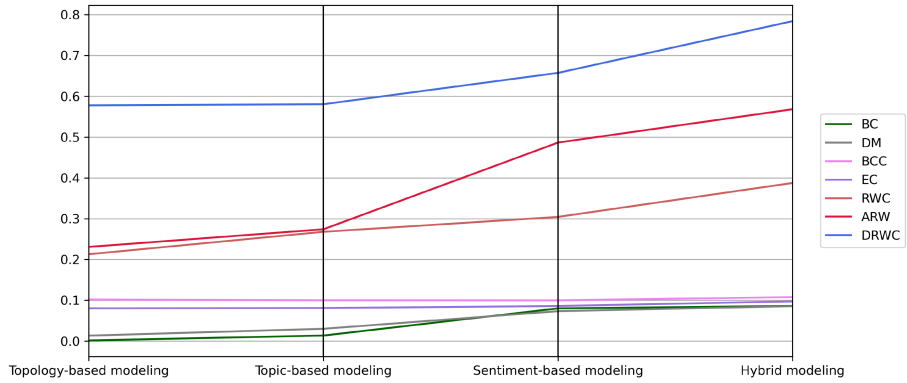


Fig. 7. Changes in network-based metric results across different graph modeling of the topic *minority discrimination*.

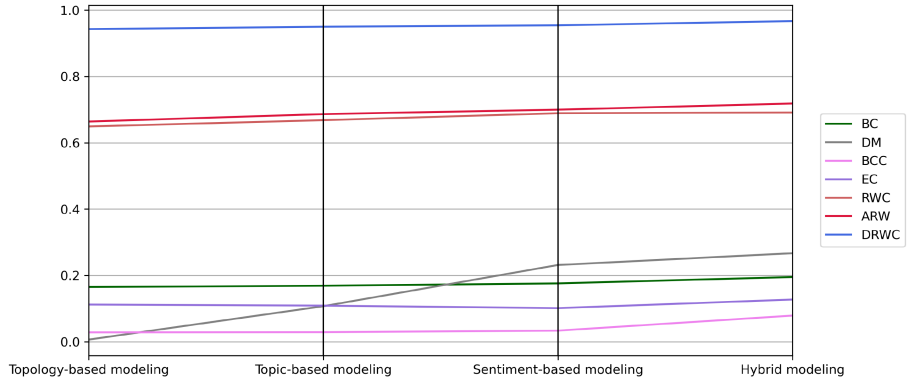


Fig. 8. Changes in network-based metric results across different graph modeling of the topic *vaccinations–Twitter*.

Table 5. Sentiment-based Metric Results

Social Graph	PN	RPN	PNPNT
gun control	0.5723	0.5723	0.2319
minority discrimination	0.9537	0.9537	0.3680
political sphere	0.8377	0.8377	0.3106
vaccinations–Reddit	0.9503	0.9503	0.3851
vaccinations–Twitter	0.9465	0.9465	0.3993

Table 6. Hybrid Metrics Results

Social Graph	$\tilde{Z}$	ECR1	ECR2
gun control	0.7748	0.5417	0.4887
minority discrimination	0.7560	0.6521	0.5578
political sphere	0.7512	0.4823	0.6182
vaccinations–Reddit	0.7924	0.8905	0.8621
vaccinations–Twitter	0.7576	0.9465	0.8310

the overall controversial state of a network. This is because it takes into account not only texts expressing opinions with more intense tones (both negative and positive) but also more moderate positions. In cases where the sentiment's proximity to the two extreme poles is less pronounced, as in the datasets considered here, the PN and PNT metrics may risk assigning a high level of controversy to networks where opposition to sentiment concerns only a minority portion of the network.

4.3 Hybrid Metrics

As illustrated in Section 3.3, hybrid metrics involve the polarization index derived by Zarate et al. [47], denoted as $\tilde{Z}$, and the metric proposed by Morini et al. [35], i.e., the Echo Chamber Risk, denoted as ECR, both defined in the range $[0, 1]$. Since Morini et al.'s metric [35] refers to a mesoscopic-scale approach that assigns a value to each echo chamber rather than to the network as a whole, the results for each graph are reported for the two largest echo chambers, denoting them as ECR1 and ECR2. As evident from Table 6, hybrid metrics allow for the identification of echo chambers in the graphs related to *Reddit* where structure-based metrics failed to capture the state of controversy.

Figure 9 illustrates, for each metric, the differences in values assumed by the graphs associated with the topics *vaccinations–Twitter* and *vaccinations–Reddit*. It is observed that, on the same topic, the most significant differences occur with respect to structural metrics, which, in most cases, attribute a higher level of controversy to the *vaccinations–Twitter* graph. For hybrid metrics, the difference is present to a lesser extent, while for sentiment-based metrics, the values for the two social graphs are nearly equal.

4.4 Overview Discussion on the Results

In the pursuit of identifying echo chambers, certain metrics have proven more effective than others depending on the application context/domain. Metrics belonging to network-based approaches have shown greater suitability for detecting controversy in datasets from Twitter. In particular, within this subset of data, a higher level of controversy was observed for graphs related to retweets, while on the *vaccinations–Twitter* graph modeled from mentions, lower values were obtained for most metrics. Some of these metrics, such as DM, BCC, and EC, do not effectively capture controversy on this graph.

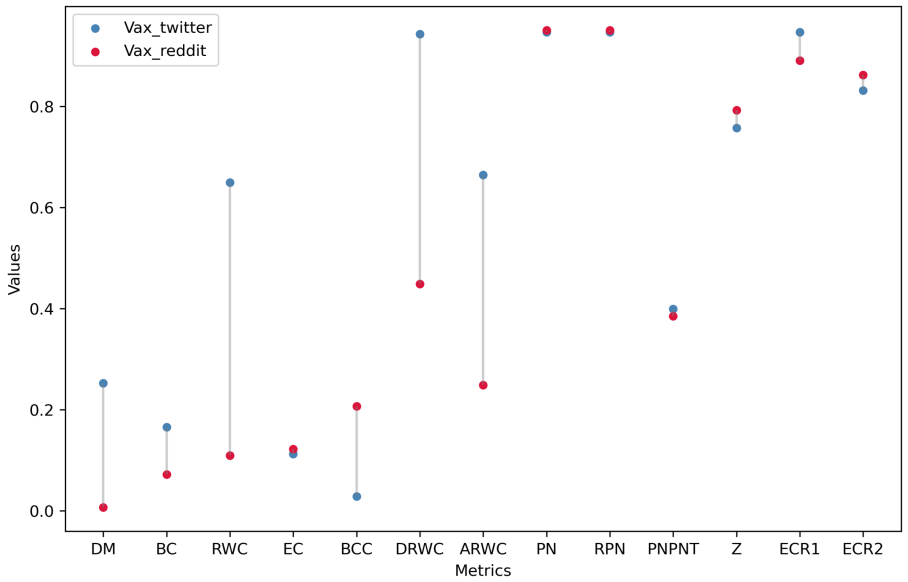


Fig. 9. Differences in metrics for the graphs *vaccinations–Twitter* and *vaccinations–Reddit*.

All network-based metrics indicate datasets from Reddit as non-controversial, demonstrating that network-based approaches, predominantly tested on Twitter in previous studies, may not generalize well to other platforms. The main reason lies in the assumption of network-based metrics that a direct link between two users represents strong ideological proximity. Therefore, they expect that in a graph with controversy, communities have dense connections internally but are connected to each other by a limited number of edges. This condition is met for retweet graphs capturing an endorsement dynamic but not for Reddit graphs modeled based on comments. Responding to a post or a comment in a subreddit does not necessarily indicate agreement with an opinion; instead, it can be used as a form of challenge or reply.

The integration of sentiment and discussed topics has increased the value of network-based metrics, but the best results on Reddit graphs are achieved with hybrid metrics. In particular, $\tilde{Z}$ and ECR metrics have proven effective in identifying echo chambers on both considered platforms. The scores obtained on the *vaccinations–Twitter* and *vaccinations–Reddit* graphs discussing the same topic were similar on both metrics. The ECR metric stands out as the most interesting among the proposed solutions, and its possible application in a meso-scale context allows evaluating each individual community as an echo chamber, disregarding areas of the network where the debate is less intense or characterized by a plurality of perspectives.

5 Conclusions and Further Research

This study examined and compared various metrics for identifying echo chambers in social networks proposed by distinct approaches in the literature so far. These approaches focus on both topological aspects of the network and the content being disseminated there, going so far as to model social graphs differently, which also impacts the proposed metrics. In addition, because there is a gap in the literature in a comparative perspective between platforms and domains, data from both Twitter and Reddit were used, which are currently the most accessible and versatile for gathering large volumes of data, including textual data, and are distinguished by similar user interaction patterns across multiple topics. In particular, Twitter datasets were identified by hashtags

related to specific events or controversial issues, while Reddit datasets were selected from known, highly polarized subreddits.

The results demonstrated that network-based metrics, especially those based on random walks, were effective in identifying echo chambers on Twitter, in particular on the "retweet" interaction network, highlighting the significant role of the network structure. However, these same metrics proved inadequate for other kinds of interaction networks, in particular in the Reddit context, where diverse social and communication dynamics prevent directly associating topological with ideological proximity. The use of alternative models involving semantic aspects led to a general improvement in structural metrics across all networks, particularly those related to Reddit. This suggests that content analysis is a key factor in identifying echo chambers on this platform. Sentiment-based and hybrid metrics, in particular, more clearly detected the presence of echo chambers in Reddit graphs.

There are aspects of the work that can be improved and further studied in the future. For example, the available data, especially those from the work by Garimella et al. [19], often used as a ground truth by many works in the literature, lacks textual content and can only be utilized for network-based metric calculations based on purely structural graph modeling. Furthermore, despite attempts to diversify the topic nature, there is room for expansion in terms of topics considered. Another constraint is the reliance on off-the-shelf techniques for tasks like graph partitioning, Sentiment Analysis, or clustering, with limited exploration of metric behavior in connection to a broader range of techniques. In particular, the study's partitioning phase focuses solely on the bipartite view of networks for consistency, but the meso-scale approaches of hybrid methodologies could be applied to contexts with more than two communities. Furthermore, the static treatment of echo chambers in this work, viewing networks as environments where polarization has already occurred, neglects dynamic considerations like how metric-recorded controversy may change with the introduction of new information into echo chambers, presenting an avenue for future exploration. Finally, a promising direction for future research could involve extending this study to other social media platforms, particularly Twitter-like but decentralized platforms such as Mastodon.[17] By focusing on such platforms, we could explore how decentralization and the absence of a single algorithmic feed shape user interactions, content dissemination, and the formation of polarized communities, offering deeper insights into the dynamics of online discourse.

Code and Data Availability

The code and data are made publicly available (in compliance with European GDPR, https://gdpr.eu/, regarding the release of data) at the following link: https://github.com/ikr3-lab/echo-chambers-twitter-reddit.

References

[1] M. Al-Ayyoub, A. Rabab'ah, Y. Jararweh, M.N. Al-Kabi, and B.B. Gupta. 2017. Studying the controversy in online crowds' interactions. *Applied Soft Computing* 66 (2017), 557–563.

[2] Abeer AlDayel and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Information Processing & Management* 58, 4 (2021), 102597.

[3] J. An, D. Quercia, and J. Crowcroft. 2014. Partisan sharing: Facebook evidence and societal consequences. In *Proceedings of the 2nd ACM Conference on Online Social Networks*. Dublin, Ireland, 13–24.

[4] K. Beelen, E. Kanoulas, and B. Velde. 2017. Detecting controversies in online news media. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval.*

[5] Fernando H. Calderón, Li-Kai Cheng, Ming-Jen Lin, Yen-Hao Huang, and Yi-Shin Chen. 2019. Content-based echo chamber detection on social media platforms. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 597–600.

[17] https://mastodon.social/

[6] Daejin Choi, Selin Chun, Hyunchul Oh, Jinyoung Han, and Ted "Taekyoung" Kwon. 2020. Rumor propagation is amplified by echo chambers in social media. *Scientific reports* 10, 1 (2020), 310.
[7] M. Cinelli, G. Francisci Morales, A. Galeazzi, W. Quattrociocchi, and M. Starnini. 2021. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences of the United States of America* 118, 9 (2021), e2023301118.
[8] S. Citraro and G.Eva Rossetti. 2019. Attribute-aware network segmentation. In *Proceedings of the International Conference on Complex Networks and Their Applications*. Lisbon, Portugal, 141–151.
[9] M. Coletto, K. Garimella, A. Gionis, and C. Lucchese. 2017. A Motif-Based Approach for Identifying Controversy. *International Conference on Weblogs and Social Media* 11, 1 (2017), 496–499.
[10] M.R. Conover, J. Ratkiewicz, M.R. Francisco, B. Gonçalves, F. Menczer, and A. Flammini. 2011. Political polarization on twitter. *International Conference on Weblogs and Social Media* 5, 1 (2011), 89–96.
[11] A. Cossard, G. D. F. Morales, K. Kalimeri, Y. Mejova, D. Paolotti, and M. Starnini. 2020. Falling into the echo chamber: The Italian vaccination debate on Twitter. In *Proceedings of the International AAAI Conference on Web and Social Media* 14 (2020), 130–140.
[12] Wesley Cota, Silvio C. Ferreira, Romualdo Pastor-Satorras, and Michele Starnini. 2019. Quantifying echo chamber effects in information spreading over political communication networks. *EPJ Data Science* 8, 1 (2019), 1–13.
[13] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi. 2015. Echo chambers in the age of misinformation. Retrieved from https://arxiv.org/abs/1509.00189
[14] M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H.E. Stanley, and W. Quattrociocchi. 2016. The spreading of misinformation online. *Proceedings of the National Academy of Sciences of the United States of America* 113, 3 (2016), 554–559.
[15] E. Dubois and G. Blank. 2018. The echo chamber is overstated: The moderating effect of political interest and diverse media. *Information, Communication and Society* 21, 5 (2018), 729–745.
[16] H. Emamgholizadeh, M. Nourizade, M. S. Tajbakhsh, et al. 2020. A framework for quantifying controversy of social network debates using attributed networks: Biased random walk (BRW). *Soc. Netw. Anal. Min.* 10, 90 (2020).
[17] Dieter Frey. 1986. Recent research on selective expo- sure to information. In *Proceedings of the Advances in Experimental Social Psychology* (2nd ed.). Elsevier, 41–80.
[18] N. E. Friedkin and E. Johnsen. 1990. Social influence and opinions. *The Journal of Mathematical Sociology* 15, 3–4 (1990), 193–206.
[19] K. Garimella, G. Francisci Morales, A. Gionis, and M. Mathioudakis. 2018. Quantifying controversy on social media. *ACM Transactions on Social Computing* 1, 1 (2018), 1–27.
[20] Steve Gregory. 2010. Finding overlapping communities in networks by label propagation. *New Journal of Physics* 12, 10 (2010), 103018.
[21] P. Guerra, W. Meira, C. Cardie, and R. Kleinberg. 2013. A measure of polarization on social media networks based on community boundaries. *International Conference on Weblogs and Social Media* 7, 1 (2013), 215–224.
[22] C. J. Hutto and E. Gilbert. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the 8th International Conference on Weblogs and Social Media*. 14.
[23] Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. 2014. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PloS One* 9, 6 (2014), e98679.
[24] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of Tricks for Efficient Text Classification. Retrieved from https://arxiv.org/abs/1607.01759
[25] George Karypis. 1997. METIS: Unstructured graph partitioning and sparse matrix ordering system. *Technical Report* (1997).
[26] J. M. Kleinberg. 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)* 46, 5 (1999), 604–632.
[27] Solomon Kullback. 1997. *Information Theory and Statistics*. Courier Corporation.
[28] Kuan-Chieh Lo, Shih-Chieh Dai, Aiping Xiong, Jing Jiang, and Lun-Wei Ku. 2021. Escape from an echo chamber. In *Proceedings of the Companion Proceedings of the Web Conference 2021*. 713–716.
[29] Jens Koed Madsen, Richard M. Bailey, and Toby D. Pilditch. 2018. Large networks of rational agents form persistent echo chambers. *Scientific Reports* 8, 1 (2018), 12391.
[30] A. Matakos, E. Terzi, and P. Tsaparas. 2017. Measuring and moderating opinion polarization in social networks. *Knowl Discov* 31, 5 (2017), 1480–1505.
[31] M. McPherson, L. Smith-Lovin, and J. M. Cook. 2001. Birds of a feather: Homophily in social networks. *Annual Review of Sociology* 27, 1 (2001), 415–444.
[32] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal* 5, 4 (2014), 1093–1113.

[33] Yelena Mejova, Amy X. Zhang, Nicholas Diakopoulos, and Carlos Castillo. 2014. Controversy and Sentiment in Online News. Retrieved from https://arxiv.org/abs/1409.8152
[34] A. J. Morales, J. Borondo, J. C. Losada, and R. M. Benito. 2015. Measuring political polarization: Twitter shows the two sides of Venezuela. *Chaos* 25, 3 (2015), 033114.
[35] V. Morini, L. Pollacci, and G. Rossetti. 2021. Toward a standard approach for echo chamber detection: Reddit case study. *Applied Sciences* 11, 12 (2021), 5390.
[36] R. S. Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology* 2, 2 (1998), 175–220.
[37] Eli Pariser. 2011. *The Filter Bubble: How the New Personalized Web is Changing What We Read and How We Think*. Penguin.
[38] Walter Quattrociocchi, Antonio Scala, and Cass R. Sunstein. 2016. Echo chambers on Facebook. *Available at SSRN 2795110* (2016).
[39] M. Rosvall, D. Axelsson, and C. T. Bergstrom. 2009. The map equation. *European Physical Journal-special Topics* 178, 1 (2009), 13–23.
[40] Allaparthi Sriteja, Prakhar Pandey, and Vikram Pudi. 2017. Controversy detection using reactions on social media. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, IEEE, 884–889.
[41] Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. In *Proceedings of the ICLR*.
[42] S. Terragni, E. Fersini, B. Galuzzi, P. Tropeano, and A. Candelieri. 2021. OCTIS: Comparing and optimizing topic models is simple!. In *EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the System Demonstrations*. 263–270. Association for Computational Linguistics (ACL)..
[43] José Van Dijck and Thomas Poell. 2013. Understanding social media logic. *Media and Communication* 1, 1 (2013), 2–14.
[44] José Van Dijck and Thomas Poell. 2015. Social media and the transformation of public space. *Social Media+ Society* 1, 2 (2015), 2056305115622482.
[45] G. Villa, G. Pasi, and M. Viviani. 2021. Echo chamber detection and analysis. *Soc. Netw. Anal. Min.* 11, 78 (2021).
[46] Claire Wardle and Hossein Derakhshan. 2017. *Information Disorder: Toward an Interdisciplinary Framework for Research and Policymaking*. Vol. 27. Council of Europe Strasbourg.
[47] J. M. O. Zarate, M. Giovanni, E. Feuerstein, M. Brambilla, J. M. O. EBoDe Zarate, M. Giovanni, E. Feuerstein, and M. Brambilla. 2020. Measuring controversy in social networks through NLP. In *Proceedings of the International Symposium on String Processing and Information Retrieval*. 194–209 pages. oks, 194–209..
[48] Jianming Zhu, Peikun Ni, Guangmo Tong, Guoqing Wang, and Jun Huang. 2021. Influence maximization problem with echo chamber effect in social network. *IEEE Transactions on Computational Social Systems* 8, 5 (2021), 1163–1171.

Received 31 January 2024; revised 11 September 2024; accepted 24 October 2024