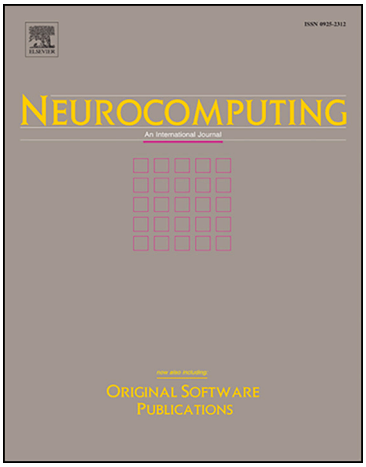

Ideological isolation in online social networks: A survey of computational definitions, metrics, and mitigation

Xiaodan Wang, Yanbin Liu, Shiqing Wu, Ziying Zhao, Yuxuan Hu, Weihua Li, Quan Bai

PII: S0925-2312(26)01731-5
DOI: https://doi.org/10.1016/j.neucom.2026.134333
Reference: NEUCOM 134333

To appear in: *Neurocomputing*

Received Date: 31 January 2026
Revised Date: 9 June 2026
Accepted Date: 21 June 2026

Please cite this article as: Wang X, Liu Y, Wu S, Zhao Z, Hu Y, Li W, Bai Q, Ideological isolation in online social networks: A survey of computational definitions, metrics, and mitigation, *Neurocomputing* (2026), doi: https://doi.org/10.1016/j.neucom.2026.134333.

Ideological Isolation in Online Social Networks: A Survey of Computational Definitions, Metrics, and Mitigation

Xiaodan Wang[a,b], Yanbin Liu[b], Shiqing Wu[c], Ziying Zhao[b], Yuxuan Hu[e], Weihua Li[b,*], Quan Bai[d,**]

[a]*School of Engineering, Yanbian University, Yanji, 133002, Jilin Province, China*
[b]*Department of Data Science and Artificial Intelligence, Auckland University of Technology, Auckland, 1010, New Zealand*
[c]*Faculty of Data Science, City University of Macau, Macau SAR, China*
[d]*School of Information and Communication Technology, University of Tasmania, Hobart, 7001, Tasmania, Australia*
[e]*Integrated Marine Observing System, University of Tasmania, Hobart, 7004, Tasmania, Australia*

Abstract

Ideological isolation in online social networks, including selective exposure, echo chambers, filter bubbles, tunnel vision, and polarization, has become a central concern for computational and neural modeling of information ecosystems. With the rapid adoption of graph learning, representation learning, and feedback-driven recommender systems, a growing body of work has proposed diverse metrics and models to quantify and mitigate these phenomena. However, existing studies and surveys rely on heterogeneous definitions and incompatible measurements, making empirical findings difficult to compare and obscuring how different forms of ideological isolation arise in learning-based systems. This survey provides a computationally grounded and comprehensive review of existing approaches to defining, analyzing, measuring, and mitigating ideological isolation in online social networks. We examine the mechanisms underlying content personalization, user behavior, and network structure that drive exposure concentration and attention narrowing. We then systematically review methodological approaches for detecting and quantifying ideological isolation, covering network-, content-, and behavior-based metrics, and synthesize empirical findings across platforms to assess their applicability and limitations. We further organize computational mitigation strategies, including network-topological interventions and recommendation-level controls, compare mitigation families and their trade-offs, and examine the dual role of large language models. The key ethical considerations in the design and deployment of diversity-aware systems are also discussed. By resolving the definition-metric-intervention mismatch that characterizes existing work, this survey provides a principled foundation for the design, evaluation, and deployment of neural and learning-based systems aimed at diagnosing and mitigating ideological isolation in online social networks.

Keywords: Online Social Networks, Ideological Isolation, Selective Exposure, Filter Bubble, Echo Chamber, Recommender Systems

1. Introduction

Online social networks have become the primary infrastructure for producing, filtering, and consuming information at scale. Modern platforms rely heavily on learning-based systems to personalize content exposure and maximize engagement. While these systems enable efficient information access, they also reshape exposure dynamics, potentially concentrating attention on limited subsets of viewpoints [28]. As a result, users may increasingly encounter ideologically aligned content and like-minded peers, while rarely engaging with dissenting views. This condition, referred to as *ideological isolation* [58], manifests through related phenomena including filter bubbles [7, 36], echo chambers [102, 138], tunnel vision [30], exposure bias [69], and polarization [35], which jointly shape how beliefs form, reinforce, and spread online.

Ideological isolation poses several intertwined challenges. First, information overload and selection make it difficult to maintain balanced exposure under recommendation and limited attention [83, 107, 142]. Second, network structure (e.g., modular communities, homophily) shapes the pathways and speed of cross-cutting information flow, often reinforcing within-group interaction [139]. Third, algorithmic filtering and amplification can magnify alignment signals derived from past behavior, creating self-reinforcing feedback loops [28, 142]. Fourth, cognitive and social dynamics, including selective exposure, confirmation bias, and the false-consensus effect, encourage users to favor attitude-consistent content and avoid contrary perspectives [107]. Finally, external amplifiers, such as coordinated campaigns or bots, can distort exposure diversity [86, 151, 162]. Addressing these challenges requires an interdisciplinary approach that connects computational modeling, measurement, and model design.

Over the past decade, these issues have motivated a

*Corresponding author: weihua.li@aut.ac.nz
**Corresponding author: quan.bai@utas.edu.au

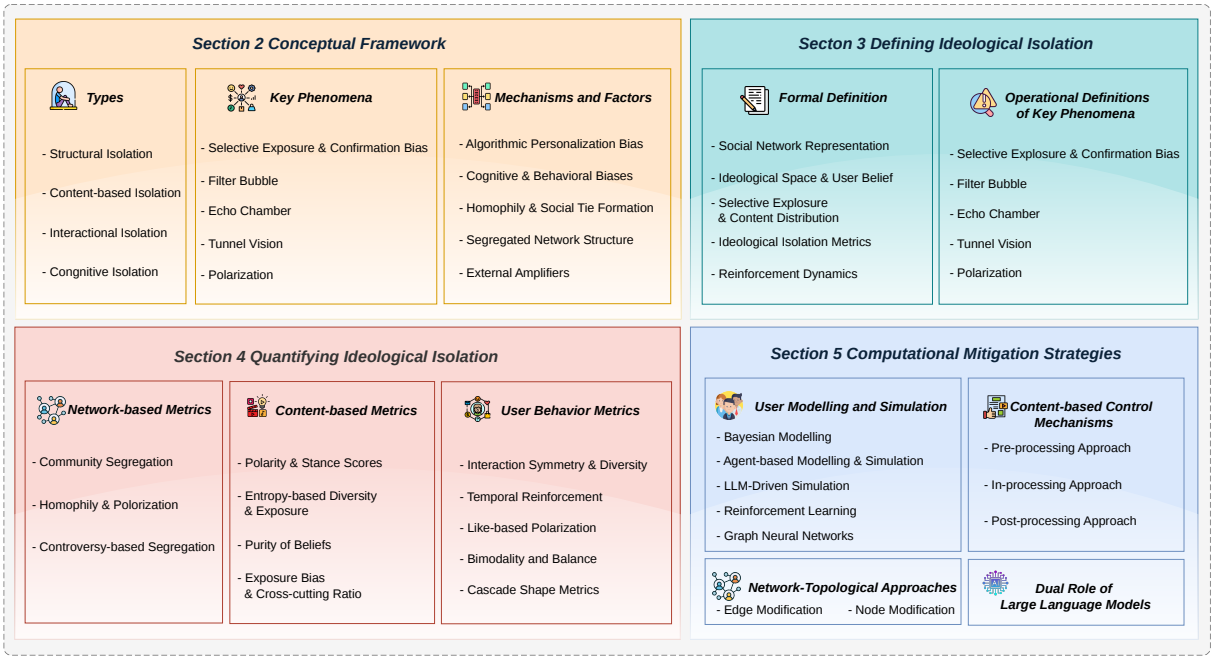

Figure 1: Overview of ideological isolation in online social networks. The figure illustrates the structure of this survey, spanning conceptual framing, formal definition, quantitative measurement, and computational mitigation.

growing body of research [65, 127] across computer science, information science, and the social sciences. However, existing work remains fragmented, with inconsistent definitions, heterogeneous measurements, and mitigation strategies scattered across graph-based approaches and recommendation system adjustments. In this survey, we provide a computationally grounded methodological synthesis of ideological isolation. We organize the literature around (i) a compact conceptual framework that distinguishes four interrelated isolation types; (ii) formal definitions that place users and content in a shared ideological space and specify visibility, exposure, and reinforcement dynamics; (iii) operational metrics spanning network, content, and behavior indicators; and (iv) computational mitigation strategies at both the network topology and recommendation algorithm levels.

To provide an end-to-end view of ideological isolation, we structure the paper as shown in Figure 1:

1. **Conceptual Framework (Section 2)**. We introduce a unified taxonomy of ideological isolation comprising four types: structural, content-based, interactional, and cognitive, and organize key phenomena (selective exposure & confirmation, filter bubbles, echo chambers, tunnel vision, and polarization) within this framework. We further analyze the computational, network, and behavioral mechanisms that drive these phenomena, including algorithmic personalization, cognitive and behavioral biases, homophily and tie formation, network segregation, and external amplifiers. Figure 1 summarizes these relationships.
2. **Definitions and Problem Formulations (Section3)**. We develop a formal computational model of ideological isolation by representing users, content, and exposure in a shared ideological space and specifying visibility, exposure, and reinforcement dynamics, with quantitative isolation indices. Building on these formal foundations, we provide operational definitions and precise conditions for key phenomena associated with ideological isolation.
3. **Quantifying Ideological Isolation (Section 4)**. We review network-based (e.g., segregation, homophily, and controversy-oriented measures), content-based (e.g., polarity, diversity, and exposure-related measures), and behavior-based (e.g., interaction, temporal, and cascade-based measures) metrics, and harmonize notation to support comparability across methods.
4. **Mitigation Strategies (Section 5)**. We review computational mitigation strategies, including user modeling and simulation (e.g., agent-based and learning-based models), network-topological interventions (e.g., edge and node modification), and content-based control (e.g., pre-, in-, and post-processing). We further analyse the dual role of large language models as both drivers and tools for mitigation, and provide a qualitative cross-family comparison. We discuss trade-offs between accuracy, utility, and exposure diversity.
5. **Ethical Considerations (Section 6)**. We analyze the ethical dimension of computational mitigation, covering algorithmic paternalism, user authorization

and transparency, and cross-cultural and regulatory variation in what "ideological diversity" means in practice.

6. **Research Challenges and Directions (Section 7)**. We outline key open problems, including cross-platform measurement, causal identification under exposure bias, human-in-the-loop evaluation, robustness to adversarial amplification, and practical deployment constraints for diversity-aware systems.

Differences from Existing Surveys. Existing surveys largely treat these themes in isolation, focusing on recommender-system phenomena such as echo chambers [51], filter-bubble dynamics [7], and exposure bias [69], but lack an integrated computational framework that connects formal definitions, measurement, and mitigation. In contrast, this survey provides: (i) a concept-to-operations bridge that maps isolation types to concrete metrics and intervention strategies; (ii) a unified formalism for exposure, diversity, and reinforcement that aligns network, content, and behavior dimensions; and (iii) a mitigation taxonomy that spans both network topology and recommendation pipelines with explicit measurement targets.

Main Contributions. The contributions of this survey are summarized as follows:

- **Unified Conceptual Framework.** We introduce a compact taxonomy of ideological isolation comprising structural, content-based, interactional, and cognitive types, and relate them to their underlying computational, network, and behavioral mechanisms.
- **Formal Computational Model.** We develop unified formal definitions of content visibility, exposure centroids and variance, and reinforcement dynamics, and specify operational conditions for ideological isolation phenomena, including selective exposure, filter bubbles, echo chambers, tunnel vision, and polarization.
- **Principled Quantification.** We systematically organize network-, content-, and behavior-based metrics with harmonized notation and guidance for interpretation and normalization, and synthesize published findings across major English-, Chinese-, and Indian-language platforms to indicate which metric families best diagnose each isolation type.
- **Mitigation Design Space.** We structure computational mitigation strategies at the network-topological and recommendation-system levels, analyzing trade-offs and deployment considerations for increasing cross-cutting exposure and semantic diversity, and compare across the mitigation families.
- **Ethical Considerations.** We connect computational mitigation to its normative implications, including algorithmic paternalism, user authorization and transparency, and cross-cultural and regulatory variation in what ideological diversity means in practice.

2. Conceptual Framework

This section presents a unifying framework for studying ideological isolation that integrates key phenomena, types, and mechanisms and clarifies their relationships. We begin with the phenomena observed in practice, organize them into types of ideological isolation that provide a coherent taxonomy for analysis, and then examine the mechanisms and factors that give rise to these types and drive the observed phenomena.

2.1. Key Phenomena

Selective exposure and confirmation bias. Individuals tend to prefer belief-congruent information; on modern platforms, however, this tendency is co-produced with feed design. Large-scale evidence from Facebook shows that user choices and ranking algorithms reduce cross-cutting exposure, yielding ideologically aligned consumption even when diverse items are available [9]. At the same time, the recommender-systems literature emphasizes that exposure (what is shown) is distinct from preference (what is liked or clicked), and that feedback loops in exposure can entrench visibility imbalances and perceived consensus, an agenda now formalized under fairness-aware recommendation [69]. Moving beyond descriptive correlations, recent causal-inference approaches explicitly model item exposure and user satisfaction as separate processes to debias learning from logs, helping disentangle human confirmation tendencies from system-induced selection effects [94].

Filter Bubble. Filter bubbles are commonly described as an exposure environment shaped by personalization in which the breadth of viewpoints presented to a user narrows over time [18]. Empirically, web-scale tracking shows that platform pathways (e.g., social referrals and search) can increase the ideological distance of news exposure, indicating that bubbles are primarily an exposure-level phenomenon rather than merely a matter of individual preference [36]. At the same time, the recommender-systems literature cautions that evidence is mixed across platforms and metrics: systematic reviews report both indications of narrowing exposure and counter-examples where recommendation pipelines maintain diversity, underscoring the need to distinguish what is shown from what is preferred and to measure diversity explicitly [7]. Finally, user-interface controls that allow users to adjust system curation can increase awareness of bubble effects and reduce extremity, although their impact on political diversity remains mixed [96]. Recent simulation-based work further disentangles filter bubbles from homogenization by decomposing their effects into inter-user and intra-user diversity.

It shows that traditional recommendation algorithms primarily influence filter bubbles through inter-user diversity, while having a limited impact on intra-user diversity [6].

Echo Chamber. Echo chambers are generally understood as social contexts characterized by dense connections among ideologically like-minded users, where content exposure is largely confined within the community, leading to high source concentration and limited cross-community exchange. Comparative reviews show that this phenomenon is typically examined through the network lens (e.g., homophily and community structure) and interaction patterns, although evidence for its prevalence and effects varies across platforms and measurement choices [138]. Recent syntheses map the field's core dimensions, attributes, mechanisms, modeling and detection approaches, and metrics, and highlight the absence of a single canonical definition, underscoring the need to specify how echo chambers are operationalized in a given study [26]. A recent systematic review attributes disagreements about existence and impact to divergent conceptualizations and operationalizations (e.g., exposure, topology, or attitudes), showing that empirical conclusions depend strongly on the chosen measurement strategy and context [51].

Tunnel Vision. Tunnel vision in social networks refers to the aspect-level narrowing of online discourse, where collective attention converges on a limited subset of an issue's facets while alternative angles and context are sidelined, reducing informational diversity [150]. Unlike filter bubbles or echo chambers, which emphasize the sources of content and ideological alignment, tunnel vision emphasizes what is being discussed, i.e., a content-level concentration of frames or aspects, even in the absence of explicit community segregation.

Polarization. Polarization denotes the separation of audiences or communities into distinct, often opposing, camps with limited middle ground. In practice, polarization manifests as multi-modal distributions of opinions and stances on controversial topics, as well as fragmented communities with sparse bridging ties. Distributional analyses of controversies on social media reveal clear splits into discernible camps, providing a content-level signature of polarization [125]. Unlike the other phenomena, polarization is an emergent, system-level outcome, typically characterized at the population or network scale [63].

2.2. Types of Ideological Isolation

Following the overview of key phenomena in Section 2.1, this subsection introduces four interrelated types of ideological isolation: structural, content-based, interactional, and cognitive. These types are not mutually exclusive and provide a conceptual basis for organizing observed phenomena and connecting measurement and mitigation in later sections. As clarified below and depicted in Figure 2, these mappings are not one-to-one: a single phenomenon may span multiple types, and multiple phenomena may contribute to the same type, reflecting the interconnected nature of ideological isolation.

Structural Isolation. Users tend to reside within segregated or highly modular clusters in the social network. This form of isolation is primarily associated with echo chambers, which arise from dense intra-community connectivity that restricts cross-group exposure. It is also closely related to polarization, particularly when the network fragments into distinct and opposing camps. Moreover, AI-driven personalization can further intensify structural isolation by shaping interaction pathways through largely invisible algorithmic structures.

Content-based Isolation. This type is most directly linked to filter bubbles, as algorithmically curated feeds narrow the ideological breadth of content a user sees, and to tunnel vision, through the discourse-level narrowing of the topics and aspects users encounter. Users are consistently exposed to ideologically skewed, repetitive, or redundant content. This type of isolation is closely related to *filter bubbles*, and the cognitive narrowing described by *tunnel vision*. Both algorithmic ranking and user preferences reinforce such exposure patterns.

Interactional Isolation. Interactional isolation captures the behavioral dimension of ideological separation. It is primarily associated with echo chambers, where interactions are confined among like-minded peers. Selective exposure and confirmation bias also fall within this type, as both reflect homophilic engagement patterns that limit exposure to dissenting views. Polarization contributes insofar as cross-cutting engagement declines between opposing groups. Collectively, network structure and user behavior jointly reduce ideological diversity in interactions, with ties formed mainly among like-minded individuals.

Cognitive Isolation. Cognitive isolation captures the psychological deepening of ideological separation. It is primarily associated with selective exposure and confirmation bias, reflecting the tendency to favor belief-consistent information and avoid opposing views. Tunnel vision also belongs to this type, as sustained attention narrowing reduces openness to alternative perspectives at the cognitive level. Additionally, the cognitive outcomes of echo chambers also contribute to this type, as repeated like-minded peer exposure deepens confirmation bias and inflates perceived consensus.

2.3. Mechanisms and Contributing Factors of Ideological Isolation

Having outlined the key phenomena and types of ideological isolation, we now examine the mechanisms and contributing factors that generate and reinforce these patterns in online social networks.

Algorithmic Personalization Bias. AI-powered recommender systems selectively curate content based on prior engagement [158], inevitably reinforcing filter bubbles and reducing content diversity [7, 58, 148]. Algorithmic bias and cognitive bias are key contributors to excessive personalization in online social networks, particularly in recommender systems. Algorithmic bias is introduced during system design and implementation, while cognitive bias (e.g., confirmation bias) influences user interactions and feedback loops [47]. Although fairness and diversity have received increasing attention in recommender-system research, most systems still emphasize personalization, thereby contributing to ideological isolation [7, 69]. Simulation studies show that algorithmic curation and selective exposure can fragment public discourse into ideological "tunnels" [28]. Model-to-data analyses further indicate that interactions between algorithmic curation and user selectivity can reproduce and amplify polarized and fragmented information landscapes [142].

Cognitive and Behavioral Biases. Users are inclined toward selective exposure and confirmation bias, seeking attitude-consistent information while avoiding opposing views [74]. Echo chambers can restrict users to homogeneous content, driven jointly by network homophily and structure, algorithmic curation, and selective engagement [51], thereby fostering biases such as the false consensus effect (FCE), in which individuals overestimate how much others share their views [100]. Luzsa and Mayr show that higher FCE is associated with prolonged social media use, homogeneous networks, and lower ambiguity tolerance [101]. Additionally, individuals may adopt group-conforming opinions either to reduce cognitive dissonance or to gain social acceptance [63], a phenomenon that has also been encoded in opinion-aware influence maximization models [27]. The lack of dissenting views in users' feeds reduces critical scrutiny and encourages further ideological rigidity [101].

Homophily and Social Tie Formation. Individuals tend to form connections with others who share similar beliefs, leading to interactional isolation and the formation of echo chambers. Simulation models demonstrate that users may sever ties with disagreeing peers, thereby reinforcing homophily [71]. This behavior can be incentivized in polarized environments, where adopting provocative or oversimplified messages attracts greater attention and followers [63, 122]. As like-minded agents cluster within echo chambers, attempts to shift user preferences often fail due to entrenched social structures [145].

Segregated Network Structure. Community clusters with high modularity restrict cross-group information flow, exacerbating topological isolation [116]. Ferraz de Arruda et al. proposed a framework that simulates how algorithmic filtering and homophilic interactions amplify opinion polarization. They validated the model on real Twitter networks, demonstrating that such systems fragment discourse into ideologically segregated clusters [28]. These network structures limit exposure to diverse viewpoints and reinforce ideological rigidity.

External Amplifiers. Social bots, coordinated campaigns, and misinformation introduce artificial signals and distort information diversity, deepening ideological silos. In particular, coordinated information tactics, such as bots or sponsored content, can exacerbate selective exposure and algorithmic bias, making it harder for users to access diverse, high-quality information [29, 62, 132]. These amplifiers operate synergistically with platform algorithms and social dynamics to distort public discourse, creating environments in which ideologically isolated communities can thrive. In addition to these mechanisms, some studies examine content creation and reception as dual filtering pathways that drive tunnel vision [4]. Content is shaped by subjective values, organizational rules, and external incentives, such as market or political pressures. Meanwhile, users may actively avoid content they did not select, further fragmenting shared discourse and undermining common ground [4].

While the mechanisms above are presented individually, empirical evidence suggests that they frequently co-occur and interact. For instance, algorithmic personalization amplifies the effects of cognitive biases by curating feeds that satisfy users' preference for belief-consistent content, creating a feedback loop between system-level filtering and individual-level selective exposure [28, 142]. Similarly, homophily-driven tie formation strengthens segregated network structures, as users who preferentially connect with like-minded peers contribute to the very modular topology that restricts cross-group information flow [139]. External amplifiers, such as social bots, exploit these existing dynamics by injecting content into already homophilic clusters, thereby magnifying both algorithmic and behavioral biases [132]. Quantitative evidence for individual mechanisms varies in maturity: algorithmic personalization effects have been measured through large-scale controlled experiments [9, 56], homophily and segregation are routinely quantified via modularity and assortativity indices (see Section 4.1), and cognitive biases have been assessed through survey-based and behavioral studies [75, 101]. However, disentangling the relative contributions of these mechanisms and their interaction effects in observational settings remains an open challenge, which we revisit in Section 7.

Beyond the five mechanisms above, recent work has identified large language models (LLMs) as a distinct and increasingly influential driver of ideological isolation. Widely deployed conversational LLMs exhibit measurable political bias [112, 128, 129], which can propagate at scale when the same models are used to draft, summarize, or rewrite political content. Persona-conditioned generation has been shown to induce affective polarization and produce filter-bubble-like outputs even in single-turn inter-

actions [80], while LLM-agent-based simulations of short-video recommender loops reproduce the temporal narrowing characteristic of filter bubbles [135]. Because LLMs can also be deployed as mitigation tools, we treat their dual role as a distinct topic in Section 5.4.

In summary, this section establishes a conceptual framework for describing ideological isolation through observable phenomena and organizing it into four interrelated types. As illustrated in Figure 2, upstream mechanisms, including algorithmic personalization, cognitive and behavioral biases, homophily and social tie formation, segregated network structures, and external amplifiers, give rise to selective exposure and confirmation bias, filter bubbles, echo chambers, tunnel vision, and polarization. These phenomena, in turn, correspond to structural, content-based, interactional, and cognitive forms of ideological isolation. This layered mapping from mechanisms to phenomena to types underpins the remainder of the survey.

3. Defining Ideological Isolation

This section formalizes ideological isolation as a computational construct. We define the networked setting, the ideological space for beliefs and content, and the exposure process that generates user feeds. On this basis, we introduce an isolation index that combines belief–exposure alignment with exposure diversity, along with a reinforcement dynamic for belief updating under repeated exposure. We then state operational criteria for five phenomena: selective exposure, filter bubbles, echo chambers, tunnel vision, and polarization, enabling consistent diagnosis across datasets and platforms.

3.1. Formal Definition

In this subsection, we present formal definitions of ideological isolation, focusing on its structural and dynamic properties in online social networks. We introduce key concepts, including user belief representations, content exposure profiles, and network interaction patterns, along with their associated notations and metrics. These formal elements provide a unified lens for capturing related phenomena such as filter bubbles, echo chambers, and tunnel vision, enabling quantitative analysis of how algorithmic filtering and social interactions shape the evolution of ideological landscapes.

3.1.1. Social Network Representation

Let the online social network be represented as a directed graph:

$$G = (U, E),$$

where $U = \{u_1, u_2, \ldots, u_N\}$ is the set of users, and $E \subseteq U \times U$ is the set of directed edges representing social relationships, such as follower/followee, friend, and subscription [98].

For each user $u_i \in U$, we define $\Gamma(u_i) = \{u_j \in U \mid (u_j, u_i) \in E\}$ as the set of in-neighbors, i.e., users whose content is visible to u_i. Depending on the platform's structure, this may reflect incoming edges, e.g., followees on X or friends whose posts appear in WeChat Moments. This neighborhood plays a central role in shaping user exposure on platforms that do not rely heavily on algorithmic recommendations, such as WeChat, where the content feed is largely provided by $\Gamma(u_i)$, occasionally interleaved with advertisements or sponsored content [87].

Meanwhile, each user u_i maintains a content feed $\mathcal{F}_i(t)$ at time t, which is shaped by the underlying logic of the platform. On many platforms, such as TikTok and YouTube, this feed is curated by recommender systems that leverage prior engagement history, social proximity, and behavioral signals to prioritize content [56, 144]. However, on other platforms, such as WeChat, $\mathcal{F}_i(t)$ primarily consists of content posted by the user's social connections (e.g., friends), with limited algorithmic intervention apart from occasional advertisements [169]. In such cases, the structure of the user's social network itself plays a dominant role in shaping ideological exposure, reinforcing the effects of homophily and social filtering over algorithmic personalization.

3.1.2. Ideological Space and User Beliefs

We define a continuous ideological space $\mathcal{I} \subset \mathbb{R}^d$ to represent the latent ideological dimensions along which both users and content can be situated. Each user $u_i \in U$ is associated with an ideological belief vector $\mathbf{b}_i \in \mathcal{I}$, which reflects their political orientation, social values, or other issue-based stances [85]. The dimension d can vary depending on the granularity of ideological representation. For example, $d = 1$ may represent a left-right political spectrum, while $d > 1$ allows modeling multiple orthogonal ideologies (e.g., economic, social, environmental) [87]. Similarly, let $\mathcal{C} = \{c_1, c_2, \ldots, c_M\}$ denote the global set of content items posted on the platform. Each item $c_j \in \mathcal{C}$ may represent a post, comment, video, or article. At any time t, each user u_i is exposed to a subset of these content items, forming their personalized content feed $\mathcal{F}_i(t) \subseteq \mathcal{C}$.

Each content item $c_j \in \mathcal{C}$ is embedded in the same ideological space and is represented by a vector $\mathbf{v}_j \in \mathcal{I}$. These embeddings can be derived from textual content, metadata, creator attributes, or user engagement patterns using machine learning models, such as topic modeling, sentiment-aware embeddings, or supervised classifiers trained on ideological annotations [111]. To formalize ideological alignment, we define a content-user affinity function $\phi : \mathcal{I} \times \mathcal{I} \to \mathbb{R}$ that measures the proximity or relevance of content to a user's beliefs. A simple way is the negative Euclidean distance:

$$\phi(\mathbf{b}_i, \mathbf{v}_j) = -\| \mathbf{b}_i - \mathbf{v}_j \|, \tag{1}$$

where $\|\cdot\|$ denotes the Euclidean norm, so that ideologically closer content receives higher affinity scores. Alter-

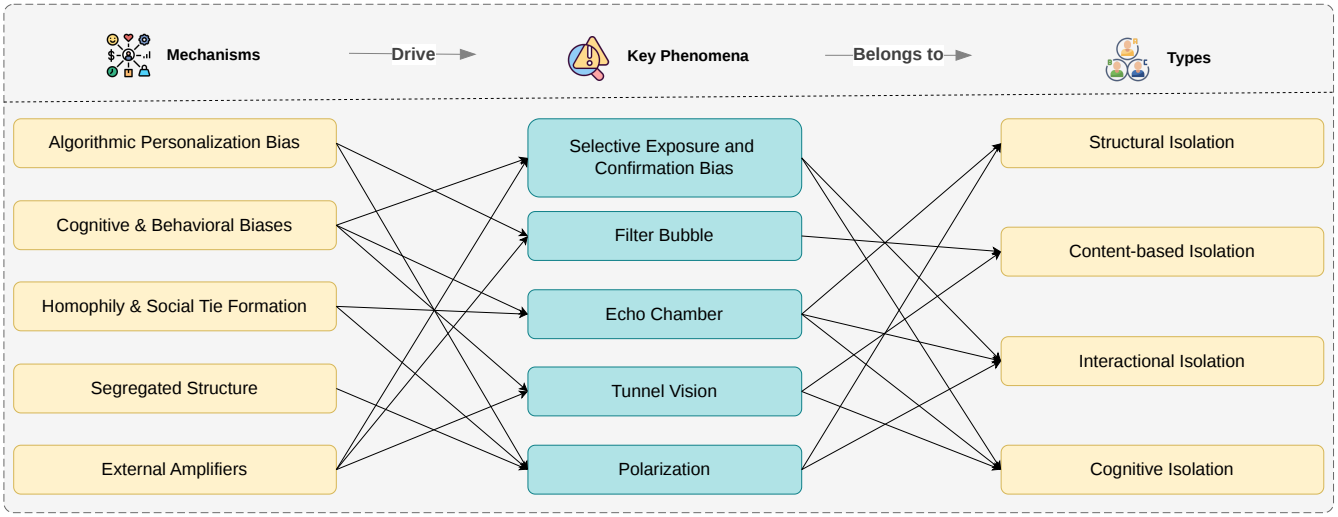

Figure 2: Mechanisms–Phenomena–Types map of ideological isolation. This figure maps five mechanisms, algorithmic personalization bias, cognitive & behavioral biases, homophily & tie formation, segregated network structure, and external amplifiers, to key ideological isolation phenomena, including selective exposure and confirmation bias, filter bubbles, echo chambers, tunnel vision, and polarization, and further organizes these phenomena into four types: structural, content-based, interactional, and cognitive.

natively, cosine similarity or kernel-based functions can be used to capture nonlinear ideological closeness.

We define a content visibility function $P(c_j \mid u_i, t)$, which denotes the probability that a content item $c_j \in \mathcal{C}$ appears in the feed $\mathcal{F}_i(t)$ of user u_i at time t. This probability is shaped by multiple interacting forces, including user-specific preferences, social connectivity, and platform-level mechanisms. For algorithmically curated platforms, $P(c_j \mid u_i, t)$ is often determined by a ranking function that incorporates engagement signals, prior interactions, ideological affinity $\phi(\mathbf{b}_i, \mathbf{v}_j)$, and temporal relevance. On socially driven platforms, e.g., WeChat, the visibility is constrained by the network topology, such that $P(c_j \mid u_i, t) > 0$ primarily if c_j is authored by some neighbor $u_j \in \Gamma(u_i)$, where $\mathcal{C}_j \subseteq \mathcal{C}$ denotes content posted by user u_j [87]. Moreover, platform-imposed constraints, such as attention limits, freshness filtering, or injected advertisements, further restrict $\mathcal{F}_i(t) \subseteq \mathcal{C}$, resulting in a highly selective exposure process. These visibility dynamics provide the foundation for understanding ideological alignment, belief reinforcement, and exposure narrowing in subsequent sections.

3.1.3. Selective Exposure and Content Distribution

Recall that given $\mathcal{F}_i(t) \subseteq \mathcal{C}$, then $\mathcal{F}_i(t) = \{c_1, \ldots, c_k\}$ denotes the set of content shown to user u_i at time t. The exposure distribution of u_i is:

$$P_i(t) = \{p_j(t) \,|\, p_j(t) = \Pr[c_j \in \mathcal{F}_i(t)]\}\,,$$

following prior work that formalizes exposure as a probability distribution over feed content in social platforms [9].

The ideological exposure profile of user u_i is defined as:

$$\boldsymbol{\mu}_i(t) = \sum_{c_j \in \mathcal{F}_i(t)} p_j(t) \cdot \mathbf{v}_j\,,$$

which represents the expected ideological position of the content user u_i is exposed to at time t, analogous to continuous exposure modeling used in recent computational studies of information spread and recommender influence [15, 106].

These formalizations provide a foundation for understanding selective exposure, belief reinforcement, and narrowing of ideological exposure in subsequent sections.

3.1.4. Ideological Isolation Index

This formal expression defines a conceptual isolation index that integrates belief-exposure alignment and content diversity. It provides the theoretical foundation for empirical measurement approaches [10].

$$\mathrm{II}_i(t) = \|\boldsymbol{\mu}_i(t) - \mathbf{b}_i\| + \alpha \cdot \mathrm{Var}_{c_j \in \mathcal{F}_i(t)}(\mathbf{v}_j), \qquad (2)$$

here $\boldsymbol{\mu}_i(t)$ is the ideological exposure profile of user u_i, defined as the expected ideological position of content in their feed; $\mathbf{b}_i$ is the user's ideological belief vector; $\|\boldsymbol{\mu}_i(t) - \mathbf{b}_i\|$ measures the ideological proximity between the user's beliefs and the average ideology of the content they are exposed to, smaller values indicate higher alignment and potential reinforcement of beliefs; $\mathrm{Var}_{c_j \in \mathcal{F}_i(t)}(\mathbf{v}_j)$ denotes the variance of the ideological positions of the content items in the feed, quantifying the degree of ideological diversity, specifically:

$$\mathrm{Var}_{c_j \in \mathcal{F}_i(t)}(\mathbf{v}_j) = \frac{1}{|\mathcal{F}_i(t)|} \sum_{c_j \in \mathcal{F}_i(t)} \|\mathbf{v}_j - \boldsymbol{\mu}_i(t)\|^2\,,$$

where $\alpha > 0$ is a tunable parameter that controls the weight assigned to content diversity.

Thus, the first term $\|\boldsymbol{\mu}_i(t) - \mathbf{b}_i\|$ captures ideological reinforcement, while the second term $\alpha \cdot \mathrm{Var}_{c_j \in \mathcal{F}_i(t)}(\mathbf{v}_j)$ penalizes ideological narrowness. A user is considered more ideologically isolated when they are exposed to content closely aligned with their own beliefs and lacking ideological variance, i.e., when both terms are small.

The two terms in Eq. (2) operate on different scales and must be balanced to avoid degenerate behaviour. The alignment term $\|\boldsymbol{\mu}_i(t) - \mathbf{b}_i\|$ is a distance in $\mathcal{I} \subset \mathbb{R}^d$, while the variance term has units of squared distance. When $\alpha \to 0$, the index reduces to a pure alignment measure, misclassifying users who see diverse but belief-aligned content as isolated, while the latter disregards the role of belief–exposure alignment. When $\alpha \to \infty$, it is dominated by content variance, ignoring belief–exposure alignment. To ensure both terms contribute comparably, a principled heuristic is to set $\alpha = \sigma_{\text{align}}/\sigma_{\text{var}}$, where σ_{align} and σ_{var} are the population-level standard deviations of the respective terms. The index is bounded below by zero and has no fixed upper bound; when a normalized score is needed, dividing by the maximum observed value provides a $[0, 1]$ scale.

3.1.5. Reinforcement Dynamics

The evolution of a user's ideological beliefs can be modeled as a gradual adjustment influenced by the content they are exposed to. Specifically, we define the belief update function for user u_i at time t as:

$$\mathbf{b}_i(t+1) = \mathbf{b}_i(t) + \eta \cdot (\boldsymbol{\mu}_i(t) - \mathbf{b}_i(t)),$$

where $\mathbf{b}_i(t) \in \mathbb{R}^d$ denotes the ideological belief vector of user u_i at time t; $\boldsymbol{\mu}_i(t) \in \mathbb{R}^d$ refers to the ideological centroid of the content feed $\mathcal{F}_i(t)$, representing the average ideological position of content the user sees; $\eta \in [0, 1]$ is the susceptibility parameter, which controls the extent to which a user updates their beliefs in response to their exposure [66].

This can be viewed as a weighted interpolation between the user's current belief and the ideological signal of the content they consume. When $\eta = 0$, the user is completely resistant to change, and their belief remains static. When $\eta = 1$, the user fully adopts the content exposure average.

Over time, if the content a user is exposed to is consistently aligned with their current beliefs (i.e., $\boldsymbol{\mu}_i(t) \approx \mathbf{b}_i(t)$), the update will be small in magnitude but will reinforce the same belief direction, leading to ideological entrenchment. Conversely, if the exposure is diverse or ideologically distant, the user's beliefs may shift, depending on the value of η.

Importantly, when the variance of the ideological content in $\mathcal{F}_i(t)$ is low (i.e., content is homogeneous), the directional pull of $\boldsymbol{\mu}_i(t)$ becomes stable and focused. As a result, repeated updates using this mechanism will steadily reinforce the user's current ideological stance and reduce openness to opposing views. This process captures the essence of reinforcement dynamics observed in echo chambers and filter bubbles.

The summary of notations is given in Table 1.

3.2. Operational Definitions of Key Phenomena

Based on the formal definitions presented earlier, we generally characterize the existence of key ideological phenomena in online social networks, including selective exposure, filter bubbles, echo chambers, tunnel vision, and polarization. These problem definitions provide a unified structural and belief-based perspective, although alternative formulations may be employed in specific settings, depending on platform mechanics, user behavior, or analytical goals.

3.2.1. Selective Exposure and Confirmation Bias

Selective exposure refers to the tendency of individuals to seek out information that aligns with their existing beliefs, often driven by confirmation bias. Large-scale studies indicate that content exposure and recommendations are typically ideologically aligned with users' prior beliefs, reflecting selective exposure and confirmation bias [9].

$$\mathbb{E}_{c_j \in \mathcal{F}_i(t)}[\phi(\mathbf{b}_i, \mathbf{v}_j)] \gg 0\,. \tag{3}$$

Alternatively, using the ideological exposure profile $\boldsymbol{\mu}_i(t)$, the condition becomes:

$$\|\boldsymbol{\mu}_i(t) - \mathbf{b}_i\| \leq \epsilon, \tag{4}$$

where $\boldsymbol{\mu}_i(t)$ is the expected ideological position of content in user u_i's feed, and $\epsilon > 0$ is a small threshold indicating strong alignment with user belief $\mathbf{b}_i$.

3.2.2. Filter Bubble

Filter bubbles arise when algorithmic systems prioritize ideologically aligned content, thus limiting diversity [3]. This occurs when content visibility is determined primarily by ideological similarity:

$$P(c_j \mid u_i, t) \propto \phi(\mathbf{b}_i, \mathbf{v}_j), \tag{5}$$

and the ideological variance of exposed content is low:

$$\mathrm{Var}_{c_j \in \mathcal{F}_i(t)}(\mathbf{v}_j) \leq \delta, \tag{6}$$

where $\delta > 0$ is a small threshold for ideological diversity.

3.2.3. Echo Chamber

An echo chamber refers to an environment in which a user's opinions, ideological stance, or beliefs are reinforced through repeated interactions with neighbors or information sources that share similar tendencies and attitudes [51].

$$\forall u_j \in \Gamma(u_i), \quad \|\mathbf{b}_j - \mathbf{b}_i\| \leq \epsilon, \qquad |\Gamma(u_i)| \geq \gamma\,, \tag{7}$$

$$c_j \in \mathcal{F}_i(t) \Rightarrow c_j \in \mathcal{C}_k \text{ for some } u_k \in \Gamma(u_i)\,. \tag{8}$$

Table 1: Summary of Notations

Symbol	Description
$G = (U, E)$	Online social network represented as a directed graph with users U and edges E.
U	Set of users: $\{u_1, u_2, \ldots, u_N\}$.
$E \subseteq U \times U$	Directed edges denoting social relationships (e.g., follower/followee, friends).
$\Gamma(u_i)$	In-neighbors of user u_i, i.e., users whose content is visible to u_i.
$\mathcal{F}_i(t) \subseteq \mathcal{C}$	Personalized content feed of user u_i at time t.
$\mathcal{I} \subseteq \mathbb{R}^d$	Ideological space shared by users and content.
$\mathbf{b}_i \in \mathbb{R}^d$	Ideological belief vector of user u_i.
$\mathcal{C}$	Set of all content items (posts, comments, videos, etc.).
$\mathcal{C}_j \subseteq \mathcal{C}$	Content authored or reposted by user u_j.
c_j	A single content item.
$\mathcal{C}_a \subseteq \mathcal{C}$	Aspect-specific content subset, i.e., items c_j that pertain to aspect a.
$\mathbf{v}_j \in \mathbb{R}^d$	Ideological embedding of content item c_j.
$\phi(\mathbf{b}_i, \mathbf{v}_j)$	Affinity score between user belief and content, e.g., $-\|\mathbf{b}_i - \mathbf{v}_j\|$.
$P(c_j \mid u_i, t)$	Probability of content c_j appearing in u_i's feed at time t.
$P_i(t)$	Exposure distribution over content in $\mathcal{F}_i(t)$.
$\boldsymbol{\mu}_i(t)$	Ideological exposure profile of user u_i; expectation over content vectors in their feed.
$\mathrm{Var}_{c_j \in \mathcal{F}_i(t)}(\mathbf{v}_j)$	Variance of ideological embeddings of content in the feed.
$\mathrm{II}_i(t)$	Ideological isolation score of user u_i at time t.
$\alpha > 0$	Weighting parameter controlling the importance of content diversity in the isolation metric.
Polarization(G)	Network-level polarization index based on ideological distance between connected users.
$\mathbb{I}[u_j \in \Gamma(u_i)]$	Indicator function: 1 if u_j is a neighbor of u_i, 0 otherwise.
$\eta \in [0, 1]$	Susceptibility parameter controlling belief update strength.
$\mathbf{b}_i(t+1)$	Updated ideological belief of user u_i after exposure at time t.

Here, all neighbours u_j of user u_i (within the visible set $\Gamma(u_i)$) hold similar ideological beliefs $\mathbf{b}_j$ within a small tolerance ϵ, and the neighbourhood size exceeds a minimum threshold γ; consequently, the content displayed in the user's feed $\mathcal{F}_i(t)$ primarily comes from users within the same neighbourhood $\Gamma(u_i)$ [26].

3.2.4. Tunnel Vision

Tunnel vision is a discourse-level concentration of attention in which prolonged selective exposure, amplified by filter bubbles and echo chambers, leads users to focus predominantly on a single aspect of an event while neglecting alternatives [150].

Recall that $\mathcal{F}_i(t) \subseteq \mathcal{C}$ denotes the feed shown to user u_i at time t, with exposure probabilities $p_j(t) = \Pr[c_j \in \mathcal{F}_i(t)]$. For each aspect a, define the subset

$$\mathcal{C}_a = \{\, c_j \in \mathcal{C} : \text{ item } c_j \text{ pertains to aspect } a \,\}.$$

We say that user u_i exhibits tunnel vision at time t when the largest aspect share of exposure exceeds $1 - \gamma$:

$$\max_a \sum_{c_j \in \mathcal{F}_i(t) \cap \mathcal{C}_a} p_j(t) \;\geq\; 1 - \gamma, \tag{9}$$

for some small $\gamma > 0$. Equivalently, there exists an aspect a^* such that items about a^* account for at least a $(1 - \gamma)$ fraction of the user's exposure. This minimal criterion captures aspect concentration succinctly and can be complemented by entropy-based measures when a continuous score is desired.

3.2.5. Polarization

Polarization is a macro-level phenomenon in which ideological distances increase across the network. It is present when the average pairwise ideological distance between connected users is high [55]:

$$\text{Polarization}(G) = \frac{1}{|U|^2} \sum_{i,j} \|\mathbf{b}_i - \mathbf{b}_j\| \cdot \mathbb{I}[u_j \in \Gamma(u_i)], \tag{10}$$

where $\|\mathbf{b}_i - \mathbf{b}_j\|$ represents the belief distance between user u_i and user u_j, typically computed using Euclidean distance. $\mathbb{I}$ is an indicator function and $\Gamma(u_i)$ is the set of visible or interacting neighbors of u_i, specifically:

$$\mathbb{I}[u_j \in \Gamma(u_i)] = \begin{cases} 1, & \text{if } u_j \in \Gamma(u_i) \\ 0, & \text{otherwise} \end{cases}.$$

Polarization exists if

$$\text{Polarization}(G) \geq \epsilon, \tag{11}$$

and the social network exhibits community fragmentation, with high modularity driven by ideological clustering [114].

4. Quantifying Ideological Isolation

In this section, we present a comprehensive set of quantitative approaches to assess ideological isolation in online

social networks. It opens with Figure 3 as an orienting key: network-, content-, and user-behavior metrics diagnose structural, content-based, interactional, and cognitive isolation, respectively, and guide the mitigation families introduced in Section 5. They provide a multifaceted understanding of how isolation emerges, persists, and manifests at both individual and community levels. All the symbols used for this section are listed in Table 2.

Table 2: Key symbols used in Section 4

Symbol	Description
Q	Modularity of a partition.
$\Phi(S)$	Conductance of subset $S \subset U$.
$\lambda_2(\mathcal{L})$	Algebraic connectivity (Fiedler value) of Laplacian $\mathcal{L}$.
$\mathcal{L}$	(Unnormalized) graph Laplacian.
EI	External-Internal index of homophily.
r_{attr}	Categorical assortativity (ideology/camp mixing on edges).
$DL(M)$	Map Equation (Infomap) description length for partition M.
$C_D(u_i)$	Degree centrality of node u_i.
$C_E(u_i)$	Eigenvector centrality of node u_i.
$C_C(u_i)$	Closeness centrality of node u_i.
$d(i,j)$	Shortest-path distance between nodes u_i and u_j.
$C_B(u_i)$	Betweenness centrality of node u_i.
RWC	Random Walk Controversy score on two-way partition.
P_{ab}	Transition prob. for walk from a to b in $\{X, Y\}$.
BC	Boundary connectivity between partitions.
Δ_{ij}	Opinion distance $\lvert p_i - p_j \rvert$ between users u_i and u_j.
H_{topic}	Topic entropy.
Spread	Avg. pairwise cosine distance.
Purity	Share of users whose inferred ideology matches the community majority.
$\mathcal{U}_c$	User set of a community.
$\mathbb{I}[\cdot]$	Indicator function used in Purity.
H	Entropy of a user's exposure distribution.
$\mathbf{p}^{(u_j)}$	Normalized exposure distribution for user u_j.
H_{u_j}	User-level exposure entropy.
EB_i	Exposure bias of user u_i.
$\mathbf{v}_i \in \mathbb{R}^d$	Embedding vector (user/content) in d-dimensional space.
ES_u	Engagement Symmetry for user u.
$\text{IE}_u(t)$	Interaction entropy at time t for user u (over $p_i^u(t)$).
d_t	Euclidean gap between successive recommendations.
$D_{\text{KL}}(P_t \Vert P_{t+1})$	KL divergence between exposure distributions at t and $t{+}1$.
$H(P_t)$	Entropy of exposure distribution at time t.
$\rho(u)$	Like-based polarization score for user u.
$\text{BMC}(t)$	Bimodality coefficient at time t.
$\text{BR}(t)$	Balance ratio at time t.
SV	Structural virality.

4.1. Network-based Metrics

4.1.1. Community segregation

Community segregation occurs when a social network fragments into densely interconnected groups that share few links with the rest of the graph. Because most interactions, information flows, and social reinforcement occur within these clusters, individuals are repeatedly exposed to like-minded peers while seldom encountering discordant viewpoints. Such structural insulation amplifies confirmation bias and is therefore a key mechanism through which *ideological isolation*, the tendency to hold increasingly homogeneous opinions, emerges at the individual level.

The strength of segregation is often quantified by *modularity* Q and related indices (e.g., the M-value), with conductance and algebraic connectivity [77]. Modularity is defined as

$$Q \;=\; \frac{1}{2m}\sum_{i,j}\Big(A_{ij} - \frac{k_i k_j}{2m}\Big)\,\delta(c_i, c_j)\,, \tag{12}$$

where A_{ij} is the edge weight, k_i is the degree of node i, c_i is its community assignment, and m is the total number of edges. A high Q indicates that the network can be partitioned into well-separated communities, a signature of echo chambers.

Conductance offers a complementary view by measuring how well-separated a subset of nodes is from the rest of the graph [167]. For a set of nodes $S \subset U$, conductance is defined as

$$\Phi(S) = \frac{|\partial S|}{\min(\text{vol}(S), \text{vol}(\bar{S}))}\,,$$

where ∂S is the set of edges with one endpoint in S and one in $\bar{S} = U \setminus S$, and $\text{vol}(S) = \sum_{i \in S} k_i$ is the total degree (volume) of nodes in S. A lower $\Phi(S)$ indicates a more segregated community.

Another important metric is the *algebraic connectivity*, where D is the diagonal degree matrix and A the adjacency matrix [34]. Smaller $\lambda_2(L)$ indicates weaker global connectivity and the presence of structural bottlenecks—i.e., the graph admits good cuts and can contain well-separated clusters. For degree-heterogeneous networks, the *normalized* Laplacian,

$$\hat{L} \;=\; I - D^{-1/2} A D^{-1/2},$$

provides a scale-invariant alternative; its Fiedler value $\lambda_2(\hat{L})$ plays an analogous role and is linked to cut quality via Cheeger-type bounds. While λ_2 itself is not ideology-specific, ideological isolation can be assessed spectrally by evaluating the Rayleigh quotient of the ideology partition with $\hat{L}$,

$$\mathcal{R}(x) \;=\; \frac{x^\top \hat{L}\, x}{x^\top x}\,,$$

which is equivalent to the normalized cut/conductance of that split; smaller values then indicate sparser cross-ideology connectivity.

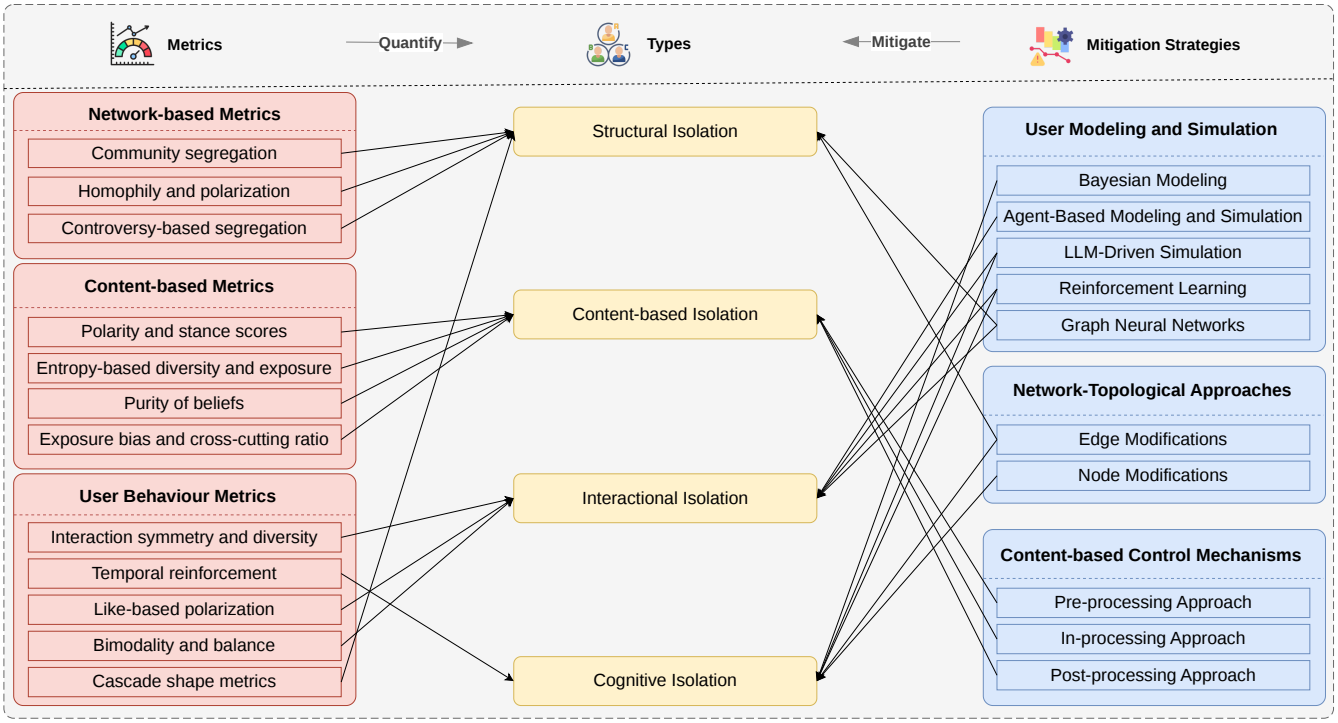

Figure 3: Metrics–Types–Mitigations map of ideological isolation. This figure aligns three metric families (network-based, content-based, user behavior) with four isolation types (structural, content-based, interactional, cognitive) and indicates how mitigation strategies (user modeling & simulation, content-based controls, network-topological approaches) act on the diagnosed type.

These measures provide a multifaceted view of the network's community structure and its potential to foster ideological isolation.

4.1.2. Homophily and polarization

Homophily refers to the social phenomenon where individuals are more likely to form connections with others who share similar attributes, beliefs, or opinions. In online social networks, this leads to preferential attachment between like-minded users, structurally reinforcing local agreement and limiting exposure to diverse viewpoints. As a result, homophily serves as a foundational mechanism behind the formation of *echo chambers*, which in turn intensify *polarization*, the divergence of opinions across distinct user groups.

To quantify the extent of homophilic connections, various metrics are used at different levels. One widely adopted metric is the *External-Internal (EI) index*, defined as:

$$EI = \frac{E - I}{E + I},$$

where E is the number of edges linking a node or group to others with different attributes (external connections), and I is the number of edges linking to those with similar attributes (internal connections). The EI-index ranges from -1 (perfect homophily) to $+1$ (perfect heterophily), with 0 indicating an equal number of internal and external connections. A low or negative EI-index suggests that users are predominantly connected to similar peers, thus reinforcing ideological homogeneity and isolating them from diverse perspectives [17, 70].

As a complementary homophily signal, the *assortativity coefficient* provides a normalized measure for a node attribute x_i (e.g., ideology or stance), defined as [115]:

$$r_{\text{attr}} = \frac{\sum_{ij}(A_{ij} - k_i k_j/2m)x_i x_j}{\sum_{ij}(k_i\delta_{ij} - k_i k_j/2m)x_i x_j}. \tag{13}$$

Here A_{ij} is the (possibly weighted) edge between u_i and u_j, k_i is the degree (or strength) of u_i, $m = \frac{1}{2}\sum_{ij} A_{ij}$ is the total number of edges, δ_{ij} is the Kronecker delta, and x_i encodes the attribute (e.g., ± 1 label or a standardized scalar). The numerator captures attribute covariance across edges beyond the configuration-model baseline $k_i k_j/2m$; the denominator normalizes by the maximum variance under that baseline, yielding a Pearson-type correlation bounded in $[-1, 1]$. Values $r_{\text{attr}} > 0$ indicate assortative (like-with-like) ties, $r_{\text{attr}} < 0$ indicate disassortative (cross-cutting) ties, and $r_{\text{attr}} \approx 0$ is consistent with random mixing. For directed or weighted graphs, we use A_{ij} as weights and may compute separate in- and out-versions by replacing k_i with the in- or out-strengths to match the exposure direction.

However, while homophily captures pairwise similarity, it does not explicitly identify emergent clusters or communities in the network. Therefore, it is often complemented by community detection algorithms, such as Louvain and

Infomap, which help uncover the structural boundaries that amplify homophilic tendencies and contribute to the formation of echo chambers.

4.1.3. Controversy-based segregation

While structural and interaction-based metrics shed light on community cohesion and node influence, they do not directly capture the extent to which a topic divides a network into opposing sides. Controversy-based segregation specifically quantifies the degree of polarization or fragmentation in a discussion, particularly in contexts such as political discourse or social issues. Such polarization reflects high ideological isolation, in which users from different viewpoints rarely interact or engage with one another, reinforcing echo chambers and deepening filter bubbles.

Random Walk Controversy (RWC) [41] score is used to measure polarization in a two-partition network, where the graph is divided into two opposing communities, X and Y, such that $X \cup Y = U$. The RWC is defined as:

$$\text{RWC} = P_{xx}P_{yy} - P_{xy}P_{yx}\,, \tag{14}$$

where P_{ab} denotes the probability that a random walk starting in community $a \in \{X, Y\}$ ends in community $b \in \{X, Y\}$.

High RWC (and variants) implies that random-walk exposure remains confined to like-minded communities, a direct operationalization of ideological isolation through constrained cross-ideological visibility.

Several extensions of the RWC have been proposed to refine the understanding of polarized interactions [113, 143]:

Authoritative-RWC (A-RWC) Focuses on walks between authoritative nodes, emphasizing the separation of influential discourse leaders.

$$\text{A-RWC} = P^{(A)}_{xx}P^{(A)}_{yy} - P^{(A)}_{xy}P^{(A)}_{yx}\,,$$

where $P^{(A)}_{ab}$ denotes the probability that a random walk among authoritative nodes starting from community a ends in community b, with $a, b \in \{X, Y\}$.

Displacement-RWC (D-RWC) Incorporates how information flows are displaced due to centrality-weighted movements across partitions, revealing more subtle dynamics of content segregation.

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

where N is a sampled set of randomly selected vertices from the network (e.g., 60% of each community), l_{rw} is the fixed length of the random walk, and $n(v)_{\text{cc}}$ is the number of community changes encountered in the walk starting from node v.

Higher D-RWC values indicate fewer inter-community transitions and stronger polarization [143].

Boundary Connectivity (BC) [41, 46] quantifies how boundary nodes split their ties between in-group and out-group neighbors. Let B be the set of boundary nodes and I the set of internal (same-side) nodes. For a boundary node u, let $d_i(u)$ be the number of edges from u to I, and $d_b(u)$ the number of edges from u to B. Then

$$\text{BC} = \frac{1}{|B|}\sum_{u\in B}\left(\frac{d_i(u)}{d_i(u) + d_b(u)} - 0.5\right)\,,$$

which lies in $[-0.5, 0.5]$. Values $\text{BC} < 0$ indicate little or no polarization; values $\text{BC} > 0$ mean boundary nodes, on average, preferentially connect inward (to internal same-side nodes) rather than across groups, signaling that controversy is likely present.

These four metrics differ meaningfully in their applicable scenarios. The original RWC provides a global partition-level score, suitable when a clean two-way community split is available . A-RWC refines this by restricting walks to authoritative nodes, making it appropriate when elite-level separation is the concern rather than general-population flow. D-RWC instead counts community-boundary crossings along fixed-length walks, making it preferable when community sizes are imbalanced or when global transition probabilities may mask local mixing. The critical distinction between A-RWC and D-RWC is what they measure: A-RWC asks who the walk reaches, while D-RWC asks how often it crosses boundaries. They address different research questions and are not interchangeable. BC complements all three walk-based measures by directly examining whether boundary nodes preferentially connect inward or outward; unlike the others, it requires no random walks and is therefore robust on small or sparse networks.

Betweenness centrality [104] captures a node's role as a bridge between different parts of the network:

$$C_B(u_i) = \sum_{s\neq i\neq t}\frac{\sigma_{st}(i)}{\sigma_{st}}\,,$$

where σ_{st} is the number of geodesic (shortest) paths from u_s to node u_t, and $\sigma_{st}(i)$ is the number of those paths passing through u_i. Nodes with high betweenness centrality can act as gatekeepers or facilitators of cross-group communication. When such bridging nodes are scarce, structurally peripheral, or bypassed in information flows, structural ideological isolation intensifies.

In these views, isolation arises when edges concentrate within like-minded subsets, boundary nodes preferentially attach inward, and random-walk mass becomes trapped within communities rather than diffusing across them. These signals quantify how the graph's wiring constrains cross-ideological visibility, even when the network is globally connected.

4.2. Content-based Metrics

4.2.1. Polarity and stance scores

In the absence of explicit group labels (e.g., political affiliation or community membership), user ideology can be

inferred from the content they post, share, or endorse. *Polarity scores* and *stance detection* offer effective means to estimate opinion leaning based on linguistic and emotional cues in user-generated content. These methods rely on sentiment analysis, political lexicons, or supervised classification models to assign a score or class label that reflects the user's opinion orientation [32].

Polarity scores ranging from "far-left" to "far-right" can be inferred from language features and retweet behavior in social media [67]. Let $p_i \in [-1, 1]$ denote the polarity score of user u_i, where -1 indicates one extreme (e.g., left-leaning or anti-stance) and $+1$ the opposite extreme (e.g., right-leaning or pro-stance). The polarity score can be computed using various strategies:

1. Sentiment analysis of shared articles or posts.
2. Aggregated stance of liked or re-posted content.
3. Classifiers trained on labeled political datasets.

We treat the scalar polarity as a one-dimensional belief, $p_i \equiv \mathbf{b}_i$. Then the opinion distance between users u_i and u_j follows Eq. (1) in Section 3:

$$\Delta_{ij} = \|\mathbf{b}_i - \mathbf{b}_j\| = |p_i - p_j| \,,$$

where a larger Δ_{ij} indicates stronger divergence in stance. Aggregating these distances across a community or between communities can offer a quantitative view of ideological segregation.

In this context, content-based polarity and stance scores serve as proxies for user ideology, enabling researchers to detect echo chambers and filter bubbles through the lens of expressed opinions, even in the absence of network labels or explicit community structures.

4.2.2. Entropy-based diversity and exposure

This subsection quantifies informational breadth from two complementary angles: (i) *topic/semantic diversity*, which characterizes the range of content that exists to be seen [53], and (ii) *exposure entropy*, which captures the distribution of what users actually see in their feeds [59].

Topic Entropy (LDA-based): Let $\mathcal{T} = \{\tau_1, \ldots, \tau_K\}$ be the topic set, and let θ denote a topic-mixture distribution over $\mathcal{T}$ (e.g., from LDA). We define

$$H_{\text{topic}} \;=\; -\sum_{k=1}^{K} \theta(\tau_k)\, \log \theta(\tau_k)\,. \tag{15}$$

Lower H_{topic} indicates a more concentrated (less diverse) thematic profile. This measure reflects the semantic opportunity set available in the corpus or in a user's candidate pool.

Semantic Entropy (BERT-based): For a set of semantic vectors $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$, diversity can be captured by computing pairwise cosine distances d_{cos} and measuring their spread:

$$H_{\text{semantic}} = \text{Entropy}\left(d_{cos}(\mathbf{e}_i, \mathbf{e}_j)\right) \,,$$

where the entropy is computed over the distribution of pairwise distances.

Semantic Spread: Defined as the average pairwise distance in embedding space:

$$\text{Spread} = \frac{2}{n(n-1)} \sum_{i<j} d_{cos}(\mathbf{e}_i, \mathbf{e}_j)\,,$$

where a lower spread suggests semantically redundant content consumption.

Consider content grouped into K topics $\mathcal{T} = \{\tau_1, \ldots, \tau_K\}$ with exposure probabilities $p(\tau_k)$. The entropy of exposure is

$$H \;=\; -\sum_{k=1}^{K} p(\tau_k)\, \log p(\tau_k)\,. \tag{16}$$

At the user level u_i, let $p_i(\tau_k)$ denote the probability that u_i is exposed to (or consumes) topic τ_k; then

$$H_{u_i} \;=\; -\sum_{k=1}^{K} p_i(\tau_k)\, \log p_i(\tau_k)\,.$$

Lower H_{u_i} signifies concentrated exposure, whereas higher values indicate a more even distribution across topics.

While H_{topic} captures the semantic supply of potentially visible content, H_{u_j} captures realized exposure. Ideological isolation is evidenced when both entropies are low, or when H_{u_j} is markedly smaller than H_{topic}, implying that ranking mechanisms and/or user choices narrow attention to like-minded content despite broader semantic availability.

4.2.3. Purity of beliefs

This metric evaluates the ideological homogeneity within a user community by assessing the proportion of users whose inferred ideology aligns with the community's majority ideology [110].

$$\text{Purity} = \frac{1}{|\mathcal{U}_c|} \sum_{u \in \mathcal{U}_c} \mathbb{I}[\hat{y}_u = y_{\text{majority}}]\,, \tag{17}$$

where $\mathcal{U}_c$ is the set of users in a community (or echo chamber); $\hat{y}_u$ is the inferred ideology label of user u; y_{majority} is the dominant ideology label in $\mathcal{U}_c$; $\mathbb{I}[\cdot]$ is the indicator function that returns 1 if the condition is true, and 0 otherwise.

A higher purity score indicates greater ideological alignment among members, signaling semantic and ideological homogeneity within the group.

4.2.4. Exposure bias and cross-cutting ratio

A defining feature of echo chambers and filter bubbles is *selective exposure*, in which users tend to encounter content that aligns with their preexisting beliefs. In contrast, opposing viewpoints are filtered out by algorithms or social behavior. To quantify this bias in content exposure,

researchers measure the proportion of ideologically opposing (cross-cutting) content that a user encounters within their information stream, such as a social media feed or content recommendation list.

One of the most direct measures of this phenomenon is the *exposure bias* [9], defined as:

$$\mathrm{EB}_i = 1 - \frac{C_i}{T_i}\,.$$

For user u_i, let C_i be the number of ideologically cross-cutting posts encountered by user u_i, and let T_i be the total number of posts seen by user u_i.

An exposure bias value near 1 indicates highly selective exposure, suggesting the user is insulated from dissenting viewpoints, while values closer to 0 reflect balanced ideological exposure.

Conversely, the *cross-cutting exposure ratio* can be defined simply as:

$$\text{Cross-Cutting Ratio} = \frac{C_i}{T_i}\,,$$

which serves as an inverse indicator of ideological isolation.

Low cross-cutting ratios or high exposure bias values are commonly interpreted as evidence of *echo chamber* or *tunnel vision* effects, where users are primarily exposed to information that reinforces their ideological views. This metric has been employed in influential studies, such as [9], which analyzed Facebook user behavior and found that both algorithmic filtering and user choice contributed to reduced cross-cutting exposure.

4.3. User Behavior Metrics

4.3.1. Interaction symmetry and diversity

User behavior is a key driver of ideological isolation, particularly through interaction patterns that exhibit selective exposure [74]. In digital environments, users often gravitate toward content that aligns with their beliefs, while avoiding opposing viewpoints, thereby reinforcing echo chambers and deepening confirmation bias. *Interaction symmetry* aims to capture these behavioral tendencies by analyzing how users engage with the ideological diversity of the content presented to them [9, 22].

The *Engagement Symmetry* score for user u_i, denoted as ES_i, is defined as:

$$\mathrm{ES}_i = \frac{Int_i^{same} - Int_i^{opp}}{Int_i^{tot}}\,, \quad Int_i^{tot} = Int_i^{same} + Int_i^{opp}$$

Here Int_i^{same} and Int_i^{opp} denote the (weighted) counts of u_i's interactions with ideologically aligned and opposing content, respectively. A higher ES_i indicates stronger engagement with ideologically aligned content (echo chamber tendency), a lower ES_i indicates preference for opposing views, while values near zero reflect balanced engagement across ideological lines.

Symmetry reflects which side a user prefers, and entropy reflects how broadly they engage across stances or categories.

To formalize this, entropy can be calculated over a user's interaction distribution across content categories or ideological stances [120]. Given user u_i's probability distribution of interactions across a set $\mathcal{C}$ of categories or topics at time t, denoted by $\{\, p_c^{(u_i)}(t) \,\}_{c\in\mathcal{C}}$, the *Interaction Entropy (IE)* is computed as:

$$\mathrm{IE}_{u_i}(t) \;=\; -\sum_{c\in\mathcal{C}} p_c^{(u_i)}(t)\, \log p_c^{(u_i)}(t)\,.$$

A high $\mathrm{IE}_{u_i}(t)$ indicates a user engages with a wide range of content, signaling openness and exposure diversity. In contrast, a low $\mathrm{IE}_{u_i}(t)$ implies behavioral concentration, suggesting that the user may be entrenched in a narrow ideological niche.

4.3.2. Temporal reinforcement

Temporal reinforcement refers to the phenomenon where recommender systems gradually deliver increasingly similar content to users over time, thereby reinforcing users' prior preferences and narrowing their information exposure. This phenomenon can be quantified by observing changes in content diversity across recommendation sequences.

Ge et al. demonstrate this effect by analyzing the Euclidean distance between the embeddings of consecutively recommended items [42]. In a high-dimensional semantic space (e.g., BERT or collaborative filtering embeddings), each item is represented as a vector. The Euclidean distance between successive items $\mathbf{x}_t$ and $\mathbf{x}_{t+1}$ is defined as:

$$d_t = \|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2\,.$$

This temporal narrowing contributes directly to the emergence of echo chambers and filter bubbles, and can culminate in tunnel vision—prolonged selective exposure that concentrates attention on a single aspect of an event, further reducing viewpoint diversity and amplifying ideological isolation. Through algorithmic filtering and behavioral feedback loops, personalization increasingly homogenizes future recommendations, increasing temporal stability of exposure and reinforcing confirmation bias and polarization.

To quantify this stability, researchers track changes in exposure distributions (e.g., topics, sentiments, ideologies) over time. A notable decline in diversity suggests an increase in informational confinement and the strengthening of echo chambers. One common approach to measure this temporal change is the use of *Kullback-Leibler (KL) divergence*, which quantifies how much one probability distribution diverges from another [13].

Given a user u_i's exposure distributions $P_i(t)$ and $P_i(t+1)$ at consecutive timepoints t and $t+1$, the KL-

divergence is defined as:

$$D_{\mathrm{KL}}(P_i(t)\|P_i(t+1)) = \sum_{k=1}^{K} P_i(t,\tau_k)\log\frac{P_i(t,\tau_k)}{P_i(t+1,\tau_k)}\,,$$

where τ_k denotes topic k, and $P_i(t)$ is the probability that u_i is exposed to τ_k at time t. A high D_{KL} indicates a significant shift in the user's content exposure between timepoints, which may signal exploration or a change in user behavior. In contrast, a persistently low D_{KL} indicates temporal stability, often reflects algorithmic reinforcement, and increases the homogeneity of content consumption.

Additionally, to track the overall diversity of exposure over time, we use the entropy definition in Eq. (16). We track temporal change via

$$\Delta H_{u_i}(t) = H_{u_i}(t{+}1) - H_{u_i}(t), \qquad \widehat{H}_{u_i}(t) = \frac{H_{u_i}(t)}{\log K}\,.$$

A persistently decreasing $\widehat{H}_{u_i}(t)$ with low but stable D_{KL} indicates temporal reinforcement and narrowing exposure, consistent with entropy-based analysis of echo chambers and selective exposure [26].

4.3.3. Like-based polarization

Like-based polarization is a user-level metric that quantifies ideological alignment based on observable engagement behaviors, specifically the number of "likes" on social media platforms [16]. In many online environments, users are exposed to content representing opposing narratives. For example, scientific information versus conspiracy theories, or left-leaning versus right-leaning political discourse.

Brugnoli et al. [16] propose a simple yet effective polarization score to capture this behavior:

$$\rho(u) \;= \frac{x-y}{x+y}\,, \tag{18}$$

where x denotes the number of likes a user u gives to content aligned with one narrative (e.g., conspiracy); y refers to the number of likes given to content from the opposing narrative (e.g., science). The resulting score $\rho(u) \in [-1,1]$ reflects the user's ideological leaning: a higher $\rho(u)$ indicates stronger alignment with one ideological side (e.g., conspiracy), a lower $\rho(u)$ indicates alignment with the opposing side (e.g., science), and values near zero represent balanced engagement across both narratives.

A bimodal distribution of $\rho(u)$ values across the population, with peaks near -1 and 1, indicates the presence of two polarized user groups, with few users engaging meaningfully with both sides. This behavioral pattern reflects the fragmentation of public discourse into ideologically homogeneous communities, a hallmark of echo chambers.

Moreover, this metric directly connects to filter bubbles: as platforms personalize content based on prior likes, users become increasingly surrounded by narratives that reinforce their initial preferences. The result is a feedback loop that magnifies polarization at both the individual and community levels.

4.3.4. Bimodality and balance

Bimodality and balance are statistical metrics used to capture the degree and structure of opinion divergence in a population [123]. In online discourse, bimodality refers to a distribution of user opinions with two distinct peaks, indicating the presence of two dominant ideological camps. This suggests that users are not uniformly distributed across the opinion spectrum but instead cluster around opposing poles. Such separation signifies strong *behavioral polarization*, where engagement patterns and content preferences reflect entrenched ideological positions.

Bimodality is often accompanied by asymmetric group sizes, with one ideological camp significantly outnumbering the other. This asymmetry is captured through the notion of *balance*, which quantifies the relative size of opposing opinion groups. When combined, bimodality and balance offer a nuanced view of polarization, revealing not only the separation between camps but also the structural marginalization that may lead to ideological isolation or filter bubble effects.

To quantify bimodality in opinion distributions, the *Bimodality Coefficient (BMC)* [12, 123] is used:

$$\mathrm{BMC}(t) = \frac{\gamma_1(t)^2+1}{\gamma_2(t) + \frac{3(n(t)-1)^2}{(n(t)-2)(n(t)-3)}}\,,$$

where $\gamma_1(t)$ is the skewness, $\gamma_2(t)$ is the kurtosis, and $n(t)$ is the number of users at time t. A value of $\mathrm{BMC}(t) > 0.55$ typically indicates a bimodal or skewed distribution [37]. In online settings, this often corresponds to two ideologically homogeneous camps with few users expressing moderate views. As such, a high bimodality coefficient is interpreted as a strong signal of opinion polarization and the emergence of echo chambers.

Complementing the separation captured by the bimodality coefficient $\mathrm{BMC}(t)$, the *Balance Ratio (BR)* assesses the symmetry of group sizes by taking the ratio of the smaller to the larger of the two dominant clusters [55]:

$$\mathrm{BR}(t) = \frac{\min\{\,|\mathcal{G}_1(t)|,\ |\mathcal{G}_2(t)|\,\}}{\max\{\,|\mathcal{G}_1(t)|,\ |\mathcal{G}_2(t)|\,\}}\,,$$

where $\mathcal{G}_1(t)$ and $\mathcal{G}_2(t)$ are the two dominant ideological clusters detected at time t. The balance ratio lies in $(0,1]$: values near 1 indicate equally sized camps (symmetric polarization), while lower values indicate a dominant majority versus a much smaller minority. A low balance ratio heightens ideological isolation for the minority group, as fewer like-minded peers and items constrain within-camp exposure and support, thereby increasing minority social/structural isolation.

4.3.5. Cascade shape metrics

Information cascades, chains of resharing or reposting, leave structural footprints of diffusion that reveal whether engagement is broad or confined to like-minded communities [26]. Cascade shape metrics quantify these dynamics through depth, breadth, structural virality, and density, offering a behavioral-structural lens on ideological isolation and selective propagation. Deep, cross-community cascades suggest broad exposure, whereas shallow, wide cascades confined to tightly connected clusters indicate echo-chamber reinforcement. We next introduce four key metrics.

Cascade Depth (D) [82] refers to the longest path from the original sharer to any node in the cascade. It captures how far content travels hierarchically:

$$D = \max_{u \in U} \text{depth}(u)\,,$$

where U is the set of nodes in the cascade tree.

Cascade Breadth (B) [82] indicates the maximum number of nodes at any single level of the cascade:

$$B = \max_{l} |U_l|\,,$$

where U_l is the set of nodes at level l of the cascade.

Structural Virality (SV) [45] refers to the average pairwise distance between all nodes in the cascade. This metric balances both depth and breadth:

$$SV = \frac{1}{n(n-1)} \sum_{i \neq j} d(i,j)\,,$$

where n is the number of nodes in the cascade and $d(i,j)$ is the shortest path between nodes u_i and u_j.

Cascade Density [115] measures the ratio of actual to possible edges in the cascade graph. A high density indicates intense local resharing:

$$\text{Density} = \frac{2|E|}{|U|(|U|-1)}\,,$$

where E and U are the edge and node sets, respectively.

Low structural virality, coupled with high density and shallow depth, is typical of cascades confined within echo chambers, where content is predominantly reshared among a homogeneous group. In contrast, cascades that reach high depth and exhibit low density are more likely to cross ideological or community boundaries, reflecting greater diversity in content exposure.

4.4. Cross-Platform Empirical Synthesis

The metric families introduced in Sections 4.1 to 4.3 have been applied to a variety of platforms with substantially different outcomes, motivating a brief empirical synthesis. We consolidate findings from published comparative studies into two reference tables: a platform-level synthesis (Table 3) and a fit map between metric families and isolation types (Table 4).

Evidence from comparative studies shows that the effectiveness of ideological-isolation metrics varies substantially across platforms. Cinelli et al. [26] analyzed over one billion content items on Facebook, Twitter, Reddit, and Gab and found clear echo-chamber signatures on Facebook and Twitter but not on Reddit and Gab. Impicciché and Viviani [64] extended this with a head-to-head comparison of network-based, content-based, and hybrid echo-chamber detection metrics on the same Twitter and Reddit datasets. They report that network-based metrics reliably flag Twitter as controversial, while the same metrics classify Reddit as non-controversial. This discrepancy is explained by differences in interaction granularity rather than by an absence of ideological clustering. The practical implication is that no single metric is universally valid across platforms. Robust diagnosis requires reporting multiple complementary indicators, such as network homophily, exposure entropy, and interaction symmetry.

Studies conducted outside English-language platforms suggest that ideological isolation manifests differently across linguistic, cultural, and platform contexts. Most foundational measurement work has been conducted on English-language platforms, e.g., Facebook, X, Reddit, Gab, but a growing body of computational research has examined Chinese-language and Indian-language ecosystems with different platform mechanics and cleavage structures. On Weibo, Song et al. [134] compared echo chambers across Twitter and Weibo and showed that personality traits interact differently with platform mechanics, yielding distinct clustering structures across the two platforms. On short-video platforms, Gao et al. [38] quantified echo-chamber effects on Douyin, TikTok, and Bilibili, finding homogeneous clustering dominant on Douyin; Li et al. [91] separately mapped TikTok's political co-follow network and identified structurally isolated, like-minded communities. WeChat, which relies more on social-graph-driven exposure than on algorithmic ranking [87, 169], tends to exhibit "echo without chambers": interactional concentration without dense structural community boundaries. Recent work models information cocoons as an emergent outcome of human-AI adaptive dynamics on these platforms [57, 124]. On Indian regional platforms, ShareChat has been characterized as a multilingual social network spanning 14 languages and exhibiting substantial cross-lingual content diffusion [2]. These findings suggest that ideological-isolation phenomena are observed across diverse linguistic and regulatory contexts. However, the choice of metrics and the operationalization of ideology must be adapted to local cleavages, cultural conditions, and platform mechanics.

Different metric families vary in their ability to diagnose particular forms of ideological isolation. Table 4 consolidates which metric families most effectively diagnose each isolation type. The mapping is qualitative.

Although no single metric family captures all forms of ideological isolation, combining network-based metrics with content- or behavior-based measures generally pro-

Table 3: Empirical synthesis of ideological-isolation findings across major platforms, drawn from published comparative studies. Entries indicate strength of reported evidence: ✓ strong, ∼ mixed, × weak or absent. **EC** = echo chamber, **FB** = filter bubble, **Pol.** = polarization, **CC** = cross-cutting exposure.The CC column reports the observed level of cross-cutting exposure (Low/Med).

Platform	EC	FB	Pol.	CC	Dominant metric family	Key references
Facebook	✓	✓	✓	Low	Network + content	[9, 26]
Twitter / X	✓	N/A	✓	Low	Network + behavior	[26, 41, 64]
Reddit	×	N/A	∼	Med	Content + behavior	[26, 64]
Gab	×	N/A	∼	Low	Network	[26]
TikTok / Douyin	✓	✓	✓	Low/Med	Behavior + content	[38, 91, 144]
Weibo	✓	✓	∼	Low	Network + content	[109, 134, 146]
WeChat	∼	×	∼	Med	Network (social graph)	[87, 169]
ShareChat	∼	×	∼	N/A	Content (multilingual)	[2]

vides a more comprehensive diagnosis. Likewise, mitigation strategies exhibit different trade-offs in effectiveness, scalability, interpretability, and deployment complexity, suggesting that their suitability depends on the target platform and intervention objective.

Table 4: Suitability of metric families for diagnosing each isolation type. "H" = strong primary signal; "M" = useful complementary signal; "L" = weak signal. **Struct.** = Structural; **Inter.** = Interactional; **Cogn.** = Cognitive

Metric family	Struct.	Content	Inter.	Cogn.
Community segregation (Q, Φ, λ_2)	H	L	M	L
Homophily / assortativity (EI, r_{attr})	H	L	H	L
RWC and variants	H	L	M	L
Topic / semantic entropy	L	H	L	M
Exposure bias / cross-cutting ratio	L	H	H	M
Engagement symmetry / interaction entropy	L	M	H	M
Temporal reinforcement (KL, ΔH)	L	M	H	H
Bimodality / balance	H	M	M	L
Cascade shape (D, SV, density)	H	L	M	L

5. Computational Mitigation Strategies for Ideological Isolation

This section focuses on computational approaches designed to mitigate ideological isolation in online social networks. As illustrated in Figure 3, the mitigation strategies discussed here correspond directly to the metric families and isolation types introduced in Section 4, thereby providing an operational link between diagnosis and intervention. We organize the discussion into three complementary perspectives. User modeling and simulation approaches replicate belief dynamics and user interactions to evaluate the effectiveness of interventions. Network-topological approaches aim to reshape connectivity by adding, removing, or reweighting edges to enhance cross-cutting exposure. Content-based control mechanisms intervene within recommendation and ranking pipelines through pre-, in-, and post-processing to promote informational diversity while maintaining relevance. These strategies form a computational toolkit for reducing the self-reinforcing mechanisms that sustain ideological isolation.

5.1. User Modeling and Simulation

To develop strategies for mitigating ideological isolation, it is essential to understand and model users' behavior, i.e., how they respond to information. Such modeling and simulation not only capture the dynamics of user interaction and belief evolution but also provide feedback for evaluating the effectiveness of the proposed strategies [28]. In this subsection, we survey several popular approaches for modeling and simulating users' behaviors.

5.1.1. Bayesian Modeling

To model how users perceive and disseminate information in social networks, Bayesian frameworks have been widely adopted to capture the belief-updating process driven by interactions with neighbors [68, 105].

Generally, each user u maintains a prior belief distribution over potential message intents or themes, denoted by $\{m_i\}$. This prior, $P(m_i)$, reflects the user's historical preferences or typical communicative behavior. At each timestep, the user observes recent messages from their neighbors, denoted by $\mathbf{M}_{\Gamma(u)}$, where $\Gamma(u)$ is the set of user u's neighbors. These neighbor messages provide evidence that may influence users' beliefs and guide their response generation. Using Bayes' theorem, the user updates their belief distribution over possible messages as:

$$P(m_i \mid \mathbf{M}_{\Gamma(u)}) \propto P(m_i) \cdot P(\mathbf{M}_{\Gamma(u)} \mid m_i)\,,$$

where $P(m_i)$ is the prior probability of producing message m_i, and $P(\mathbf{M}_{\Gamma(u)} \mid m_i)$ denotes the likelihood of observing the neighbor messages $\mathbf{M}_{\Gamma(u)}$ given intent m_i. This likelihood can be estimated based on semantic similarity, topical relevance, or discourse coherence between m_i and the neighbor messages.

This Bayesian mechanism enables users to adapt their views based on local interactions and to propagate information aligned with their updated beliefs, thereby simulating a dynamic social information flow.

5.1.2. Agent-Based Modeling and Simulation

Agent-Based Models (ABMs) are widely used to study the emergence of ideological isolation by explicitly representing individual traits and behaviors. A critical aspect of these models is understanding how agents influence, and are influenced by, their neighbors [11, 166]. Prior research highlights personal preference as a key factor, alongside the influence of surrounding agents [89, 155]. To capture both intrinsic predispositions and social interaction effects, two mechanisms, *Prior Commitment Level* (PCL) and *Comprehensive Social Influence* (CSI), are incorporated into ABM frameworks [89, 166].

Mathematically, each agent u_i is part of a social network $G = (U, E)$ with a set of neighbors $\Gamma(u_i)$. The PCL pcl_{jx} measures agent u_j's predisposition toward item i_x based on past ratings, normalized as:

$$pcl_{jx} = \begin{cases} \frac{r_{jx} - \min(R_j)}{\max(R_j) - \min(R_j)}, & \max(R_j) \neq \min(R_j) \\ 0.5, & \text{otherwise} \end{cases} \tag{19}$$

where R_j is the set of ratings given by u_j. Higher pcl_{jx} implies stronger favorability, while lower values indicate the opposite. The CSI reflects the cumulative persuasive strength from neighbors $\Gamma(u_j)$ who hold a given opinion:

$$csi_{jx} = 1 - \prod\nolimits_{v_m \in \Gamma(u_j)} (1 - ipp_{mj}) \,, \tag{20}$$

where ipp_{mj} is the interpersonal persuasion power from v_m to u_j. The probability of u_j adopting a new opinion is a weighted combination of personal predisposition (PCL) and social influence (CSI), modulated by a trade-off parameter $\lambda_j \in [0, 1]$:

$$p_{jx} = \lambda_j \cdot pcl_{jx} + (1 - \lambda_j) \cdot \frac{csi_{jx}}{\sum_{x'} csi_{jx'}} \,. \tag{21}$$

This formulation allows ABMs to capture how individual tendencies and peer influence jointly drive opinion dynamics.

Another approach to designing agents in ABM is to define them through a set of explicit rules that govern how they perceive, process, and act upon information [90, 147]. At each time step t, each agent receives *influence messages* $msg_p(u_j \to u_i, t)$ from neighbors $u_j \in \Gamma(u_i)$, where each message carries topical relevance and an opinion indicator.

Formally, a message msg_p can be represented as:

$$S_{msg_p} = \{m_p(\tau_k) \mid \tau_k \in T\}, \quad m_p(\tau_k) \in [0, 1] \,, \tag{22}$$

where $m_p(\tau_k)$ is the degree of association with topic τ_k. The agent retains a posting history $PM^{(u_i)}$ and a received message set $I^{(u_i)}$, which shape its current context $C_L(u_i, t)$. Decision-making is then modeled as:

$$\text{NextMessage}(u_i, t) = f\big(I^{(u_i)}, PM^{(u_i)}, \Gamma(u_i)\big) \,, \tag{23}$$

where $f(\cdot)$ encodes behavioral rules (e.g., Perceive recent messages from neighbors, Incorporate own past expressed opinions, Select and deliver the next influence message).

By leveraging ABM, the network's evolution is driven by individual agents' behaviors [88]. Over time, this evolution can reveal ideological patterns, which can be evaluated using established metrics of ideology.

5.1.3. LLM-Driven Simulation

An emerging direction in ABMs is the integration of large language models (LLMs) to simulate realistic, context-aware user behaviors [60, 165]. In this paradigm, each agent u_i operates within a social network G and maintains repositories of received messages R_i^{in} and posted messages R_i^{out}. At each timestep t, an agent decides whether to broadcast, adopt, or evolve a message based on probabilistic thresholds. When triggered, the agent uses an LLM as a function to generate a response m_x conditioned on both the message content c_x and the agent's profile u_i:

$$m_x = \text{LLM}(c_x, u_i) \,.$$

This mechanism enables the simulation of nuanced discourse generation, where responses are shaped by both the social context (e.g., influence from neighbors $\Gamma(u_i)$) and the agent's prior outputs. By iteratively applying

$$R_i^{out}(t{+}1) = R_i^{out}(t) \cup \{m_x\}, \quad R_j^{in}(t{+}1) = R_j^{in}(t) \cup \{m_x\}$$

for targeted neighbors u_j, the model captures message propagation, evolution, and conversational focus shifts over time. Compared with traditional probability-based ABMs, this approach enables agents to produce language-rich, semantically coherent interactions, thereby allowing the study of how influence diffusion shapes collective discourse and topic dominance in online environments. On the other hand, because LLM behavior is more uncertain and its use requires greater time and computational resources, applying LLM-based Monte Carlo simulation to large-scale networks is challenging. However, if the model can tolerate variability in individual behaviors, LLM-driven simulation remains a suitable option.

5.1.4. Reinforcement Learning

A complementary way to model users is to treat each user agent as a reinforcement-learning (RL) decision maker embedded in a social network, as RL naturally captures the sequential, feedback-driven nature of online interactions, allowing agents to adapt their strategies over time to maximize long-term engagement or influence [131, 133, 156]. Mathematically, at each time step t, agent u_i receives an observation:

$$o_t^{(i)} = g\big(M_t^{(i)}, x_t^{(i)}, c_t\big) \,,$$

where $M_t^{(i)}$ is the set of messages from neighbors $\Gamma(u_i)$ at timestep t, $x_t^{(i)}$ encodes the agent's internal state (e.g., preferences/history), and c_t denotes contextual features (topic, time, forum). The agent selects an action from a finite set:

$$a_t^{(i)} \in \mathcal{A} = \{\text{ignore, reshare, reply, click, compose}, ...\} \,,$$

according to a policy $\pi_\theta(a \mid o)$ (e.g., Deep Q-Learning (DQN) or actor–critic).

The immediate utility of an action is captured by a reward function that balances engagement, preference alignment, and effort/risk:

$$r_t^{(i)} \;=\; \alpha\, E\big(m_t^{(i)}\big) \;+\; \beta\, S\big(m_t^{(i)}, P^{(i)}\big) \;-\; \eta\, C\big(a_t^{(i)}\big)\,,$$

where $m_t^{(i)}$ is the produced/forwarded message, $E(\cdot)$ measures observable engagement (e.g., click, reply), $S(\cdot, P^{(i)})$ measures semantic alignment with the agent's preference profile $P^{(i)}$, and $C(\cdot)$ penalizes effort, delay, or moderation risk. To reflect social spillovers, the reward can include a neighborhood term:

$$\bar{r}_t^{(i)} \;=\; r_t^{(i)} \;+\; \lambda \sum_{j \in \Gamma(u_i)} w_{ij}\, \Delta\Psi_t^{(j)}\,,$$

where $\Delta\Psi_t^{(j)}$ is a change in neighbor u_j's engagement or stance and w_{ij} is an influence weight. The objective is to maximize the discounted return

$$J(\theta) \;=\; \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{T} \gamma^t\, \bar{r}_t^{(i)}\right]\,,$$

where $\gamma \in [0,1]$ denotes the discount factor, and $J(\theta)$ is optimized either via value-based learning with the Bellman target:

$$Q(o_t^{(i)}, a_t^{(i)}) \;\leftarrow\; r_t^{(i)} + \gamma \max_{a'} Q(o_{t+1}^{(i)}, a'),$$

or via policy gradient:

$$\nabla_\theta J(\theta) \;=\; \mathbb{E}\Big[\nabla_\theta \log \pi_\theta(a_t^{(i)} \mid o_t^{(i)})\,(G_t^{(i)} - b(o_t^{(i)}))\Big]\,,$$

with return $G_t^{(i)}$ and baseline $b(\cdot)$ for variance reduction. Message diffusion is modeled by updating neighbors' inboxes stochastically, e.g.,

$$R_j^{\text{in}}(t{+}1) \;=\; R_j^{\text{in}}(t) \cup \{m_t^{(i)}\} \quad \text{with probability } p_{ij}\,.$$

Because observations are partial, $o_t^{(i)}$ is often encoded with a history aggregator (RNN/transformer) to approximate a belief state. This RL formalization yields agents that *read* incoming messages, *decide* among response options, and *adapt* behavior over time to maximize long-term social utility, providing a principled alternative to purely rule- or probability-based ABMs.

5.1.5. Graph Neural Networks

Graph Neural Networks (GNNs) provide a powerful framework for modeling user behaviors in online social networks, where agents (users) are naturally represented as nodes and their connections as edges [126]. In this setting, each agent u_i receives messages from its neighbors $\Gamma(u_i)$, processes them according to its internal state, and decides whether and what message to deliver in the next step. GNNs capture both the structural information of the network and the evolving states of users, enabling the simulation or prediction of realistic message-propagation patterns.

Specifically, at time t, each user u_i has a feature vector $\mathbf{h}_i^{(t)}$ representing its current state (e.g., preference profile, recent message embeddings). Incoming messages from neighbors can be encoded into message vectors $\mathbf{m}_{ij}^{(t)}$, aggregated as:

$$\mathbf{m}_i^{(t)} = \text{AGGREGATE}\left(\left\{\mathbf{m}_{ij}^{(t)} \mid u_j \in \Gamma(u_i)\right\}\right)\,, \tag{24}$$

where AGGREGATE$(\cdot)$ can be a mean, sum, or attention-weighted sum. The node update rule in a GNN layer combines the agent's current state and aggregated neighbor messages:

$$\mathbf{h}_i^{(t+1)} = \sigma\Big(\mathbf{W}_1 \mathbf{h}_i^{(t)} + \mathbf{W}_2 \mathbf{m}_i^{(t)}\Big) \tag{25}$$

where $\mathbf{W}_1$ and $\mathbf{W}_2$ are learnable weight matrices and $\sigma(\cdot)$ is a non-linear activation function. In attention-based GNNs, $\mathbf{m}_i^{(t)}$ can be computed using:

$$\alpha_{ij} = \frac{\exp\Big(\text{LeakyReLU}(\mathbf{a}^\top[\mathbf{W}\mathbf{h}_i^{(t)} \,\|\, \mathbf{W}\mathbf{h}_j^{(t)}])\Big)}{\sum_{k \in \Gamma(u_i)} \exp\Big(\text{LeakyReLU}(\mathbf{a}^\top[\mathbf{W}\mathbf{h}_i^{(t)} \,\|\, \mathbf{W}\mathbf{h}_k^{(t)}])\Big)}\,, \tag{26}$$

$$\mathbf{m}_i^{(t)} = \sum_{j \in \Gamma(u_i)} \alpha_{ij}\, \mathbf{W}\mathbf{h}_j^{(t)}\,, \tag{27}$$

where α_{ij} is the attention weight representing the relative influence of neighbor u_j on u_i.

The decision to deliver a new message is then modeled as a classification or generation problem:

$$p(a_i^{(t)} \mid \mathbf{h}_i^{(t+1)}) = \text{softmax}\Big(\mathbf{W}\mathbf{h}_i^{(t+1)}\Big)\,, \tag{28}$$

where $a_i^{(t)}$ denotes the action (e.g., ignore, reshare, modify, or create a new message) and $\mathbf{W}$ is a learnable parameter matrix.

In predictive settings, the GNN is trained to maximize the likelihood of observed actions; in simulation settings, $a_i^{(t)}$ is sampled from $p(\cdot)$ to update the network state. By iteratively updating node states and simulating message exchanges, GNN-based models can reproduce or forecast complex message diffusion behaviors that depend on both network topology and user-level dynamics, outperforming purely probabilistic or rule-based models in capturing higher-order interaction effects [33, 159].

5.2. Network-Topological Approaches

From a network topology perspective, two primary strategies are commonly employed to mitigate ideological isolation and enhance diversity of information exposure: (1) *Edge modifications*, and (2) *Node modifications* (via influence spread).

5.2.1. Edge Modifications

Edge modifications involve *adding*, *removing*, or *reweighting* specific edges in the network to alter existing interaction patterns. These interventions are designed to weaken the structural mechanisms, such as community boundaries or echo chambers, that reinforce ideological isolation.

Edge Addition. Adding cross-partition edges, links between nodes in opposing communities X and Y, can weaken segregation by increasing cross-group transition probabilities (P_{xy}, P_{yx}) and decreasing within-group persistence (P_{xx}, P_{yy}). Using the RWC definition in Equation 14, the edge-addition problem can be formulated as minimizing polarization after augmenting the graph with a small set of cross-cutting links [40]:

$$\mathrm{RWC}(G) \;=\; P_{xx}(G)\,P_{yy}(G) \;-\; P_{xy}(G)\,P_{yx}(G),$$

$$\min_{\Delta E \subseteq (X\times Y)\cup(Y\times X)} \mathrm{RWC}(G \oplus \Delta E) \quad \text{s.t.} \quad |\Delta E| \le k,$$

where $G \oplus \Delta E$ denotes the graph after adding the directed edges ΔE. A practical strategy is greedy selection by marginal gain:

$$\Delta\mathrm{RWC}(u,v) \;=\; \mathrm{RWC}(G) \;-\; \mathrm{RWC}(G \oplus \{(u,v)\})\,,$$

choosing up to k candidate edges $(u,v) \in X \times Y \cup Y \times X$ that yield the largest decrease in RWC. Depending on the mitigation goal, the same objective can be extended to target influential nodes by replacing RWC with A$-$RWC, or to discourage long runs within a single community by replacing it with D$-$RWC, thereby prioritizing edges that specifically reduce elite-level separation or increase inter-community transitions along random walks.

Building on this formulation, several heuristics and optimization methods have been proposed to identify effective cross-community links that minimize RWC or related polarization indices. Typical methods include:

High-degree bridging [40] Given a partition of graph G, i.e., $V' = \{X \cup Y\}$ with $\{X \cap Y\} = \varnothing$, define the candidate cross-community edges

$$E_\times \;=\; \{(u,v) \in X \times Y \;:\; (u,v) \notin E\}\,.$$

A simple and effective scoring rule is the degree-product:

$$s_{\deg}(u,v) \;=\; k(u)\,k(v),$$

where $k(\cdot)$ is node degree. Select the top-k edges by $s_{\deg}$. This favors bridges between structurally influential nodes, which (i) increases the cut size between X and Y by k, (ii) reduces modular segregation, and (iii) raises cross-community random-walk flow.

Greedy selection for edge addition Let $F(G)$ be a polarization objective to *minimize* (e.g., RWC in Equation 14 or modularity Q in Equation 12). Define the marginal gain of adding a candidate edge e to the current graph G_t as

$$\Delta_F(e \,|\, G_t) \;=\; F(G_t) \;-\; F(G_t + e)\,.$$

The greedy rule (for a budget of k edges) is

$$e_t^\star \;=\; \arg\max_{e \in E_\times} \Delta_F(e \,|\, G_{t-1})\,,$$

$$G_t \leftarrow G_{t-1} + e_t^\star, \quad t = 1, \ldots, k\,.$$

Edge Removal. A line of work studies *removing* a small number of *intra-community* links to weaken echo chambers while preserving overall connectivity. Recall that opinion dynamics are often modeled with the Friedkin–Johnsen (FJ) model, which is introduced in Equation 29.

The task is commonly cast as selecting a small deletion set ΔE (budget k) that minimizes a chosen *echo-chamber index* $\mathrm{EC}(G)$, subject to structural constraints:

$$\min_{\Delta E \subseteq E_{\text{intra}},\, |\Delta E| \le k} \mathrm{EC}\big(G \setminus \Delta E\big)\,,$$

subject to $(G \setminus \Delta E)$ remains connected and inter-community edges are retained. This formulation summarizes existing approaches that target reinforcing links rather than bridges, aiming to attenuate echo-chamber effects without fragmenting the network.

Targeted removal of high-betweenness intra-community edges Let $b(e)$ be the *edge betweenness* (number of shortest paths using e) [44]:

$$b(e) \;=\; \sum_{s \ne t} \frac{\sigma_{st}(e)}{\sigma_{st}}\,,$$

where σ_{st} is the number of shortest $s \to t$ paths and $\sigma_{st}(e)$ those that traverse e. Restrict candidates to $E_{\text{intra}} = \{(u,v) \in E \,:\, c_u = c_v\}$ and (under a connectivity constraint) remove the top-k edges by $b(e)$. The intuition is that deleting intra-community "highways" weakens insular clusters without destroying cross-community bridges.

Greedy conflict minimization [24] Let $R(G)$ denote a conflict/polarization *risk* (e.g., the average-case risk under Friedkin–Johnsen dynamics). At each step, remove the *intra*-community edge with the most significant marginal risk drop:

$$e^\star \;=\; \arg\max_{e \in E_{\text{intra}}} \Big(R(G) \;-\; R(G \setminus \{e\})\Big)\,.$$

A concrete instantiation uses steady-state opinions $\mathbf{x}^*(G)$ from the FJ model and an intra-edge disagreement objective

$$D_{\text{intra}}(G) \;=\; \sum_{(i,j) \in E_{\text{intra}}} \big(x_i^* - x_j^*\big)^2\,,$$

then greedily deletes edges maximizing $D_{\text{intra}}(G) - D_{\text{intra}}(G \setminus \{e\})$, subject to preserving connectivity. The intuition is that removing the links that most reduce within-group reinforcement weakens echo chambers.

Structural-bias (random-walk) reduction [50] Let $V_{\text{opp}(u)}$ be the opposite side of u. Define the *polarized bubble radius* (expected hitting time to the opposite side)

$$\text{BR}(u) \;=\; \mathbb{E}\big[\tau_{u\to v} \,\big|\, v \in V_{\text{opp}(u)}\big]\,,$$

and the global structural bias $\Phi(G) = \sum_{u\in V} \text{BR}(u)$. Select intra-community deletions that maximize the reduction

$$\Delta_\Phi(e) \;=\; \Phi(G) \;-\; \Phi\big(G \setminus \{e\}\big)\,,$$

again with connectivity safeguards. This is based on the intuition that deleting edges that keep random walks "circulating" inside one side shortens stochastic access to opposing views.

Edge Edition. Edge editing adjusts the weights of existing links to influence the dynamics of opinion. This approach fine-tunes interpersonal influence strengths, simulating how recommended systems or moderation algorithms amplify or suppress interactions.

Chitra and Musco augment the FJ opinion model by introducing a network administrator that adaptively reweights edges to minimize disagreement, and an algorithmic filtering mechanism that can induce filter bubbles [25]. In round r, users first update to the FJ equilibrium on the current graph:

$$z^{(r)} \;=\; \big(L^{(r-1)} + I\big)^{-1} s\,,$$

where $L^{(r-1)}$ is the Laplacian, s are innate opinions, and z are expressed opinions. Disagreement is

$$D_{G,z} \;= z^\top L z\,.$$

Given $z^{(r)}$, the administrator then chooses a new graph (i.e., edge weights W) by

$$G^{(r)} \;=\; \arg\min_{G\in\mathcal{S}} \; D_{G,\,z^{(r)}}\,,$$

subject to realistic constraints $\|W - W^{(0)}\|_F \le \varepsilon \|W^{(0)}\|_F$ and row-sum (degree) preservation $\sum_j W_{ij} = \sum_j W^{(0)}_{ij}$ for all i. Alternating these steps can *increase polarization dramatically* on real Reddit/Twitter graphs; a simple regularized variant (adding $\gamma\|W\|_F^2$) limits the filter-bubble effect while barely affecting disagreement.

5.2.2. Node Modifications

Node-level interventions to mitigate ideological isolation typically follow two approaches: (i) modifying existing node features, such as susceptibility, thresholds, or receptivity, to influence how opinions spread, often within an influence maximization framework; and (ii) introducing or designating special nodes, such as seeds or stubborn "mediator" agents, to inject cross-cutting information. The former is more prevalent in the literature, while the latter offers a complementary strategy for disrupting echo chambers. Both approaches are based on the social influence diffusion models.

Social Influence Diffusion. Social influence diffusion models describe how behaviors, opinions, or information propagate through networks. Two widely used frameworks are the *Independent Cascade* (IC) and *Linear Threshold* (LT) models [72].

In the *IC model*, each directed edge (u_j, u_i) has an activation probability $p_{ji} \in [0,1]$. Let A_t be the set of active nodes at time t. The probability that an inactive node u_i becomes active at $t+1$ is:

$$P\big(u_i \text{ activates at } t+1\big) = 1 - \prod_{u_j \in A_t \cap \Gamma(u_i)} \big(1 - p_{ji}\big)\,,$$

where $\Gamma(u_i)$ is the set of in-neighbors of u_i. Once active, a node remains active for all subsequent steps.

In the *LT model*, each node u_i has a threshold $\theta_{u_i} \in (0,1]$ and receives influence weights $w_{u_j u_i}$ from its neighbors, with $\sum_{u_j\in\Gamma(u_i)} w_{u_j u_i} \le 1$. At each step, an inactive node u_i becomes active if the total influence from its active neighbors reaches or exceeds its threshold:

$$\sum_{u_j \in \Gamma(u_i)\cap A_t} w_{u_j u_i} \;\ge\; \theta_{u_i}\,.$$

As in the IC model, activation is irreversible.

Both models naturally capture the stochastic and progressive nature of influence and are submodular under common assumptions, enabling efficient seed selection through greedy algorithms with approximation guarantees [72]. In the context of breaking echo chambers, IC and LT can be used to model the spread of interventions, such as pro-diversity content or belief adjustments, across communities.

Susceptibility tuning via influence maximization. The key idea is to target highly insular communities with interventions that lower their susceptibility to within-group influence, thereby weakening echo chambers and promoting cross-cutting exposure [117]. The diffusion process (IC/LT) is introduced for delivering interventions with the FJ opinion dynamics:

$$x(t+1) = \Lambda W x(t) + (I - \Lambda)s\,, \tag{29}$$

where $\Lambda = \text{diag}(\lambda_i)$ encodes node susceptibilities ($\lambda_i \in [0,1]$), W is the influence weight matrix, and s are innate opinions. The steady state is

$$x^\star(\Lambda) = (I - \Lambda W)^{-1}(I - \Lambda)s\,.$$

Given a seed set $S \subseteq U$ with $|S| \le k$, diffusion yields an exposure level $\eta_i(S)$ for each node u_i. Exposure updates

susceptibilities via a decreasing function

$$\lambda_i' = g(\lambda_i, \eta_i(S)), \quad \frac{\partial g}{\partial \eta} \leq 0\,,$$

e.g., $g(\lambda, \eta) = \lambda(1 - \alpha\eta)$ or $g(\lambda, \eta) = \max\{\lambda - \alpha\eta, \lambda_{\min}\}$. Let $\Lambda' = \mathrm{diag}(\lambda_i')$ and $x^\star(S)$ be the resulting FJ equilibrium.

The influence maximization problem is then:

$$\max_{S:|S|\leq k} \; f_{\text{spread}}(S) \quad \text{s.t.} \quad \mathrm{EC}(x^\star(S)) \text{ is minimized}\,,$$

where $f_{\text{spread}}(S)$ is the standard submodular spread objective under IC/LT [72] and $\mathrm{EC}(\cdot)$ measures post-intervention echo-chamber intensity [40].

Diversity-aware seeding across communities. Diversity-aware or fairness-aware influence maximization seeks seed sets that not only maximize total spread but also distribute exposure across distinct subpopulations (e.g., communities), so that isolated or minority groups receive meaningful reach, thereby mitigating echo chamber effects and improving cross-cutting exposure. Formalizations typically account for group structure (e.g., detected or given communities) and optimize spread subject to balance or coverage criteria rather than concentrating influence in a few large or central groups [84, 141, 171].

Let the node set U be partitioned into communities $\{C_1, \ldots, C_m\}$. Under IC/LT diffusion, let $\sigma_c(S)$ be the expected number of activations in C_c from seed set S. A common diversity-aware objective is a concave, weighted coverage:

$$\max_{S:\,|S|\leq k} \; \sum_{c=1}^{m} w_c\, f\big(\sigma_c(S)\big)\,,$$

where $f(\cdot)$ is a concave nondecreasing rewarding balance and $w_c \geq 0$ encodes group priorities (e.g., to up-weight underserved communities). To avoid size bias, a normalized max-min fairness variant is often used:

$$\max_{S:\,|S|\leq k} \; \min_{c} \frac{\sigma_c(S)}{|C_c|}\,,$$

which maximizes the worst-case fraction of communities affected. These formulations encourage the spread of seeds across communities, allowing influence to penetrate otherwise insular groups [84, 141].

When the objective is maximizing the influence coverage, the spread function is monotone and submodular, enabling a simple greedy algorithm with a $(1-1/e)$ approximation guarantee [72]. Efficient implementations build on this structure, notably CELF for lazy marginal-gain evaluation [82] and reverse-influence-sampling methods such as TIM and IMM, which achieve near-optimal accuracy with substantially reduced running time [137]. In contrast, diversity- or fairness-aware objectives (e.g., max-min coverage across communities, even after normalization by group size) are generally not submodular. Consequently, classic guarantees do not carry over, and one must resort to specialized formulations and approximation schemes developed for fair or community-diversified influence maximization [84, 141, 168]. A complementary line of work blends spread with diversity through concave utility aggregation over group-level reach, for which tailored approximation strategies have also been proposed within the influence maximization framework [168].

Opinion-aware seeding with stubborn nodes. The goal is to steer long-run beliefs by placing a small number of stubborn (*zealot*) nodes whose opinions remain fixed, thereby anchoring the dynamics. In opinion-exchange models, introducing fixed-opinion agents can reshape the equilibrium of the remaining population [157]. With appropriate placement and values, this can reduce isolation across communities by weakening within-group reinforcement and injecting bridging signals. This phenomenon is well established for DeGroot/FJ-type processes and for binary variants of the voter model with stubborn nodes [43, 164].

Formally, let $Z \subseteq U$ be the zealot set with fixed opinions o_Z, and let $R = U \setminus Z$ denote adaptive nodes. Under a linear update (DeGroot/FJ) with block partitioning of the influence matrix,

$$x_R^\star \;=\; H\, o_Z \;+\; h\,,$$

where the matrices H and vector h depend on network topology and weights (e.g., via the fundamental matrix $(I - W_{RR})^{-1}$). Thus, the equilibrium on R is an affine image of o_Z, making both the *locations* and *opinions* of zealots the key control variables.

A design problem of interest is to choose up to k zealots so as to minimize a post-equilibrium fragmentation measure:

$$\min_{Z\subseteq V,\,|Z|\leq k} \; \mathrm{EC}\big(x^\star(Z)\big)\,,$$

where $\mathrm{EC}(\cdot)$ quantifies echo-chamber intensity (e.g., controversy or cross-community separation). Although this combinatorial objective is generally challenging, several related goals exhibit favorable structure; for example, when the target is to increase mean opinion toward a reference value, the set function over Z is monotone submodular, enabling greedy selection with approximation guarantees. Empirically and theoretically, the optimized placement of zealots can shift the means and reduce polarization in real networks [61].

5.3. Content-based Control Mechanisms

Content-based control steers each user's exposure toward broader yet still relevant items, anchored in the exposure centroid $\boldsymbol{\mu}_i(t)$ and its dispersion. We organize methods into three families: pre-processing (data and candidate-pool shaping), in-processing (training-time objectives with constraints), and post-processing (re-ranking

or allocation with explicit control), following the established taxonomy in Wang et al. [152]. These interventions must navigate the accuracy–diversity–utility trade-off, which we address through multi-objective formulations, constrained losses, and constraint-aware re-ranking. Recent surveys further systematize how fairness and diversity objectives interact across user-side and item-side perspectives in recommendation pipelines [170].

5.3.1. Pre-processing Approach

Pre-processing approaches mitigate ideological isolation by directly manipulating input data before model training, making it model-agnostic and relatively easy to integrate [14]. Most existing works classify pre-processing methods into three categories: *data relabeling*, *re-sampling*, and *data modification* [92, 152]. These strategies address bias in training data distributions, thereby indirectly reducing filter bubbles, selective exposure, and polarization.

Data Relabeling. Data relabeling modifies observed user feedback or labels to correct bias. For example, popularity bias can lead to frequently exposed items receiving disproportionately high labels, thereby reinforcing homogenized recommendations. Relabeling adjusts such labels to reflect underlying user preference rather than exposure frequency [23, 73]. This approach is intuitive and lightweight, but its effectiveness depends on the accuracy of relabeling heuristics.

Data Re-sampling and Re-weighting. Resampling methods adjust the distribution of training data by undersampling overrepresented items or oversampling underrepresented content. By balancing exposure frequencies, they enhance diversity and reduce the reinforcement of filter bubbles. A closely related strategy is *re-weighting*, which adjusts the loss contribution of each training instance rather than altering the data itself. For instance, Gupta et al. propose a re-weighting technique that balances the training loss between popular and non-popular items, ensuring equitable representation during model learning and reducing the reinforcement of filter bubbles [48], while Abdollahpouri et al. provide a systematic empirical analysis of popularity bias and its connection to fairness and calibration in recommendation [1].

Data Modification. Data modification extends beyond resampling and relabeling by actively transforming or augmenting training data. A representative example is counterfactual data augmentation, where new user-item interactions are generated to simulate alternative exposure histories. For instance, Wang et al. propose a sequential recommendation framework that constructs synthetic interaction sequences to increase exposure diversity and mitigate bias arising from imperfect or sparse training data [153]. Recent studies treat data modification as a core pre-processing strategy, noting its role in intervening on interaction or feature representations, but also cautioning about potential distribution shifts or noise introduced by synthetic data when preserving recommendation accuracy [92].

In summary, pre-processing approaches are attractive due to their model-agnostic nature and ease of deployment. Relabeling focuses on correcting biased labels, resampling and re-weighting balance exposure distributions, and modification leverages augmentation or adversarial techniques for stronger debiasing. However, pre-processing alone cannot fully eliminate systemic bias and is often complemented by in-processing and post-processing approaches to comprehensively mitigate ideological isolation.

5.3.2. In-processing Approach

In-processing methods embed the mitigation strategy inside the learning or scoring pipeline. Specifically, they modify either the training objective or the scoring function so that the recommender optimizes relevance and mitigation criteria jointly during learning or inference. This makes them expressive, targeting bias, diversity, or fairness, while preserving end-to-end optimization. These methods typically augment the loss or scores as:

$$\mathcal{L} = \mathcal{L}_{rec} + \lambda \cdot \mathcal{L}_{mitigation}, \quad \hat{y}_{ui} = f_{rec}(u,i) + \Omega(u,i)\,, \tag{30}$$

where $\mathcal{L}_{rec}$ is the standard recommendation loss, $\mathcal{L}_{mitigation}$ encodes bias control, fairness constraints, or diversity regularization, and $\lambda > 0$ tunes the trade-off. $\hat{y}_{ui}$ represents the *predicted* relevance score for user u on item i. The base score $f_{rec}(u,i)$ models user–item relevance, and $\Omega(u,i)$ is an in-model adjustment, such as debiasing corrections, diversity terms, or counterfactual offsets. We group in-processing into five families as follows.

Regularization-based Mitigation. Here, explicit penalties or corrective terms are added to reduce exposure/popularity bias during learning. For instance, discrete choice modeling frameworks adjust user-item click probabilities by incorporating exposure frequency:

$$P(i|u,E) = \frac{\exp(\mu_{ui} - \delta \cdot E_i)}{\sum_{j \in E} \exp(\mu_{uj} - \delta \cdot E_j)}\,, \tag{31}$$

thereby attenuating the over-representation of frequently exposed items [78]. In addition to the commonly adopted fairness and exposure constraints, Yang et al. introduce content-suppression mechanisms to reduce redundancy and enhance diversity in recommendations [163].

Latent Representation Modulation. These methods perturb latent representations to encourage diversity along designated semantic directions while preserving relevance. The TD-VAE-CF model achieves this by shifting embeddings along Concept Activation Vectors (CAVs):

$$z'_u = z_u + \alpha \cdot v_{CAV}, \quad \alpha > 0\,, \tag{32}$$

which allows the system to enhance diversity while preserving overall relevance [39].

Counterfactual Corrections. Counterfactual in-processing disentangles causal factors, e.g., preference vs. item popularity, and subtracts the non-preferential component at score time. In MACR, a debiased score removes a direct popularity effect:

$$\hat{y}_{ui}^{debias} = \hat{y}_{ui} - \gamma \cdot \sigma(\hat{y}_i)\,, \tag{33}$$

thus reducing systemic popularity bias without altering the user-item matching signal [154].

Reinforcement Learning (RL). Framing recommendation as an MDP enables temporal exposure control. With state $s_t = (u, H_t)$ (user u and history H_t), action a_t (slate or item), and composite reward:

$$R_t = \eta \cdot Acc(a_t) + (1-\eta) \cdot Div(a_t)\,, \tag{34}$$

an RL agent learns to trade off short-term accuracy and long-term exposure diversity, mitigating filter-bubble reinforcement [93].

Multi-Objective Optimization. Multi-objective formulations encode the accuracy-diversity trade-off directly:

$$\max_{\theta} \Big[\alpha \cdot Acc(\theta) + (1-\alpha) \cdot Div(\theta)\Big]\,, \tag{35}$$

where $Acc(\theta)$ measures predictive accuracy, $Div(\theta)$ quantifies exposure diversity, and α tunes the trade-off. Such formulations provide principled ways to reconcile competing objectives in recommender systems [121].

While all in-processing approaches intervene at the model level, their mechanisms differ: regularization-based and counterfactual methods correct statistical biases, latent modulation reshapes semantic representations, RL optimizes sequential decisions, and multi-objective optimization explicitly encodes trade-offs. These approaches demonstrate the versatility of in-processing techniques for increasing fairness and mitigating ideological isolation while retaining relevance.

5.3.3. Post-processing Approach

Post-processing methods mitigate ideological isolation by adjusting the items produced by a base recommender model [136]. Unlike in-processing methods that modify the learning objective or pre-processing approaches that manipulate input data, post-processing provides a flexible, model-agnostic layer for incorporating fairness, diversity, novelty, and neutrality constraints.

Given a candidate set C and target list length K, the re-ranked set $R^*(u)$ for user u is defined as:

$$R^*(u) = \arg \max_{S \subseteq C, |S|=K} \alpha \cdot \underbrace{\sum_{i \in S} f_{rec}(u,i)}_{\text{relevance}} + \beta \cdot \underbrace{D(S)}_{\text{diversity/fairness}}\,, \tag{36}$$

where $f_{rec}(u,i)$ measures the relevance of item i for user u, $D(S)$ quantifies the diversity or fairness of the set S, and $\alpha, \beta \in [0,1]$ control the trade-off. Different methodological schools instantiate this formulation in distinct ways.

Greedy Re-ranking. Greedy re-ranking incrementally constructs or revises a top-K list from the base recommender by repeatedly selecting the item that maximizes a weighted combination of relevance and the marginal gain in a set-level criterion, e.g., diversity or fairness. Two classics are MMR [20], which trades off relevance and novelty to reduce redundancy, and xQuAD [130], which increases coverage over multiple topical aspects during selection. Both are widely used for diversity-aware re-ranking and extend cleanly to recommendation settings. Recent work further enriches the diversity term with LLM-derived semantic signals while keeping the greedy scaffold. At each step, a typical rule is

$$i^* = \arg \max_{i \in C \setminus S} \lambda f_{rec}(u,i) + (1-\lambda)\Delta D(i,S)\,,$$

where $\Delta D(i,S)$ is the marginal diversity gain of adding i to S.

Recent work explores the use of an LLM as a re-ranker. Specifically, given a candidate list, the LLM is prompted to output a more diverse ranking. Empirically, LLM-based re-ranking improves diversity over random re-ranking but remains below traditional greedy methods (e.g., MMR or xQuAD) on relevance-aware metrics; importantly, the LLM operates as an alternative re-ranker rather than injecting LLM signals into $\Delta D(i,S)$ within a greedy step [21]. Beyond using LLMs as re-rankers, Wang et al. propose a responsible graph-based recommendation pipeline, where a Generative-AI module is applied to synthesize contextualized content based on the user's belief graph and nudge the users toward a more balanced information perception, mitigating belief filter bubbles [149].

Graph-based and Optimization-based Re-ranking. These methods seek a globally balanced set by solving the re-ranking objective over the entire subset S (e.g., $S \subseteq C$, $|S| = K$) rather than making greedy, per-item choices. A common instantiation builds a user-item bipartite graph and imposes flow or coverage constraints to improve catalog balance. For example, FairMatch [103] constructs a bipartite graph and uses maximum-flow optimization to boost aggregate diversity by increasing the coverage of underrepresented items. Likewise, Antikacioglu and Ravi minimize discrepancy from a target exposure distribution via minimum-cost flow, achieving explicit control over exposure and coverage [5]. Compared to greedy heuristics, these flow-based formulations offer stronger fairness guarantees and clearer constraint handling, but their higher computational cost can limit scalability in large, real-time recommender systems.

Probabilistic Re-ranking. Probabilistic approaches replace an explicit set function $D(S)$ with a likelihood over subsets. A prominent example is the *Determinantal Point Process* (DPP), which defines a distribution $P(S) \propto \det(L_S)$ over item sets S, where $L \succeq 0$ is a kernel

encoding both item quality and similarity [79]. For top-K re-ranking, a common MAP objective is

$$\max_{S \subseteq C,\ |S|=K} \log\det(L_S).$$

The determinant acts as a repulsive potential: it grows when items are individually high-quality yet mutually dissimilar, thereby discouraging redundancy. A standard parameterization is the quality-similarity factorization

$$L = Q^{1/2} S Q^{1/2}, \tag{37}$$

$$Q = \mathrm{diag}(q_1, \ldots, q_{|C|}), \quad S_{ii} = 1, \quad S_{ij} \in [-1, 1], \tag{38}$$

where q_i is a per-item relevance (quality) score and S captures pairwise similarity. Under this factorization, $\log\det(L_S)$ naturally balances accuracy (via q_i) and diversity (via S). Although the exact MAP is generally expensive on large candidate pools, greedy maximization of $\log\det(\cdot)$, a monotone submodular objective, admits an efficient $(1 - 1/e)$-approximation with incremental updates, making DPP re-ranking practical at scale.

Novelty-based and Serendipity-based Re-ranking. Beyond diversity, post-processing can target *novelty* (items the user has not seen or is unlikely to know) and *serendipity* (items that are both unexpected and useful). A common formalization augments the set objective with per-item novelty or serendipity scores [76]:

$$\max_{S \subseteq C,\ |S|=K} \alpha \sum_{i \in S} f_{rec}(u, i) \ + \ \beta \sum_{i \in S} f_{\mathrm{ser}}(u, i),$$

where $f_{rec}(u, i)$ denotes base-model relevance and $f_{\mathrm{ser}}(u, i)$ rewards useful surprise. Following prior work, f_{ser} is often modeled as the product of *usefulness* and *unexpectedness* [161]:

$$f_{\mathrm{ser}}(u, i) = \mathrm{Rel}(u, i) \cdot \mathrm{Unexp}(u, i), \tag{39}$$

$$\mathrm{Unexp}(u, i) = 1 - \max_{j \in H_u} \mathrm{sim}(i, j), \tag{40}$$

with H_u the user's history and $\mathrm{sim}(\cdot, \cdot)$ a content or embedding similarity. This construction promotes items that remain relevant while being dissimilar to what the user already knows. In practice, serendipity-oriented re-ranking can improve user satisfaction and long-term engagement, provided the degree of surprise is calibrated. The main design choice is how to instantiate Unexp, e.g., distance to user profile, distance to prior exposures, or popularity-inverse signals, and how to balance α vs. β for the target application.

In summary, heuristic approaches provide efficient local approximations, optimization-based methods offer global diversity guarantees, probabilistic approaches deliver principled mathematical formulations, and novelty-oriented methods improve long-term user engagement. All these post-processing techniques can be viewed as instances of the generalized multi-objective framework, differing mainly in how they instantiate the diversity function $D(S)$ and whether they optimize it locally, globally, probabilistically, or with novelty-oriented constraints.

5.4. Dual Role of Large Language Models

LLMs increasingly influence information exposure in online platforms, acting both as a potential driver of ideological isolation and as a tool for mitigation. This complements the simulation-oriented treatment of LLMs in the discussion of LLM-based interventions.

LLMs can also drive ideological isolation. Recent studies suggest that LLMs may contribute to homogenized exposure through several mechanisms. First, measurable political biases have been reported in widely used conversational models [112, 128, 129]. Second, LLM outputs diverge in factual emphasis and framing across users. Lazovich [80] formalizes this as an LLM analog of filter bubbles and documents affective polarization in persona-conditioned ChatGPT outputs. In a related vein, Hackenburg et al. [49] show that partisan-role-played GPT-4 messages can match, and on some issues exceed, the persuasive impact of human persuasion experts, raising the prospect of large-scale automated narrative reinforcement. Third, simulation studies integrating LLM agents into recommender systems reproduce the exposure-narrowing characteristic of filter bubbles and personalized recommendation loops [135].

LLMs can also serve as tools for mitigating ideological isolation. The same capabilities also enable new intervention strategies. LLMs can support cross-ideological content rewriting [52], bias detection and stance analysis [118], and user-controllable diversification through the generation of counter-perspective content [149]. They have also been used to facilitate more constructive political discussions [8] and to serve as diversity-aware re-rankers in recommendation systems [21].

5.5. Cross-Family Mitigation Comparison

The mitigation families surveyed in previous sections differ in effectiveness, scalability, robustness, interpretability, and deployment cost. Figure 4 provides a visual comparison across these characteristics. The ratings are based on the deployment, simulation, and benchmark studies cited in the corresponding subsections rather than a single head-to-head experiment.

Several broad patterns emerge. (*i*) Network-topological interventions can effectively reduce structural isolation but are often difficult to deploy at platform scale. (*ii*) Pre-processing approaches are model-agnostic and low-cost to deploy, but cannot eliminate systemic bias alone and are typically combined with in- and post-processing methods. (*iii*) Post-processing re-ranking remains the most practical approach because it is model-agnostic and readily integrated into existing systems. (*iv*) In-processing methods generally offer stronger diversity–accuracy trade-offs but require retraining and additional validation. (*v*) LLM-based approaches provide flexible support for content diversification and stance-aware recommendation, although they inherit biases from the underlying models.

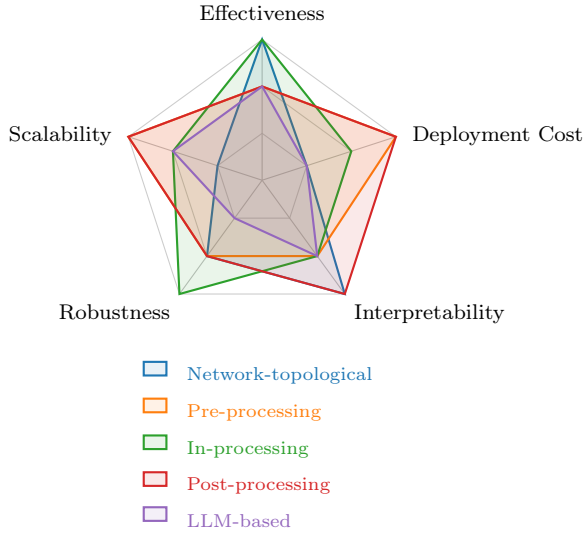

Figure 4: Radar comparison of mitigation families across five evaluation criteria. The user-modeling/simulation family is omitted for visual clarity as it is primarily a diagnostic tool rather than a user-facing intervention.

6. Ethical Considerations

Computational mitigation of ideological isolation is not value-neutral. Choices about what to diversify, how aggressively to intervene, and who controls these interventions inevitably reflect normative assumptions.

Two issues are particularly relevant to the mitigation strategies reviewed in Section 5: algorithmic paternalism and user autonomy, and cross-cultural variation in the meaning of ideological diversity.

6.1. Paternalism, Authorization, and Transparency

Many mitigation strategies reviewed in Section 5, including diversity-aware recommendation, network interventions, and LLM-driven content generation, modify users' information exposure without explicit awareness. While such interventions may reduce ideological isolation, they can also raise concerns about algorithmic paternalism when user preferences are overridden in pursuit of broader epistemic goals [19].

Two principles are particularly important. First, users should be able to control diversity-promoting interventions through informed opt-in mechanisms and adjustable settings. Empirical studies suggest that user controls over feed curation can improve perceived diversity, although their effect on actual political diversity remains mixed [96]. Second, systems should provide transparent explanations of why particular viewpoints or recommendations are presented. This is consistent with recent regulatory developments, such as the EU Digital Services Act, which emphasizes transparency and user choice in recommender systems [31, 81].

More broadly, interventions should support users' ability to navigate diverse information environments rather than silently replacing their judgement [97]. Consequently, mitigation studies should evaluate not only reductions in isolation metrics but also the transparency and explainability of the intervention process.

6.2. Cultural and Regulatory Variation

The notion of ideological diversity is not culturally invariant. Much computational work adopts a left-right ideological spectrum that reflects political divisions common in Western democracies [53]. However, salient dimensions of disagreement may instead involve religion, ethnicity, language, regional identity, or other social factors in different contexts [2]. Consequently, metrics and models developed for one setting may not transfer directly to another.

Regulatory expectations also vary across jurisdictions. The EU Digital Services Act emphasizes user-side transparency and choice, including the right to opt for non-profiling-based recommendations [31, 81]. China's Provisions on the Administration of Algorithmic Recommendation in Internet Information Services, effective March 2022, require platforms to register algorithms, offer opt-out from personalization, and explicitly avoid creating "information cocoons", framing the issue in collective rather than individual terms [160]. These differences suggest that mitigation objectives and evaluation criteria may need to be adapted to local social and regulatory contexts.

Two implications follow for computational research. First, the ideological space $\mathcal{I}$ introduced in Section 3 should be specified according to the application context rather than treated as a universal one-dimensional axis, and may require culturally grounded annotations [2]. Second, mitigation strategies should be evaluated not only by their effectiveness in reducing ideological isolation but also by their suitability for the cultural and regulatory environments in which they are deployed.

7. Open Challenges and Future Directions

Having surveyed how ideological isolation can be defined, measured, and mitigated, we now examine what is still required to develop reliable, deployable solutions. This section identifies several cross-cutting open challenges spanning theory, measurement, and system design, and outlines corresponding directions for future research.

7.1. Advancing Causal Understanding

A profound limitation of current quantification metrics is their correlational nature. While such metrics can detect the presence of ideological isolation, they often cannot disentangle its underlying causes [99]. For example, a user's ideologically homogeneous feed could arise from either the platform's personalization algorithm or the user's selective exposure and homophilic interactions [107, 108].

Distinguishing between these drivers is crucial for designing effective interventions. The question here is whether the recommender system should be reconfigured

or whether users should be nudged toward more diverse engagement patterns. Future research should adopt causal inference frameworks to distinguish between algorithmic influence and user agency. Approaches such as natural experiments, instrumental variable estimation, and counterfactual modeling, when applied to interaction logs, can help quantify the causal impact of platform-level mechanisms. Moreover, constructing causal models of belief evolution, building on the reinforcement dynamics described in Section 3.1.5, will be essential for identifying the most effective and ethically appropriate levers for change.

7.2. Balancing Competing Objectives in Recommender Systems

A central dilemma in the design of content-based control mechanisms (Section 5) is the perceived trade-off between recommendation accuracy and ideological diversity [95, 119, 121]. Many existing in-processing and post-processing approaches risk degrading the perceived utility or relevance of recommendations in pursuit of diversity. If users perceive diversified recommendations as less engaging or misaligned with their interests, they may disengage, undermining the intent to broaden ideological exposure.

A promising research direction is to optimize for a Pareto frontier between accuracy and diversity, rather than treating them as opposing objectives. This can be achieved through adaptive, personalized diversification, in which the degree of exposure to diverse content is dynamically adjusted based on individual user tolerance, susceptibility parameters, and contextual features. In addition, user-controllable interfaces, which allow individuals to calibrate their own diversity-exposure settings, could help reconcile this trade-off by aligning system interventions with user autonomy and engagement preferences.

7.3. Cross-Platform and Longitudinal Measurement of Ideological Isolation

Another critical challenge is to move beyond single-platform, one-shot studies and actually measure ideological isolation across multiple platforms and over time. In reality, people combine Facebook, X, YouTube, TikTok, news sites, search, and messaging apps into a single "media diet", so isolation is a property of this whole ecosystem, not of Twitter or Facebook alone. Comparative work already shows that echo chambers and polarization appear very different across platforms such as Facebook, Twitter/X, Reddit, and Gab, suggesting that platform-specific snapshots tell only part of the story [26]. Methodologically, cross-platform identity linkage is difficult, limiting our view of how different forms of isolation co-evolve across a person's online environment.

Future work should therefore build multiplex and longitudinal measurement frameworks. One direction is to model a user's environment as a multilayer exposure network, where each layer represents a platform. A second direction is to define isolation trajectories for individuals and groups, tracking how their exposure diversity, cross-cutting contacts, and stance distributions change before, during, and after key events. Finally, agent-based and opinion-dynamics simulations on multiplex networks can serve as testbeds for stress-testing proposed cross-platform metrics under realistic sampling and logging constraints.

7.4. Robustness to Adversarial and Strategic Amplification

A key challenge for measuring and mitigating ideological isolation is that the information environment is not purely "organic": it is actively shaped by adversarial and strategic actors. Social bots, coordinated disinformation campaigns, and manipulation of trending or recommendation systems can all artificially amplify specific narratives, hashtags, or accounts [54]. Recent work shows that relatively small, well-designed campaigns can dramatically increase exposure to harmful or misleading content, exploiting follow networks and recommendation algorithms to maximize coverage [140]. These operations can deepen echo chambers, manufacture apparent consensus, or crowd out moderate content, thereby distorting any attempt to measure "natural" structural, content-based, or interactional isolation. Yet most existing metrics and models implicitly assume that users, content, and platforms behave non-strategically; they rarely distinguish organic amplification from coordinated or adversarial activity.

Future work needs to treat adversarial and strategic amplification as a first-class concern in the study of ideological isolation. One direction is to design manipulation-aware isolation metrics and measurement pipelines that explicitly detect and discount exposures driven by bots, coordinated campaigns, or inauthentic accounts. Another is to build stress-testing and red-teaming frameworks to see how easily diversity-promoting or depolarization algorithms can be gamed. Finally, robustness methods from recommender systems and graph learning should be directly linked to isolation objectives, so that systems are evaluated not only on accuracy but also on their resistance to being weaponized to create artificial echo chambers.

8. Conclusion

This survey has examined ideological isolation in online social networks from a computational perspective, with a particular focus on formal definitions, metrics, computational modeling, and mitigation strategies.

Our review contributes in four main ways. *First*, we present a unified formal framework that characterizes ideological isolation and its underlying mechanisms, linking user beliefs, content visibility, exposure, and reinforcement dynamics within a shared ideological space. *Second*, we harmonize a multi-view measurement toolkit spanning network-based, content-based, and behavior-based indicators, using consistent notation. We further review evidence from English-, Chinese-, and Indian-language platforms. *Third*, we organize computational mitigation

strategies across modeling and system layers, including Bayesian and agent-based models, LLM-driven simulation, reinforcement learning, and graph neural networks, as well as network-topological and recommendation-pipeline interventions. We analyse the dual role of LLMs in both driving and mitigating ideological isolation. *Fourth*, we examine ethical considerations in computational mitigation, including algorithmic paternalism, user autonomy and transparency, and cross-cultural and regulatory variation in ideological diversity.

The open challenges identified in this survey point to an inherently interdisciplinary research agenda that integrates network science, machine learning, causal inference, and social science. Addressing these challenges calls for responsible, interpretable, and human-centered AI systems that balance personalization with information diversity, thereby fostering healthier digital public spheres and mitigating ideological fragmentation.

References

[1] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. The connection between popularity bias, calibration, and fairness in recommendation. In *Proceedings of the 14th ACM conference on recommender systems*, pages 726–731, 2020.

[2] Pushkal Agarwal, Kiran Garimella, Sagar Joglekar, Nishanth Sastry, and Gareth Tyson. Characterising user content on a multi-lingual social network. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 2–11, 2020.

[3] Mukhtar Ahmmad, Khurram Shahzad, Abid Iqbal, and Mujahid Latif. Trap of social media algorithms: A systematic review of research on filter bubbles, echo chambers, and their impact on youth. *Societies*, 15(11):301, 2025.

[4] Mónika Andok. Religious filter bubbles on digital public sphere. *Religions*, 14(11):1359, 2023.

[5] Arda Antikacioglu and R Ravi. Post processing recommender systems for diversity. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 707–716, 2017.

[6] Md Sanzeed Anwar, Grant Schoenebeck, and Paramveer S. Dhillon. Filter bubble or homogenization? disentangling the long-term effects of recommendations on user consumption patterns. In *Proceedings of the ACM Web Conference 2024 (WWW)*, pages 123–134, 2024.

[7] Qazi Mohammad Areeb, Mohammad Nadeem, Shahab Saquib Sohail, Raza Imam, Faiyaz Doctor, Yassine Himeur, Amir Hussain, and Abbes Amira. Filter bubbles in recommender systems: Fact or fallacy—a systematic review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 13 (6):e1512, 2023.

[8] Lisa P Argyle, Christopher A Bail, Ethan C Busby, Joshua R Gubler, Thomas Howe, Christopher Rytting, Taylor Sorensen, and David Wingate. Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale. *Proceedings of the National Academy of Sciences*, 120 (41):e2311627120, 2023.

[9] Eytan Bakshy, Solomon Messing, and Lada A Adamic. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132, 2015.

[10] Pablo Barberá. Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. *Political analysis*, 23(1):76–91, 2015.

[11] Faza Fawzan Bastarianto, Thomas O Hancock, Charisma Farheen Choudhury, and Ed Manley. Agent-based models in urban transportation: review, challenges, and opportunities. *European Transport Research Review*, 15(1): 19, 2023.

[12] Alessandro Bessi, Fabiana Zollo, Michela Del Vicario, Michelangelo Puliga, Antonio Scala, Guido Caldarelli, Brian Uzzi, and Walter Quattrociocchi. Users polarization on facebook and youtube. *PloS one*, 11(8):e0159641, 2016.

[13] Christopher M Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[14] Alessandro Bondielli and Francesco Marcelloni. A survey on fake news and rumour detection techniques. *Information sciences*, 497:38–55, 2019.

[15] Alexandre Bovet and Hernán A Makse. Influence of fake news in twitter during the 2016 us presidential election. *Nature communications*, 10(1):7, 2019.

[16] Emanuele Brugnoli, Matteo Cinelli, Walter Quattrociocchi, and Antonio Scala. Recursive patterns in online echo chambers. *Scientific Reports*, 9(1):20118, 2019.

[17] Axel Bruns. Echo chamber? what echo chamber? reviewing the evidence. In *6th Biennial Future of Journalism Conference (FOJ17)*, 2017.

[18] Laura Burbach, Patrick Halbach, Martina Ziefle, and André Calero Valdez. Bubble trouble: Strategies against filter bubbles in online social networks. In *International Conference on Human-Computer Interaction*, pages 441–456. Springer, 2019.

[19] Christopher Burr, Nello Cristianini, and James Ladyman. An analysis of the interaction between intelligent software agents and human users. *Minds and Machines*, 28:735–774, 2018.

[20] Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336, 1998.

[21] Diego Carraro and Derek Bridge. Enhancing recommendation diversity by re-ranking with large language models. *ACM Transactions on Recommender Systems*, 4(2):1–40, 2025.

[22] Yusuf Mücahit Çetinkaya, Vahid Ghafouri, Guillermo Suarez-Tangil, Jose Such, and Tuğrulcan Elmas. Cross-partisan interactions on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 324–340, 2025.

[23] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. Bias and debias in recommender system: A survey and future directions. *ACM Transactions on Information Systems*, 41(3):1–39, 2023.

[24] Xi Chen, Jefrey Lijffijt, and Tijl De Bie. Quantifying and minimizing risk of conflict in social networks. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1197–1205, 2018.

[25] Uthsav Chitra and Christopher Musco. Analyzing the impact of filter bubbles on social network polarization. In *Proceedings of the 13th international conference on web search and data mining*, pages 115–123, 2020.

[26] Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. The echo chamber effect on social media. *Proceedings of the national academy of sciences*, 118(9):e2023301118, 2021.

[27] Jialing Dai, Jianming Zhu, and Guoqing Wang. Opinion influence maximization problem in online social networks based on group polarization effect. *Information Sciences*, 609:195–214, 2022.

[28] Henrique Ferraz de Arruda, Felipe Maciel Cardoso, Guilherme Ferraz de Arruda, Alexis R Hernández, Luciano da Fontoura Costa, and Yamir Moreno. Modelling how social network algorithms can influence opinion polarization. *Information Sciences*, 588:265–278, 2022.

[29] Jinglong Duan, Shiqing Wu, Weihua Li, Quan Bai, Minh Nguyen, and Jianhua Jiang. Botdmm: Dual-channel multimodal learning for llm-driven bot detection on social media. *Information Fusion*, page 103758, 2025.

[30] Eitan Elaad. Tunnel vision and confirmation bias among police

investigators and laypeople in hypothetical criminal contexts. *Sage Open*, 12(2):21582440221095022, 2022.

[31] European Union. Regulation (EU) 2022/2065 of the European Parliament and of the Council on a single market for digital services (digital services act). Official Journal of the European Union, L 277/1, 2022. Effective 17 February 2024.

[32] Max Falkenberg, Fabiana Zollo, Walter Quattrociocchi, Jürgen Pfeffer, and Andrea Baronchelli. Patterns of partisan toxicity and engagement reveal the common structure of online political communication across countries. *Nature Communications*, 15(1):9560, 2024.

[33] Wenqi Fan, Yao Ma, Qing Li, Yuan He, Eric Zhao, Jiliang Tang, and Dawei Yin. Graph neural networks for social recommendation. In *The world wide web conference*, pages 417–426, 2019.

[34] Miroslav Fiedler. Algebraic connectivity of graphs. *Czechoslovak mathematical journal*, 23(2):298–305, 1973.

[35] James Flamino, Alessandro Galeazzi, Stuart Feldman, Michael W Macy, Brendan Cross, Zhenkun Zhou, Matteo Serafino, Alexandre Bovet, Hernán A Makse, and Boleslaw K Szymanski. Political polarization of news media and influencers on twitter in the 2016 and 2020 us presidential elections. *Nature Human Behaviour*, 7(6):904–916, 2023.

[36] Seth Flaxman, Sharad Goel, and Justin M Rao. Filter bubbles, echo chambers, and online news consumption. *Public opinion quarterly*, 80(S1):298–320, 2016.

[37] Jonathan B Freeman and Rick Dale. Assessing bimodality to detect the presence of a dual cognitive process. *Behavior research methods*, 45(1):83–97, 2013.

[38] Yichang Gao, Fengming Liu, and Lei Gao. Echo chamber effects on short video platforms. *Scientific reports*, 13(1):6282, 2023.

[39] Zhaolin Gao, Tianshu Shen, Zheda Mai, Mohamed Reda Bouadjenek, Isaac Waller, Ashton Anderson, Ron Bodkin, and Scott Sanner. Mitigating the filter bubble while maintaining relevance: Targeted diversification with vae-based recommender systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2524–2531, 2022.

[40] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Reducing controversy by connecting opposing views. In *Proceedings of the tenth ACM international conference on web search and data mining*, pages 81–90, 2017.

[41] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Quantifying controversy on social media. *ACM Transactions on Social Computing*, 1 (1):1–27, 2018.

[42] Yingqiang Ge, Shuya Zhao, Honglu Zhou, Changhua Pei, Fei Sun, Wenwu Ou, and Yongfeng Zhang. Understanding echo chambers in e-commerce recommender systems. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 2261–2270, 2020.

[43] Javad Ghaderi and Rayadurgam Srikant. Opinion dynamics in social networks with stubborn agents: Equilibrium and convergence rate. *Automatica*, 50(12):3209–3215, 2014.

[44] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.

[45] Sharad Goel, Ashton Anderson, Jake Hofman, and Duncan J Watts. The structural virality of online diffusion. *Management science*, 62(1):180–196, 2016.

[46] Pedro Guerra, Wagner Meira Jr, Claire Cardie, and Robert Kleinberg. A measure of polarization on social media networks based on community boundaries. In *Proceedings of the international AAAI conference on web and social media*, volume 7, pages 215–224, 2013.

[47] Andrew M Guess, Neil Malhotra, Jennifer Pan, Pablo Barberá, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, Deen Freelon, Matthew Gentzkow, et al. How do social media feed algorithms affect attitudes and behavior in an election campaign? *Science*, 381(6656):398–404, 2023.

[48] Shivam Gupta, Kirandeep Kaur, and Sanjay Jain. Eqbal-rs: Mitigating popularity bias in recommender systems. *Journal of Intelligent Information Systems*, 62:509–534, 2024.

[49] Kobi Hackenburg, Lujain Ibrahim, Ben M. Tappin, and Manos Tsakiris. Comparing the persuasiveness of role-playing large language models and human experts on polarized U.S. political issues. *AI & Society*, 2025. doi: 10.1007/s00146-025-02464-x.

[50] Shahrzad Haddadan, Cristina Menghini, Matteo Riondato, and Eli Upfal. RePBubLik: Reducing polarized bubble radius with link insertions. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 139–147, 2021.

[51] David Hartmann, Sonja Mei Wang, Lena Pohlmann, and Bettina Berendt. A systematic review of echo chamber research: comparative analysis of conceptualizations, operationalizations, and varying outcomes. *Journal of Computational Social Science*, 8(2):52, 2025.

[52] Yifeng He, Ziye Tang, and Hao Chen. Content fuzzing for escaping information cocoons on digital social media. *arXiv preprint arXiv:2604.05461*, 2026.

[53] Natali Helberger, Kari Karppinen, and Lucia D'acunto. Exposure diversity as a design principle for recommender systems. *Information, communication & society*, 21(2):191–207, 2018.

[54] McKenzie Himelein-Wachowiak, Salvatore Giorgi, Amanda Devoto, Muhammad Rahman, Lyle Ungar, H Andrew Schwartz, David H Epstein, Lorenzo Leggio, and Brenda Curtis. Bots and misinformation spread on social media: implications for covid-19. *Journal of medical Internet research*, 23 (5):e26933, 2021.

[55] Marilena Hohmann, Karel Devriendt, and Michele Coscia. Quantifying ideological polarization on a network using generalized euclidean distance. *Science Advances*, 9(9):eabq2044, 2023.

[56] Homa Hosseinmardi, Amir Ghasemian, Miguel Rivera-Lanas, Manoel Horta Ribeiro, Robert West, and Duncan J Watts. Causally estimating the effect of youtube's recommender system using counterfactual bots. *Proceedings of the national academy of sciences*, 121(8):e2313377121, 2024.

[57] Lei Hou, Xue Pan, Kecheng Liu, Zimo Yang, Jianguo Liu, and Tao Zhou. Information cocoons in online navigation. *iScience*, 26(1):105893, 2023.

[58] Yuxuan Hu. *Influence-based detection and mitigation of ideological isolation in online social networks*. PhD thesis, University of Tasmania, 2024.

[59] Yuxuan Hu, Shiqing Wu, Chenting Jiang, Weihua Li, Quan Bai, and Erin Roehrer. Ai facilitated isolations? the impact of recommendation-based influence diffusion in human society. In *IJCAI*, pages 5080–5086, 2022.

[60] Yuxuan Hu, Gemju Sherpa, Lan Zhang, Weihua Li, Quan Bai, Yijun Wang, and Xiaodan Wang. An llm-enhanced agent-based simulation tool for information propagation. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24, Jeju, Republic of Korea*, pages 8679–8682, 2024.

[61] David Scott Hunter and Tauhid Zaman. Optimizing opinions with stubborn agents. *Operations Research*, 70(4):2119–2137, 2022.

[62] Ferenc Huszár, Sofia Ira Ktena, Conor O'Brien, Luca Belli, Andrew Schlaikjer, and Moritz Hardt. Algorithmic amplification of politics on twitter. *Proceedings of the national academy of sciences*, 119(1):e2025334119, 2022.

[63] Luca Iandoli, Simonetta Primario, and Giuseppe Zollo. The impact of group polarization on the quality of online debate in social media: A systematic literature review. *Technological Forecasting and Social Change*, 170:120924, 2021.

[64] Paola Impicciché and Marco Viviani. Comparing echo chamber detection metrics: A cross-modeling and cross-platform analysis of Twitter and Reddit. *ACM Transactions on the Web*, 19 (2):1–32, 2025.

[65] Ruben Interian, Ruslán G. Marzo, Isela Mendoza, and Celso C Ribeiro. Network polarization, filter bubbles, and echo chambers: an annotated review of measures and reduction methods. *International Transactions in Operational Research*, 30 (6):3122–3158, 2023.

[66] Dong Jiang, Qionglin Dai, Haihong Li, and Junzhong Yang. Opinion dynamics based on social learning theory. *The European Physical Journal B*, 97(12):193, 2024.

[67] Julie Jiang, Xiang Ren, and Emilio Ferrara. Social media polarization and echo chambers in the context of covid-19: Case study. *JMIRx med*, 2(3):e29570, 2021.

[68] Antonio Jiménez-Martínez. A model of belief influence in large social networks. *Economic theory*, 59(1):21–59, 2015.

[69] Di Jin, Luzhi Wang, He Zhang, Yizhen Zheng, Weiping Ding, Feng Xia, and Shirui Pan. A survey on fairness-aware recommender systems. *Information Fusion*, 100:101906, 2023.

[70] Jonas Kaiser and Adrian Rauchfleisch. Birds of a feather get recommended together: Algorithmic homophily in youtube's channel recommendations in the united states and germany. *Social Media+ Society*, 6(4):2056305120969914, 2020.

[71] Yoshihisa Kashima, Andrew Perfors, Vanessa Ferdinand, and Elle Pattenden. Ideology, communication and polarization. *Philosophical Transactions of the Royal Society B*, 376(1822): 20200133, 2021.

[72] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '03)*, pages 137–146, Washington, DC, USA, 2003. ACM.

[73] Anastasiia Klimashevskaia, Dietmar Jannach, Mehdi Elahi, and Christoph Trattner. A survey on popularity bias in recommender systems. *User Modeling and User-Adapted Interaction*, 34(5):1777–1834, 2024. doi: 10.1007/s11257-024-09406-0.

[74] Silvia Knobloch-Westerwick, Benjamin K Johnson, and Axel Westerwick. Confirmation bias in online searches: Impacts of selective exposure before an election on political attitude strength and shifts. *Journal of Computer-Mediated Communication*, 20(2):171–187, 2015.

[75] Erik Knudsen. Modeling news recommender systems' conditional effects on selective exposure: evidence from two online experiments. *Journal of Communication*, 73(2):138–149, 2023.

[76] Denis Kotkov, Shuaiqiang Wang, and Jari Veijalainen. A survey of serendipity in recommender systems. *Knowledge-Based Systems*, 111:180–192, 2016.

[77] Nane Kratzke. How to find orchestrated trolls? a case study on identifying polarized twitter echo chambers. *Computers*, 12 (3):57, 2023.

[78] Thorsten Krause, Alina Deriyeva, Jan H Beinke, Gerrit Y Bartels, and Oliver Thomas. Mitigating exposure bias in recommender systems—a comparative analysis of discrete choice models. *ACM Transactions on Recommender Systems*, 3(2): 1–37, 2024.

[79] Alex Kulesza, Ben Taskar, et al. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3):123–286, 2012.

[80] Tomo Lazovich. Filter bubbles and affective polarization in user-personalized large language model outputs. In *Proceedings of the NeurIPS 2023 Workshop on "I Can't Believe It's Not Better: Failure Modes in the Age of Foundation Models"*, volume 239 of *Proceedings of Machine Learning Research*, 2023.

[81] Paddy Leerssen. The regulation of recommender systems under the DSA: A transition from default to multiple and dynamic controls? *Internet Policy Review / DSA Observatory*, 2024. `https://dsa-observatory.eu/2024/11/22/`.

[82] Jure Leskovec, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne VanBriesen, and Natalie Glance. Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 420–429, 2007.

[83] Gilat Levy and Ronny Razin. Echo chambers and their effects on economic and political outcomes. *Annual Review of Economics*, 11(1):303–328, 2019.

[84] Jianxin Li, Taotao Cai, Ke Deng, Xinjue Wang, Timos Sellis, and Feng Xia. Community-diversified influence maximization in social networks. *Information Systems*, 92:101522, 2020.

[85] Jinning Li, Huajie Shao, Dachun Sun, Ruijie Wang, Yuchen Yan, Jinyang Li, Shengzhong Liu, Hanghang Tong, and Tarek Abdelzaher. Unsupervised belief representation learning with information-theoretic variational graph auto-encoders. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1728–1738, 2022.

[86] Linda Li, Orsolya Vásárhelyi, and Balázs Vedres. Social bots spoil activist sentiment without eroding engagement. *Scientific Reports*, 14(1):27005, 2024.

[87] Quan Li, Zhenhui Peng, Haipeng Zeng, Qiaoan Chen, Lingling Yi, Ziming Wu, Xiaojuan Ma, and Tianjian Chen. Friend network as gatekeeper: a study of wechat users' consumption of friend-curated contents. In *Proceedings of the Eighth International Workshop of Chinese CHI*, pages 21–31, 2020.

[88] W Li, Q Bai, and M Zhang. Agent-based influence propagation in social networks. In *Proceedings of the 2016 IEEE International Conference on Agents (ICA 2016)*, pages 51–56. Matsue, 2016. ISBN 9781509039319. doi: 10.1109/ICA.2016.021.

[89] Weihua Li, Quan Bai, and Minjie Zhang. A multi-agent system for modelling preference-based complex influence diffusion in social networks. *The Computer Journal*, 62(3):430–447, 2019.

[90] Weihua Li, Quan Bai, Ling Liang, Yi Yang, Yuxuan Hu, and Minjie Zhang. Social influence minimization based on context-aware multiple influences diffusion model. *Knowledge-Based Systems*, 227:107233, 2021.

[91] Yanlin Li, Zicheng Cheng, and Homero Gil de Zúñiga. Tiktok's political landscape: Examining echo chambers and political expression dynamics. *new media & society*, page 14614448251339755, 2025.

[92] Yunqi Li, Hanxiong Chen, Shuyuan Xu, Yingqiang Ge, Juntao Tan, Shuchang Liu, and Yongfeng Zhang. Fairness in recommendation: Foundations, methods, and applications. *ACM Transactions on Intelligent Systems and Technology*, 14(5):1–48, 2023.

[93] Zhenyang Li, Yancheng Dong, Chen Gao, Yizhou Zhao, Dong Li, Jianye Hao, Kai Zhang, Yong Li, and Zhi Wang. Breaking filter bubble: A reinforcement learning framework of controllable recommender system. In *Proceedings of the ACM web conference 2023*, pages 4041–4049, 2023.

[94] Jie Liao, Min Yang, Wei Zhou, Hongyu Zhang, and Junhao Wen. Modeling item exposure and user satisfaction for debiased recommendation with causal inference. *Information Sciences*, 676:120834, 2024.

[95] Jian-Guo Liu, Kerui Shi, and Qiang Guo. Solving the accuracy-diversity dilemma via directed random walks. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 85(1):016118, 2012.

[96] Ping Liu, Karthik Shivaram, Aron Culotta, Matthew Shapiro, and Mustafa Bilgic. How does empowering users with greater system control affect news filter bubbles? In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 943–957, 2024.

[97] Philipp Lorenz-Spreen, Stephan Lewandowsky, Cass R Sunstein, and Ralph Hertwig. How behavioural sciences can promote truth, autonomy and democratic discourse online. *Nature human behaviour*, 4(11):1102–1109, 2020.

[98] Linyuan Lü and Tao Zhou. Link prediction in complex networks: A survey. *Physica A: statistical mechanics and its applications*, 390(6):1150–1170, 2011.

[99] Huishi Luo, Fuzhen Zhuang, Ruobing Xie, Hengshu Zhu, Deqing Wang, Zhulin An, and Yongjun Xu. A survey on causal inference for recommendation. *The Innovation*, 5(2), 2024.

[100] Robert Luzsa and Susanne Mayr. Links between users' online social network homogeneity, ambiguity tolerance, and estimated public support for own opinions. *Cyberpsychology,*

Behavior, and Social Networking, 22(5):325–329, 2019.

[101] Robert Luzsa and Susanne Mayr. False consensus in the echo chamber: Exposure to favorably biased social media news feeds leads to increased perception of public support for own opinions. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 15(1), 2021.

[102] Amin Mahmoudi, Dariusz Jemielniak, and Leon Ciechanowski. Echo chambers in online social networks: A systematic literature review. *IEEE Access*, 12:9594–9620, 2024.

[103] Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. Fairmatch: A graph-based approach for improving aggregate diversity in recommender systems. In *Proceedings of the 28th ACM conference on user modeling, adaptation and personalization*, pages 154–162, 2020.

[104] Alexander V Mantzaris. Uncovering nodes that spread information between communities in social networks. *EPJ Data Science*, 3(1):26, 2014.

[105] George Masterton and Erik J Olsson. Argumentation and belief updating in social networks: a Bayesian approach. In *Trends in Belief Revision and Argumentation Dynamics*. College Publications, 2013.

[106] Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems. In *Proceedings of the 27th acm international conference on information and knowledge management*, pages 2243–2251, 2018.

[107] Hannah Metzler and David Garcia. Social drivers and algorithmic mechanisms on digital media. *Perspectives on Psychological Science*, 19(5):735–748, 2024.

[108] Smitha Milli, Micah Carroll, Yike Wang, Sashrika Pandey, Sebastian Zhao, and Anca D Dragan. Engagement, user satisfaction, and the amplification of divisive content on social media. *PNAS nexus*, 4(3):pgaf062, 2025.

[109] Yong Min, Tingjun Jiang, Cheng Jin, Qu Li, and Xiaogang Jin. Endogenetic structure of filter bubble in social networks. *Royal Society open science*, 6(11):190868, 2019.

[110] Marco Minici, Federico Cinus, Corrado Monti, Francesco Bonchi, and Giuseppe Manco. Cascade-based echo chamber detection. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 1511–1520, 2022.

[111] Corrado Monti, Giuseppe Manco, Cigdem Aslay, and Francesco Bonchi. Learning ideological embeddings from information cascades. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 1325–1334, 2021.

[112] Fabio Motoki, Valdemar Pinho Neto, and Víctor Rodrigues. More human than human: Measuring ChatGPT political bias. *Public Choice*, 198(1):3–23, 2024.

[113] Pau Muñoz, Alejandro Bellogín, Raúl Barba-Rojas, and Fernando Díez. Quantifying polarization in online political discourse. *EPJ Data Science*, 13(1):39, 2024.

[114] Sreeja Nair and Adriana Iamnitchi. Cross-community affinity: A polarization measure for multi-community networks. *Online Social Networks and Media*, 43-44:100280, 2024.

[115] M. E. J. Newman. *Networks*. Oxford University Press, Oxford, 2nd edition, 2018. ISBN 9780198805090. doi: 10.1093/oso/9780198805090.001.0001.

[116] Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582, 2006.

[117] George Panagopoulos, Fragkiskos D Malliaros, and Michalis Vazirgianis. Influence maximization using influence and susceptibility embeddings. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 511–521, 2020.

[118] Lata Pangtey, Anukriti Bhatnagar, Shubhi Bansal, Shahid Shafi Dar, and Nagendra Kumar. Large language models meet stance detection: A survey of tasks, methods, applications, challenges and future directions. *arXiv preprint arXiv:2505.08464*, 2025.

[119] Javier Parapar and Filip Radlinski. Towards unified metrics for accuracy and diversity for recommender systems. In *Proceedings of the 15th ACM Conference on Recommender Systems*, pages 75–84, 2021.

[120] Giulio Pecile, Niccolò Di Marco, Matteo Cinelli, and Walter Quattrociocchi. Decoding political polarization in social media interactions. In *International Conference on Complex Networks and Their Applications*, pages 282–293. Springer, 2024.

[121] Kenny Peng, Manish Raghavan, Emma Pierson, Jon Kleinberg, and Nikhil Garg. Reconciling the accuracy-diversity trade-off in recommendations. In *Proceedings of the ACM Web Conference 2024*, pages 1318–1329, 2024.

[122] Yuan Peng, Yiyi Zhao, and Jiangping Hu. On the role of community structure in evolution of opinion formation: A new bounded confidence opinion dynamics. *Information Sciences*, 621:672–690, 2023.

[123] Roland Pfister, Katharina A Schwarz, Markus Janczyk, Rick Dale, and Jonathan B Freeman. Good things peak in pairs: a note on the bimodality coefficient. *Frontiers in psychology*, 4: 700, 2013.

[124] Jinghua Piao, Jiazhen Liu, Fang Zhang, Jun Su, and Yong Li. Human–AI adaptive dynamics drives the emergence of information cocoons. *Nature Machine Intelligence*, 5(11):1214–1224, 2023.

[125] Jiangtao Qiu, Zhangxi Lin, and Qinghong Shuai. Investigating the opinions distribution in the controversy on social media. *Information Sciences*, 489:289–302, 2019.

[126] Jiezhong Qiu, Jian Tang, Hao Ma, Yuxiao Dong, Kuansan Wang, and Jie Tang. Deepinf: Social influence prediction with deep learning. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2110–2119, 2018.

[127] Emerson Rodrigues da Cunha Palmieri. Social media, echo chambers and contingency: a system theoretical approach about communication in the digital space. *Kybernetes*, 53(8): 2593–2604, 2024.

[128] David Rozado. The political preferences of LLMs. *PLoS ONE*, 19(7):e0306621, 2024.

[129] Jérôme Rutinowski, Sven Franke, Jan Endendyk, Ina Dormuth, Moritz Roidl, and Markus Pauly. The self-perception and political biases of ChatGPT. *Human Behavior and Emerging Technologies*, 2024:7115633, 2024.

[130] Rodrygo LT Santos, Craig Macdonald, and Iadh Ounis. Exploiting query reformulations for web search result diversification. In *Proceedings of the 19th international conference on World wide web*, pages 881–890, 2010.

[131] Egemen Sert, Yaneer Bar-Yam, and Alfredo J Morales. Segregation dynamics with reinforcement learning and agent based modeling. *Scientific reports*, 10(1):11771, 2020.

[132] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. The spread of low-credibility content by social bots. *Nature communications*, 9(1):4787, 2018.

[133] Nuan Song, Wei Sheng, Yanhao Sun, Tianwei Lin, Zeyu Wang, Zhanxue Xu, Fei Yang, Yatao Zhang, and Dong Li. Online dynamic influence maximization based on deep reinforcement learning. *Neurocomputing*, 618:129117, 2025.

[134] Xiaolei Song, Siliang Guo, and Yichang Gao. Personality traits and their influence on echo chamber formation in social media: a comparative study of Twitter and Weibo. *Frontiers in Psychology*, 15:1323117, 2024.

[135] Nicholas Sukiennik, Haoyu Wang, Zailin Zeng, Chen Gao, and Yong Li. Simulating filter bubble on short-video recommender system with large language model agents. *arXiv preprint arXiv:2504.08742*, 2025.

[136] Haoran Tang, Shiqing Wu, Zhihong Cui, Yicong Li, Guandong Xu, and Qing Li. Model-agnostic dual-side online fairness learning for dynamic recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 2025.

[137] Youze Tang, Yanchen Shi, and Xiaokui Xiao. Influence maximization in near-linear time: A martingale approach. In *Proceedings of the 2015 ACM SIGMOD international conference on management of data*, pages 1539–1554, 2015.

[138] Ludovic Terrén and Rosa Borge-Bravo. Echo chambers on social media: A systematic review of the literature. *Review of Communication Research*, 9, 2021.

[139] Petter Törnberg. How digital media drive affective polarization through partisan sorting. *Proceedings of the National Academy of Sciences*, 119(42):e2207159119, 2022.

[140] Bao Tran Truong, Xiaodan Lou, Alessandro Flammini, and Filippo Menczer. Quantifying the vulnerabilities of the online public square to adversarial manipulation tactics. *PNAS nexus*, 3(7):pgae258, 2024.

[141] Alan Tsang, Bryan Wilder, Eric Rice, Milind Tambe, and Yair Zick. Group-fairness in influence maximization. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 5997–6005, 2019.

[142] Carlo M Valensise, Matteo Cinelli, and Walter Quattrociocchi. The drivers of online polarization: Fitting models to data. *Information Sciences*, 642:119152, 2023.

[143] Giacomo Villa, Gabriella Pasi, and Marco Viviani. Echo chamber detection and analysis: A topology-and content-based approach in the covid-19 scenario. *Social Network Analysis and Mining*, 11(1):78, 2021.

[144] Karan Vombatkere, Sepehr Mousavi, Savvas Zannettou, Franziska Roesner, and Krishna P Gummadi. Tiktok and the art of personalization: investigating exploration and exploitation on social media feeds. In *Proceedings of the ACM Web Conference 2024*, pages 3789–3797, 2024.

[145] Lukas Von Flüe and Sonja Vogt. Integrating social learning and network formation for social tipping towards a sustainable future. *Current Opinion in Psychology*, 60:101915, 2024.

[146] Difan Wang and Yi Qian. Echo chamber effect in rumor rebuttal discussions about COVID-19 in China: Social media content and network analysis study. *Journal of Medical Internet Research*, 23(3):e27009, 2021. doi: 10.2196/27009.

[147] Guan Wang, Weihua Li, Quan Bai, and Edmund MK Lai. Maximizing social influence with minimum information alteration. *IEEE Transactions on Emerging Topics in Computing*, 12(2):419–431, 2024.

[148] Mengyan Wang, Yuxuan Hu, Shiqing Wu, Weihua Li, Quan Bai, and Verica Rupar. Balancing information perception with yin-yang: Agent-based adaptive information neutrality model for recommendation systems. *IEEE Transactions on Computational Social Systems*, 2025.

[149] Mengyan Wang, Yuxuan Hu, Shiqing Wu, Weihua Li, Quan Bai, Zihan Yuan, and Chenting Jiang. Nudging toward responsible recommendations: A graph-based approach to mitigate belief filter bubbles. *IEEE Transactions on Artificial Intelligence*, 6(2):378–392, 2025.

[150] Xiaodan Wang, Yanbin Liu, Weihua Li, and Quan Bai. Tunnel vision in online discourse: Formalization and entropy-based quantification with llm-simulated agents. In *Pacific Rim International Conference on Artificial Intelligence*, pages 297–312. Springer, 2025.

[151] Xinyu Wang, Jiayi Li, Eesha Srivatsavaya, and Sarah Rajtmajer. Evidence of inter-state coordination amongst state-backed information operations. *Scientific reports*, 13(1):7716, 2023.

[152] Yifan Wang, Weizhi Ma, Min Zhang, Yiqun Liu, and Shaoping Ma. A survey on the fairness of recommender systems. *ACM Transactions on Information Systems*, 41(3):1–43, 2023.

[153] Zhenlei Wang, Jingsen Zhang, Hongteng Xu, Xu Chen, Yongfeng Zhang, Wayne Xin Zhao, and Ji-Rong Wen. Counterfactual data-augmented sequential recommendation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 347–356, 2021.

[154] Tianxin Wei, Fuli Feng, Jiawei Chen, Ziwei Wu, Jinfeng Yi, and Xiangnan He. Model-agnostic counterfactual reasoning for eliminating popularity bias in recommender system. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 1791–1800, 2021.

[155] Shiqing Wu, Quan Bai, and Byeong Ho Kang. Adaptive incentive allocation for influence-aware proactive recommendation. In *Pacific Rim International Conference on Artificial Intelligence*, pages 649–661. Springer, 2019.

[156] Shiqing Wu, Weihua Li, and Quan Bai. Gac: A deep reinforcement learning model toward user incentivization in unknown social networks. *Knowledge-Based Systems*, 259:110060, 2023.

[157] Shiqing Wu, Weihua Li, Hao Shen, and Quan Bai. Identifying influential users in unknown social networks for adaptive incentive allocation under budget restriction. *Information Sciences*, 624:128–146, 2023.

[158] Shiqing Wu, Guandong Xu, and Xianzhi Wang. Soac: Supervised off-policy actor-critic for recommender systems. In *2023 IEEE International Conference on Data Mining (ICDM)*, pages 1421–1426. IEEE, 2023.

[159] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S Yu. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1):4–24, 2021.

[160] Jian Xu. Opening the 'black box'of algorithms: regulation of algorithms in china. *Communication Research and Practice*, 10(3):288–296, 2024.

[161] Yuanbo Xu, Yongjian Yang, En Wang, Jiayu Han, Fuzhen Zhuang, Zhiwen Yu, and Hui Xiong. Neural serendipity recommendation: Exploring the balance between accuracy and novelty with sparse explicit feedback. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 14(4):1–25, 2020.

[162] Harry Yaojun Yan, Kai-Cheng Yang, James Shanahan, and Filippo Menczer. Exposure to social bots amplifies perceptual biases and regulation propensity. *Scientific Reports*, 13(1): 20707, 2023.

[163] Haifeng Yang, Ran Zhang, Jianghui Cai, Jie Wang, Yupeng Wang, Yating Li, Yaling Xun, and Xujun Zhao. Content suppression mechanisms-based recommendation systems. *Expert Systems with Applications*, page 128928, 2025.

[164] Ercan Yildiz, Asuman Ozdaglar, Daron Acemoglu, Amin Saberi, and Anna Scaglione. Binary opinion dynamics with stubborn agents. *ACM Transactions on Economics and Computation*, 1(4):1–30, 2013.

[165] Lan Zhang, Yuxuan Hu, Weihua Li, Quan Bai, and Parma Nand. Llm-aidsim: Llm-enhanced agent-based influence diffusion simulation in social networks. *Systems*, 13(1):29, 2025.

[166] Wen Zhang, Song Wang, Guangle Han, Ye Yang, and Qing Wang. Hiprank: Ranking nodes by influence propagation based on authority and hub. In Quan Bai, Fenghui Ren, Minjie Zhang, Takayuki Ito, and Xijin Tang, editors, *Smart Modeling and Simulation for Complex Systems*, volume 564, pages 1–14. Springer, Tokyo, 2015.

[167] Yilin Zhang and Karl Rohe. Understanding regularized spectral clustering via graph conductance. *Advances in Neural Information Processing Systems*, 31, 2018.

[168] Yu Zhang. Diversifying seeds and audience in social influence maximization. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 89–94, 2019.

[169] Yuanxing Zhang, Zhuqi Li, Chengliang Gao, Kaigui Bian, Lingyang Song, Shaoling Dong, and Xiaoming Li. Mobile social big data: Wechat moments dataset, network applications, and opportunities. *IEEE network*, 32(3):146–153, 2018.

[170] Yuying Zhao, Yu Wang, Yunchao Liu, Xueqi Cheng, Charu C. Aggarwal, and Tyler Derr. Fairness and diversity in recommender systems: A survey. *ACM Transactions on Intelligent Systems and Technology*, 16(1):1–28, 2024.

[171] Ziying Zhao, Weihua Li, Jing Ma, Jianhua Jiang, and Quan Bai. Fair influence maximization in social networks: a group-fairness-aware multi-objective grey wolf optimizer. In *2024 IEEE International Conference on Agents (ICA)*, pages 88–93. IEEE, 2024.

Declaration of Interest Statement

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The author is an Editorial Board Member/Editor-in-Chief/Associate Editor/Guest Editor for this journal and was not involved in the editorial review or the decision to publish this article.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Author Biography

Xiaodan Wang is an Associate Professor at Yanbian University, China, and a PhD candidate at Auckland University of Technology, New Zealand. Her research focuses on artificial intelligence, with particular interests in natural language processing, sentiment analysis, information propagation, and LLM-driven intelligent systems. She has published over 20 peer-reviewed papers in leading journals and international conferences. Her current research emphasizes LLM-enhanced simulation and modeling of complex social systems.

Yanbin Liu received the B.E. and M.S. degrees from Tianjin University, China, in 2013 and 2015, respectively. He received the Ph.D. degree from the University of Technology Sydney, Australia, in 2021. He is currently a senior lecturer at the Auckland University of Technology. His research interests lie in machine learning and deep learning for computer vision problems, especially learning with limited labeled data.

Shiqing Wu is an Assistant Professor in the Faculty of Data Science at City University of Macau, Macau SAR, China. Before that, he was a Postdoctoral Research Fellow in the School of Computer Science at the University of Technology Sydney, Australia, from 2022 to 2024. He received his PhD in Computer Science from the University of Tasmania, Australia, in 2022, and a joint B.Sc. in Computer Science from Auckland University of Technology, New Zealand, and China Jiliang University, China, in 2016. His research interests include agent-based modeling, social influence analysis, recommender systems, and reinforcement learning. He has published over 35 research papers in prestigious venues, such as SIGIR, ICDM, IJCAI, AAMAS, ACM MM, TKDE, and TOIS. Shiqing is an Editorial Board Member of Information Processing & Management.

Ziying Zhao received the M.S. degree in Artificial Intelligence from Jilin University of Finance and Economics, Changchun, China, in 2023. She is currently pursuing a Ph.D. degree in Computer Science at the Auckland University of Technology (AUT), Auckland, New Zealand. Her current research interests include artificial intelligence, computational intelligence, social intelligence, and complex systems.

Yuxuan Hu received the bachelor's degree in electronic information

engineering from Hunan Normal University, China, in 2016, the M.E. degree in electrical and electronic engineering from the University of Auckland, New Zealand, in 2017, and the Ph.D. degree in computer science from the University of Tasmania, Australia, in 2023. She is currently a Data Engineer with the Integrated Marine Observing System at the University of Tasmania. Her research interests include agent-based modeling, social network analysis, and recommendation systems.

Weihua Li received the Ph.D. degree in Computer Science from Auckland University of Technology (AUT), Auckland, New Zealand, in 2018, and the M.Tech. degree in Knowledge Engineering from the National University of Singapore, Singapore, in 2014. He is currently a Senior Lecturer with the Department of Data Science and Artificial Intelligence. Weihua is an active researcher in the field of artificial intelligence, with a particular focus on social computing, generative AI, agent-based modeling and simulation of complex systems, and agentic AI. He has authored over 80 publications and has been actively involved in organizing several international AI conferences.

Quan Bai received his PhD and MSc from the University of Wollongong, Australia. He is currently an Associate Professor at the University of Tasmania, where he leads a vibrant AI research group. A recognized expert in Agentic AI and machine learning, he is at the forefront of advancing next-generation intelligent systems. His research focuses on applying advanced AI methodologies to model and manage complex systems with highly interdependent components. He has published more than 180 peer-reviewed papers in leading journals and conferences and has attracted over AUD 4 million in research funding.