Published: 12 September 2025
Online AM: 01 August 2025
Accepted: 24 July 2025
Revised: 05 April 2025
Received: 20 December 2024

RESEARCH-ARTICLE

Breaking the Loop: Causal Learning to Mitigate Echo Chambers in Social Networks

DIANER YU, University of Technology Sydney, Sydney, NSW, Australia

QIAN LI, Curtin University, Perth, WA, Australia

HUAN HUO, University of Technology Sydney, Sydney, NSW, Australia

GUANGDONG XU, The Education University of Hong Kong, Hong Kong, Hong Kong

Breaking the Loop: Causal Learning to Mitigate Echo Chambers in Social Networks

DIANER YU, University of Technology Sydney, Sydney, Australia
QIAN LI, School of Electrical Engineering Computing and Mathematical Sciences, Curtin University, Perth, Australia
HUAN HUO, University of Technology Sydney, Sydney, Australia
GUANDONG XU, Centre for Learning, Teaching & Technology, The Education University of Hong Kong, Hong Kong, Hong Kong

In social networks, echo chambers form when users primarily encounter information that reinforces their existing views with limited exposure to different perspectives. This self-reinforcing isolation worsens societal issues such as division and declining public discourse. Traditional approaches attempt to mitigate echo chambers by analyzing observable interaction patterns to identify their formative mechanisms. However, they overlook unobserved implicit factors, called hidden confounders in causal inference, that significantly influence content exposure and user behaviors despite not being directly captured in the data. To address this, we propose **Causal Echo Diffusion Attenuator (CEDA)**, a novel framework that integrates causal learning with sequential recommendations to detect and adjust for hidden confounders in social networks. Generally, CEDA comprises four key components: (1) *User Dual Modelling* builds comprehensive user embeddings by combining users' attributes and structural information to fully capture behavior patterns. (2) *Causal Transformer* then estimates residual embeddings that account for hidden confounders, incorporating them into the Transformer as causal adjustments for unbiased user embeddings. (3) *Social Diffusion Predictor* uses unbiased user embeddings to jointly optimize diffusion prediction accuracy and information diversity. (4) *Targeted Interventions* strategically reshapes information flows to disrupt echo chambers based on the generated prediction and diversity insights. Extensive experiments demonstrate CEDA's superior performance in both predicting information diffusion patterns and mitigating echo chambers.

CCS Concepts: • **Theory of computation → Social networks**; • **Computing methodologies → Causal reasoning and diagnostics**;

Additional Key Words and Phrases: Social Network, Causal Inference, Sequential Recommender System

This work is partially supported by the Australian Research Council (ARC) under Grant number DP220103717, and the National Science Foundation of China under Grant number 62072257.
Authors' Contact Information: Dianer Yu, University of Technology Sydney, Sydney, Australia; e-mail: Dianer.Yu-1@student.uts.edu.au; Qian Li (corresponding author), School of Electrical Engineering Computing and Mathematical Sciences, Curtin University, Perth, Australia; e-mail: qli@curtin.edu.au; Huan Huo, University of Technology Sydney, Sydney, Australia; e-mail: Huan.Huo@uts.edu.au; Guandong Xu (corresponding author), Centre for Learning, Teaching & Technology, The Education University of Hong Kong, Hong Kong, Hong Kong; e-mail: gdxu@eduhk.hk.

ACM 1558-2868/2025/9-ART163
https://doi.org/10.1145/3757738

ACM Reference format:
Dianer Yu, Qian Li, Huan Huo, and Guandong Xu. 2025. Breaking the Loop: Causal Learning to Mitigate Echo Chambers in Social Networks. *ACM Trans. Inf. Syst.* 43, 6, Article 163 (September 2025), 27 pages.
https://doi.org/10.1145/3757738

1 Introduction

Social networks have become integral platforms for information sharing and public discourse, fundamentally shaping how people interact with information in the digital age [39, 58, 81]. While the social network platforms facilitate unprecedented connectivity and information exchange, they face a critical challenge known as echo chambers: isolated information spaces where users are predominantly exposed to content that reinforces their existing views while limiting their exposure to diverse perspectives [10, 22, 62]. Such self-reinforcing isolation mechanism creates a concerning feedback loop, where algorithmic content filtering and users' natural tendencies to connect with like-minded individuals continuously strengthen the information bubbles. The downstream effects of echo chambers raise serious societal concerns, since reduced exposure to alternative viewpoints can lead to increased polarization and degradation of public discourse quality [11, 18]. Despite various technical and social interventions proposed by prior works, effectively mitigating echo chambers while maintaining user engagement remains a fundamental challenge. This motivates our research question: *How can we effectively reduce echo chambers in social networks, fostering information diversity while preserving user engagement?*

Previous approaches to mitigating echo chambers generally fall into two directions: (1) adjusting content-ranking algorithms to promote exposure to diverse viewpoints across all users [32, 47, 49, 60] or (2) intervening in network structures by connecting isolated communities to encourage cross-group interactions [4, 11]. However, these methods focus solely on observable user interactions and fail to account for implicit external influences—known as hidden confounders in causal inference—that may simultaneously influence content exposure and user behavior [16, 19, 44]. These hidden confounders commonly arise in social networks and can introduce spurious correlations that distort the true causal structure of user engagement despite not being directly captured in the data [4, 20, 26]. One notable example occurs in political discussions on social media. When a major election debate takes place offline, it simultaneously influences thousands of users, prompting them to engage with related content. This unobserved external event may increase exposure to political content (e.g., Treatment) and results in widespread participation in political discussions (e.g., Outcome), even among users who have no history of interacting with each other. A traditional recommender model that overlooks such unobserved influences (e.g., election debate) may mistakenly infer preference alignment, reinforcing echo chambers by continuously recommending similar politically homogeneous content [21]. Similarly, in e-commerce platforms, hidden confounders emerge when external supply chain disruptions cause users to shift their attention from an unavailable product to alternatives. Without modeling this external constraint, the system may incorrectly attribute the change to evolving user preferences, leading to suboptimal recommendations. To further clarify, as shown in Figure 1, the unobserved World Cup event (e.g., acts as hidden confounder) affects users across different communities, leading them to simultaneously engage in soccer-related discussions (e.g., highlighted in light orange). Although these users lack prior direct connections, their synchronized behavioral shift is driven by the same unobserved external factor. Thus, addressing hidden confounders becomes critical for effectively mitigating echo chambers in social networks, as these unobserved factors may systematically

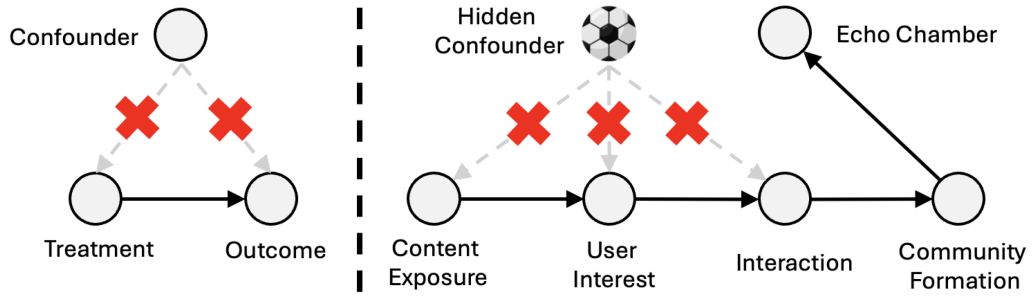

Fig. 1. Our designed causal graph illustrating how hidden confounders contribute to echo chamber formation in social networks. Black arrows are observable causal relationships, while gray dashed arrows denote the influence of hidden confounders on multiple variables. *Left*: An abstract causal diagram where a hidden confounder simultaneously affects both treatment (e.g., content exposure) and outcome (e.g., user interaction), leading to spurious correlations. *Right*: A more detailed causal pathway showing how hidden confounders influence user interest and interaction patterns, ultimately shaping community structures and reinforcing echo chambers.

generate spurious correlations in user behaviors and inadvertently reinforce information loops across social networks.

To address the fundamental challenge of mitigating echo chambers in social networks, we propose integrating causal inference principles into social network analysis. Causal inference provides a principled framework for uncovering true cause-effect relations among variables, helping us understand how hidden confounders influence the formations of echo chambers [71, 77, 78]. Specifically, hidden confounders always lead to systematic patterns in social networks where multiple users simultaneously deviate from their typical behavior in similar or identical ways [1, 17]. For example, during the World Cup event, user behavior in different communities commonly shifted towards soccer-related discussions to form echo chambers, as shown in Figure 2. Such synchronous behavioral changes cannot be explained by observable user features or social relationships, indicating the presence of unobserved implicit hidden confounders driving user behaviors. In causal inference, the discrepancy between predicted and observed outcomes can serve as an indicator of hidden confounder effects [38, 50]. Based on this causal principle, we propose a novel residual embedding method that quantifies the behavioral discrepancies by comparing behavior predictions based on observable features with actual observed user behaviors. By incorporating such hidden confounder effects into the model, we can better identify the formation mechanisms of echo chambers, thereby designing more effective interventions to enhance the diversity and maintain user engagement.

Towards this end, we propose a novel causal-based approach called **Causal Echo Diffusion Attenuator (CEDA)** to address the fundamental challenge of hidden confounders in mitigating echo chambers in social networks. Our framework integrates rigorous causal inference principles into the Transformer architecture within sequential recommendation frameworks, enabling systematic identification and adjustment for hidden confounders that influence information diffusion patterns. Specifically, CEDA has four key components: (1) *User Dual Modelling* constructs comprehensive user embeddings by combining users' attributes and their network positions, providing a holistic view of each user's behavior patterns in the social network. (2) *Causal Transformer* then estimates the residual embeddings to quantify hidden confounder effects, incorporating them as causal adjustments within the Transformer's attention mechanism to refine the comprehensive user embeddings into unbiased user embeddings. (3) *Social Diffusion Predictor* utilizes unbiased user

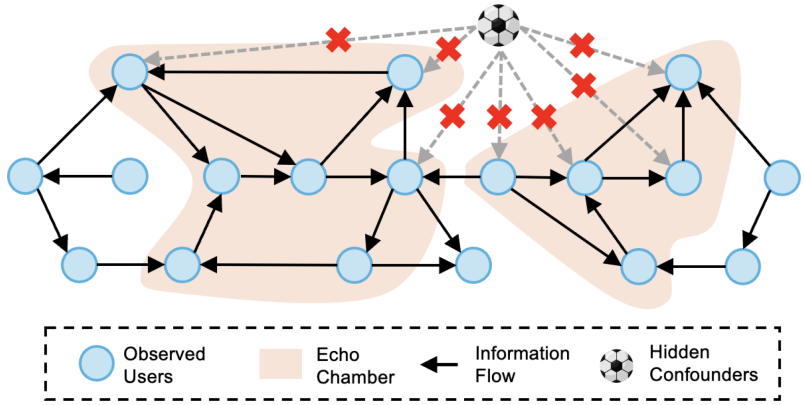

Fig. 2. A toy example illustrating how hidden confounders influence information diffusion in social networks, contributing to echo chamber formation (highlighted in light orange). The World Cup, an unobserved external event, acts as the hidden confounder by simultaneously increasing users' exposure to soccer-related content (Treatment) and prompting widespread engagement in related discussions (Outcome) across different communities. Although these users previously lacked direct social connections, their synchronized behavior emerges from the shared external influence rather than intrinsic similarity. Black edges represent observable interactions, while gray dashed lines indicate behavior shaped by the hidden confounder that cannot be explained by network structure alone. This example corresponds to the causal graph in Figure 1, where the hidden confounder jointly affects treatment and outcome, leading to spurious correlations that reinforce echo chambers.

embedding to jointly optimize prediction accuracy and diversity metrics, leading to accurate prediction of information diffusion patterns while promoting content diversity in social networks. (4) *Targeted Interventions* leverages the generated prediction and diversity insights to strategically reshape information flow patterns, effectively disrupting echo chamber formation while maintaining user engagement. Through extensive experiments on three real-world datasets, CEDA demonstrates superior performance in both predicting information diffusion and mitigating echo chambers compared to state-of-the-art methods.

Our main contributions of this work are summarized below:

- To the best of our knowledge, we are the first to apply causal learning to address the echo chambers in social networks, providing an innovative perspective on enhancing information diversity and diffusion accuracy in social networks.
- We propose a novel framework called *CEDA*, which integrates causal inference with sequential recommendation techniques to mitigate echo chambers by addressing hidden confounders often overlooked but critically influencing information flow patterns.
- We develop a dual-perspective user modeling approach that combines user attributes and structural positions within diffusion sequences, offering a comprehensive representation of user behavior patterns.
- We introduce a residual embedding-based causal adjustment mechanism within the Transformer, which quantifies the effects of hidden confounders through behavioral discrepancies, enhancing the accuracy of information diffusion prediction and enabling targeted interventions to disrupt echo chambers.
- We conduct extensive experiments on three real-world datasets to demonstrate CEDA's superior performance in mitigating echo chambers in social networks.

2 Related Work

2.1 Echo Chambers

Echo chambers in social networks refer to the self-reinforcing environments where users primarily encounter information that reinforces with their existing views [10, 22, 62]. This phenomenon emerge through several key interconnected mechanisms that create a self-perpetuating cycle: users' inherent tendency to connect with like-minded individuals, ongoing peer influences from established social connections, and algorithmic filtering systems that progressively amplify similarity within user groups [30, 81]. The proliferation of echo chambers and their detrimental effects on public discourse, such as increased polarization and reduced exposure to diverse perspectives, has motivated various mitigation strategies in recent research efforts. For example, OCR [66] implements a quadratic optimization framework to promote content diversity while maintaining user engagement through carefully calibrated content selection interventions. Taking a causal perspective, DCCF [74] employs counterfactual reasoning to systematically balance content exposure across different user groups to mitigate echo chamber effects. CECD [46] advances the field by developing a probabilistic generative model that detects echo chambers through analysis of information cascade patterns across networks. Approaching the problem through network embeddings, ECS [2] proposes an embedding distance-based methodology to quantify and guide interventions by measuring both cohesion within communities and separation between them. Building on graph neural networks, FRECH [63] utilizes a GCN to learn representations of users and echo chambers from content patterns and community interactions, serving as an intervention mechanism to recommend connections beyond users' echo chambers. Despite their contributions, these methods focus on observable data and often overlook unobserved hidden confounders, that can critically influence user behavior and content diffusion patterns, leading to biased assessments and suboptimal intervention strategies. In this work, we integrate causal inference into social network analysis, aiming to uncover hidden confounders and design more effective interventions to break echo chambers and foster information diversity.

2.2 Causal Inference

Causal inference offers a sophisticated statistical framework for understanding complex cause-and-effect relationships between variables, offering crucial advantages over traditional correlational analysis in addressing echo chambers [74, 77]. Through principled causal techniques like causal intervention and instrumental variables, causal inference enables identification and adjustment of confounding effects. This helps reveal the true causal mechanisms between variables and provides a rigorous basis for understanding complex social phenomena [50]. The effectiveness of causal approaches in handling confounders has been conclusively demonstrated through successful applications across multiple domains [14, 26]. In finance domain, Atanasov and Black employ shock-based causal models to address market conditions as confounders, which simultaneously affect both corporate governance choices and firm value [3]. In healthcare domain, Prosperi et al. estimate treatment effects from clinical data by adjusting for hidden confounders that influence both treatment decisions and patient outcomes [51]. In cybersecurity domain, Baluta et al. employ directed acyclic graphs with do-calculus to address model complexity as a confounder, which simultaneously affects both model generalization and membership inference attack accuracy [5]. These examples demonstrate the transformative potential of causal inference to address unobserved factors that significantly skew analyses and outcomes. Inspired by these successes, we extend the application of causal inference to social networks, addressing the challenge of hidden confounders that bias user behavior and information diffusion patterns. By integrating causal inference techniques into echo chamber mitigation, we aim to unravel the true mechanisms of their formation

and then design more precise and effective intervention strategies. This approach promises to bridge the gap between observable and unobservable factors, offering a comprehensive solution to enhance diversity and disrupt self-reinforcing cycles (e.g., echo chambers) in the social network.

3 Problem Formulation and Preliminary

To formulate the echo chamber problem, let $\mathcal{U}$ be the set of users in social networks, where each user $u_j \in \mathcal{U}$ is associated with a set of attributes a_j. Information propagates through the social network via sequential user interactions $S = \{u_1, \ldots, u_j, \ldots, u_J\}$, where each transition $\{u_{(j-1)}, u_j\}$ represents content propagated from user $u_{(j-1)}$ to user u_j. As mentioned, these interaction sequences may be impacted by hidden confounders that not directly observable in the dataset, but affect both content exposure and user interaction patterns. Our designed causal graph in Figure 1 shows how hidden confounders create spurious correlations between observed variables, leading to biased estimates of the true causal factor behind echo chamber formation. Therefore, our goal is to address hidden confounders in social networks to accurately predict information diffusion patterns and reveal the true causal mechanisms underlying echo chambers. Thereby, targeted interventions can be implemented to effectively promote the information diversity while maintaining the user engagement. Table 1 shows the key notations used in this work.

Specifically, we propose a novel causal-based approach that integrates causal inference into the social network analysis. Based on users' attributes a_j and their sequential positions in information diffusion paths S, we first learn a comprehensive user embedding $\boldsymbol{e}_j$ for each user u_j to capture user behavior patterns. Following causal inference principles where discrepancies between predicted and observed outcomes may indicate hidden confounder effects [50], we estimate residual embeddings $\boldsymbol{r}_j$ by computing the difference between predicted content sharing activity based on $\boldsymbol{e}_j$ and actual sharing records O_j. This discrepancy reveals hidden confounders' influence on sharing behaviors that is not directly reflected in user attributes and structural positions. The residual embeddings are then incorporated into the Transformer's attention mechanism as causal adjustment to adjust attention weights for hidden confounders, thereby refining the comprehensive user embeddings $\boldsymbol{e}_j$ into unbiased user embeddings $\boldsymbol{u}_j$. Then, we optimize our model through a joint objective to capture different aspects of information spread and diversity:

$$\mathcal{L} = \lambda_1 \mathcal{L}_r + \lambda_2 \mathcal{L}_m(\boldsymbol{u}_j, \boldsymbol{u}_q) + \lambda_3(1 - \mathcal{L}_d(\boldsymbol{u}_j)) + \lambda_4(1 - \mathcal{L}_c(\boldsymbol{u}_j)), \tag{1}$$

where λ_1, λ_2, λ_3 and λ_4 are weighting parameters. $\boldsymbol{u}_j$ and $\boldsymbol{u}_q$ are the unbiased user embedding for user u_j and u_j. $\mathcal{L}_m(\boldsymbol{u}_j, \boldsymbol{u}_q)$ measures the **Mean Absolute Error (MAE)** [28] of diffusion probability predictions between user pairs to optimize information diffusion accuracy. $\mathcal{L}_d(\boldsymbol{u}_j)$ measures the **Intra-List Diversity (ILD)** [29, 31] of user dissimilarity within diffusion sequences for maximizing interaction diversity. $\mathcal{L}_c(\boldsymbol{u}_j)$ measures the **Category Coverage (CC)** [52, 68] to quantify the diversity of content exposure in information diffusion across the network. By optimizing the $\mathcal{L}$, our model learns to make accurate predictions while promoting information diversity, enabling targeted interventions that effectively reshape information flows to break echo chambers.

4 Methodology

Figure 3 shows the overall framework of our proposed CEDA, designed to integrate causal inference for effectively addressing echo chambers in social networks. CEDA comprises four main components: (1) *User Dual Modelling* constructs comprehensive user embeddings by combining users' attributes and network positions, providing a holistic view of each user's behavior in the social network. (2) *Causal Transformer* then leverages comprehensive user embeddings to estimate residual embeddings that account for hidden confounders, incorporating them into the Transformer's attention weights as causal adjustments to produce unbiased user embeddings. (3) *Social Diffusion*

Table 1. Key Notations and Descriptions

Notations	Description
Sets and Basic Elements	
$\mathcal{U}$	Set of all users in social networks
u_j, u_q	The jth, qth user in sequential interaction path
S	Set of sequential user interactions
E	Set of all user pairs in the network
C	Set of all diffusion cascades
$\lvert I_{jq}\rvert$	Number of information transmissions from u_j to u_q
$\lvert I_j\rvert$	Total number of information transmissions from user u_j
$\lvert c\rvert$	Number of users in cascade c
Vectors and Matrices	
$\boldsymbol{a}_j \in \mathbb{R}^d$	Attribute vector of user u_j
$\boldsymbol{e}_j \in \mathbb{R}^d$	Comprehensive user embedding for user u_j
$\boldsymbol{u}_j \in \mathbb{R}^d$	Unbiased user embedding for user u_j
$\boldsymbol{r}_j \in \mathbb{R}^d$	Residual embedding for user u_j
$\boldsymbol{p}_j \in \mathbb{R}^d$	Positional encoding vector for user u_j
$\mathbf{E} \in \mathbb{R}^{J\times d}$	User embedding matrix $[\boldsymbol{e}_1, \cdots, \boldsymbol{e}_J]^\top$
$\mathbf{P} \in \mathbb{R}^{J\times d}$	Positional encoding matrix
$\mathbf{O} \in \mathbb{R}^{J\times d}$	Output matrix from attention heads
$\mathbf{1}_J \in \mathbb{R}^J$	Vector of ones for broadcasting
Functions	
$z(\boldsymbol{e}_j)$	Prediction function based on user embedding
$g(O_j)$	Mapping function for observed outcomes
Categories$(\boldsymbol{u}_j)$	Function mapping $\boldsymbol{u}_j$ to category probabilities
D_{jq}	True diffusion probability from user u_j to user u_q
$\widehat{D}_{jq}$	Predicted diffusion probability from user u_j to user u_q
$\cos(\boldsymbol{u}_j, \boldsymbol{u}_q)$	Cosine similarity between user embeddings
$\lvert\cdot\rvert_F$	Frobenius norm
Losses	
$\mathcal{L}_r$	Residual loss for capturing hidden confounders
$\mathcal{L}_m(\boldsymbol{u}_j, \boldsymbol{u}_q)$	Mean Absolute Error loss for diffusion prediction
$\mathcal{L}_d(\boldsymbol{u}_j)$	Intra-List Diversity loss for user diversity
$\mathcal{L}_c(\boldsymbol{u}_j)$	Category Coverage loss for content diversity
Intervention Sets and Thresholds	
ILD_{low}	Set of clusters with low Intra-List Diversity scores
U_{botnec}	Set of bottleneck users
θ_m	Threshold for Intra-List Diversity scores
θ_n	Minimum neighbor threshold
θ_c	Content diversity threshold
$\lambda_1, \lambda_2, \lambda_3, \lambda_4$	Loss function weighting parameters

Predictor subsequently utilizes unbiased user embeddings to jointly optimize information diffusion accuracy (via MAE), user interaction diversity (via ILD), and content category coverage (via CC), ensuring both precise predictions and diversity promotion. (4) *Targeted Interventions* finally leverages the optimized predictions and diversity metrics to find optimal intervention points and implement strategic interventions, thereby effectively breaking echo chambers while maintaining user engagement. In this way, CEDA employs causal learning to account for hidden confounders, which improves the information diffusion prediction accuracy and effectively mitigates the echo chamber problem in social networks.

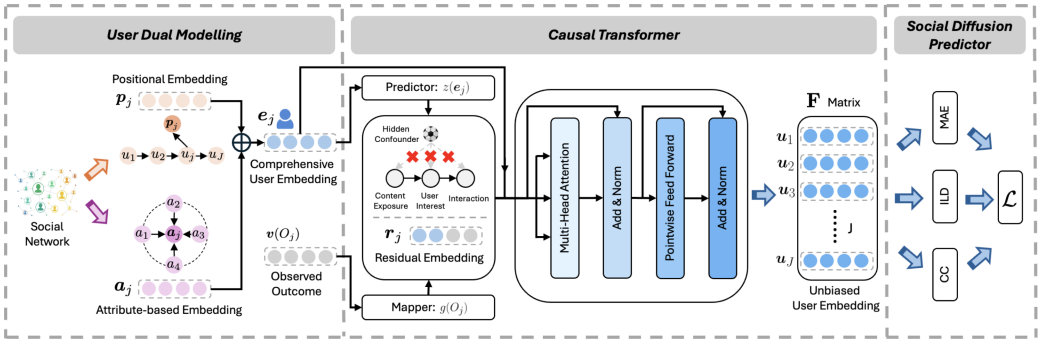

Fig. 3. The overall framework of our proposed method CEDA. First, comprehensive user embeddings are constructed based on users' attributes and structural information within social networks. Second, these comprehensive user embeddings are used to compute their corresponding residual embeddings, which capture discrepancies between predicted outcomes and observed behaviors as the influence of hidden confounders. Then, the residual embeddings are incorporated into the Transformer's attention mechanism as causal adjustments to adjust attention weights, refining the comprehensive user embeddings into unbiased user embeddings that account for hidden confounders. Finally, the unbiased user embeddings are leveraged to jointly optimize information diffusion prediction accuracy (MAE) and diversity metrics (ILD and CC), effectively mitigating echo chambers in the social network.

4.1 User Dual Modelling

To accurately model information diffusion in social networks, it is crucial to comprehensively understand user behavior patterns [64, 79]. Fundamentally, user behaviors in social networks are influenced by two key factors: individual tendencies shaped by users' attributes, which guide their content preferences and sharing patterns, and their structural positions within the network, which determine content exposure and potential interaction pathways [7, 69]. Thus, our first component *User Dual Modelling* synergistically integrates both users' attributes and their structural information in social networks to build comprehensive user embeddings for all users. First, we learn attribute-based user embeddings based on users' observable attributes, capturing individual tendencies in information sharing. Then, we learn positional encoding to capture how users' positions impact their content exposure and information flow patterns. By jointly modeling both attribute and positional information, we gain a deeper understanding of user behavior patterns in the social network.

Formally, we first create the attribute-based vector $\boldsymbol{a}_j \in \mathbb{R}^d$ for each user u_j using a retrieval function [54], which transforms categorical attributes into numerical encodings for compatibility, where d is the dimension. For instance, consider a user with two attribute categories "gender" and "age." The "gender" has values ["male," "female," "unknown"], where "male" can be encoded as [1, 0, 0]. Following social media demographic analysis [56], we could separate the "age" with values ["0–18," "19–35," "34–49," "50+"], where "19–35" can be encoded as [0, 1, 0, 0]. We then concatenate these individual attribute encodings to construct the attribute-based user representation $\boldsymbol{a}_j$ for a male user aged "19–35" as [1, 0, 0, 0, 1, 0, 0]. This numerical representation enables our model to process and analyze user attributes effectively while maintaining the semantic relationships between different demographic categories.

To effectively capture users' structural positions within information diffusion sequences, we employ a sophisticated positional encoding mechanism inspired by [45]. Specifically, we define a matrix $\boldsymbol{P} \in \mathbb{R}^{J \times d}$ where each row vector $\boldsymbol{p}_j \in \mathbb{R}^d$ encodes the sequential position of user u_j in the

sequence S as:

$$\boldsymbol{p}_j = \left[\sin\left(\frac{j}{10{,}000^{2i/d}}\right), \cos\left(\frac{j}{10{,}000^{2i/d}}\right)\right], \tag{2}$$

where j represents the sequential position of user u_j in the information diffusion sequence S. The term $10{,}000^{2i/d}$ serves as a frequency modulator, with dimension index i ranging from 0 to $d/2 - 1$, and d being the total embedding dimension. This creates wavelengths from 2π to $10{,}000 \cdot 2\pi$, allowing the encoding to capture both fine-grained local patterns and broader sequential relationships [45]. In simple terms, for each position j, we compute pairs of sine and cosine values across different frequency bands, generating a unique positional signature that preserves geometric properties regardless of sequence length. Such mathematical construction produces a position-aware representation that effectively captures both absolute positions and relative distances between users in information diffusion pathways. In this way, we can better analyze content dissemination patterns and identify potential echo chambers in social networks.

With the attribute-based representation $\boldsymbol{a}_j$ capturing user characteristics and positional encoding $\boldsymbol{p}_j$ encoding sequential position information, we now combine them to derive the comprehensive user embedding for each user:

$$\boldsymbol{e}_j = (\boldsymbol{a}_j \oplus \boldsymbol{p}_j)\mathbf{W}^E, \tag{3}$$

where $\mathbf{W}^E$ is a learnable weight matrix and $\oplus$ is the concatenation operation. $\boldsymbol{a}_j$ is the attribute-based vector that encodes user u_j's associated attributes. $\boldsymbol{e}_j$ is the comprehensive user embedding for the user u_j, which integrates both attributes and positional data, providing a holistic view of user behavior in the social network.

In summary, our dual embedding approach integrates users' attributes and structural positions within social networks to construct comprehensive user embeddings, forming a robust foundation for accurately modeling information diffusion. Since such dual perspective ensures a nuanced understanding of user behavior, enabling the model to account for not only individual content preferences but also the broader diffusion pathways shaped by network structures. As a result, our approach can enhance the model's ability to predict information flow patterns and identify key nodes or clusters for intervention strategies, thereby improving the overall effectiveness of mitigating echo chambers in social networks.

4.2 Causal Transformer

Although the estimated comprehensive user embeddings form a strong foundation for modeling user behavior, they do not account for the influence introduced by hidden confounders within the social network. Formally, hidden confounders refer to external and unobserved factors that simultaneously affect both content exposure and user interactions. Their presence introduces spurious correlations in observed behavior, which may lead to incorrect inferences about causal mechanisms underlying echo chamber formation and result in ineffective intervention strategies. To address this issue, we design our second component called *Causal Transformer*, which integrates causal inference into the Transformer architecture to explicitly model and adjust for hidden confounders within user behavior sequences. The Transformer, widely employed in **Sequential Recommender Systems (SRSs)**, is particularly effective at modeling sequential social interaction data due to its self-attention mechanism, which captures long-range dependencies and complex user relationships [8, 67]. Additionally, its multi-head attention structure allows CEDA to integrate residual embeddings in a modular and scalable way. We first compute residual embeddings by measuring the discrepancy between predicted and observed outcomes [42, 43]. These discrepancies are then incorporated as causal adjustments within the Transformer's attention mechanism, allowing CEDA

to refine attention weights and effectively account for the influence of hidden confounders. This adjustment ensures that user embeddings more accurately reflect true preferences rather than external confounding effects, ultimately improving information diffusion predictions and enabling more effective interventions to mitigate echo chambers.

4.2.1 Causal Adjustment. As mentioned, a fundamental challenge in analyzing social network behavior is identifying and mitigating hidden confounders that systematically influence user interactions and information diffusion patterns. To address this, we introduce a causal adjustment mechanism, which can be interpreted as performing a causal intervention over hidden confounders, effectively blocking their confounding effects, as shown in Figure 1. Specifically, we aim to compute a residual embedding for each user u_j to account for hidden confounder-induced discrepancies, which serves as a causal adjustment to the comprehensive user embedding $\boldsymbol{e}_j$. These discrepancies arise when a user's behavior is influenced by factors not captured in the observable features, leading to systematic deviations (e.g., a mathematical signal of hidden confounder effects) from expected behavioral patterns. For instance, consider an online discussion network in which users typically engage based on interest similarity or prior connections. If a controversial political speech takes place (e.g., a televised debate) but is not logged as an event in the dataset, it may cause a simultaneous surge in political content engagement across otherwise unrelated users. Since this engagement is driven by the shared influence of the unobserved external event—not by intrinsic preferences or social ties—it results in observed behavior that deviates from model expectations based solely on visible inputs. Without adjusting for such confounding effects from unobserved hidden confounders, a recommendation model may incorrectly infer that these users inherently prefer political content, leading to biased content exposure. Therefore, to quantify the effects from hidden confounders, we first estimate the expected user behavior $z(\boldsymbol{e}_j)$ based on observable features and then compare it with the actual observed outcome $g(O_j)$ as:

$$\boldsymbol{r}_j = \text{MLP}\left(z\left(\boldsymbol{e}_j\right) - g\left(O_j\right)\right), \tag{4}$$

$$z(\boldsymbol{e}_j) = \text{ReLU}(\boldsymbol{W}_z \boldsymbol{e}_j + \boldsymbol{b}_z), \tag{5}$$

$$g(O_j) = \text{ReLU}(\boldsymbol{W}_g \boldsymbol{v}(O_j) + \boldsymbol{b}_g), \tag{6}$$

where MLP is a multi-layer perceptron and O_j is the observed outcome for user u_j. $\boldsymbol{e}_j$ is the comprehensive user embedding generated in Equation (3). $z(\cdot)$ is a prediction function that maps the comprehensive user embedding $\boldsymbol{e}_j$ to an expected behavioral outcome. In other words, $z(\boldsymbol{e}_j)$ represents what behavior we would predict for user u_j based solely on their observable attributes and structural position within the network. Conversely, $g(\cdot)$ is a mapping function that transforms the observed outcome O_j (actual user behavior) into the same representational space as the predictions, enabling direct comparison. $\boldsymbol{v}(O_j)$ is an embedding lookup for the scalar observed outcome O_j, allowing us to map the scalar outcome to a higher-dimensional vector [25, 54]. The difference between these two functions' outputs quantifies the discrepancy between expected and actual behavior, which indicates the effects of hidden confounders [9, 35, 65]. $\mathbf{W}_z$ and $\mathbf{W}_g$ are learnable weight matrices, while $\boldsymbol{b}_z$ and $\boldsymbol{b}_g$ are bias vectors. The ReLU activation function introduces non-linearity to capture complex relationships. Inspired by [45], we train the $z()$ and $g()$ as the following loss:

$$\mathcal{L}_r = \sum_{(u_j, O_j)} \left|z\left(\boldsymbol{e}_j\right) - g\left(O_j\right)\right|^2 + \lambda\left(|\mathbf{W}_z|_F^2 + |\mathbf{W}_g|_F^2\right), \tag{7}$$

where $|\cdot|_F$ represents the Frobenius norm and λ controls regularization strength. The loss function $\mathcal{L}_r$ achieves a crucial balance by minimizing the prediction observation discrepancy while preventing overfitting through careful regularization of the weight matrices. Through systematic

optimization of this loss using stochastic gradient descent, we obtain residual embeddings $\boldsymbol{r}_j$ that effectively quantify the influence of hidden confounders on network behaviors.

4.2.2 Multi-Head Self Attention. To comprehensively capture the effects of hidden confounders from multiple perspectives, we employ a multi-head attention mechanism that processes user interaction patterns through distinct representational lenses [12, 55]. This sophisticated approach enables our model to simultaneously analyze different aspects of user behavior while accounting for hidden confounding effects in each attention head. Mathematically, for each head j, we first project the user embeddings into query, key, and value representations:

$$\text{head}_j = \text{Attention}(\mathbf{E}\mathbf{W}_j^Q, \mathbf{E}\mathbf{W}_j^K, \mathbf{V}\mathbf{W}_j^V, \boldsymbol{r}_j) \tag{8}$$

$$= \text{softmax}\left(\frac{\mathbf{E}\mathbf{W}_j^Q(\mathbf{E}\mathbf{W}_j^K)^\top - \mathbf{E}\mathbf{W}_j^Q(\boldsymbol{r}_j\mathbf{1}_J^\top)^\top}{\sqrt{d_k}}\right)\mathbf{V}\mathbf{W}_j^V, \tag{9}$$

where the matrices $\mathbf{W}_j^Q, \mathbf{W}_j^K, \mathbf{W}_j^V \in \mathbb{R}^{d\times d_k}$ are learnable projections that map each embedding from the original d-dimensional space into d_k-dimensional query, key, and value vectors, respectively. The user embedding matrix $\mathbf{E} = [\boldsymbol{e}_1, \cdots, \boldsymbol{e}_J]^\top \in \mathbb{R}^{J\times d}$ contains comprehensive representations for all users in the network. The matrix $\mathbf{V}$ is typically set equal to $\mathbf{E}$ unless further customized. $\mathbf{1}_J \in \mathbb{R}^J$ is a vector of ones that broadcasts the residual embedding $\boldsymbol{r}_j$ across all users in the sequence. In other words, $\boldsymbol{r}_j$ acts as a correction term, adjusting the attention weights to account for hidden confonuders. Finally, we combine information from all attention heads:

$$\mathbf{O} = (\text{head}_1 \oplus \ldots \oplus \text{head}_j)\mathbf{W}^J, \tag{10}$$

where the concatenation operation $\oplus$ combines the outputs from all attention heads into a matrix $\mathbf{O} \in \mathbb{R}^{J\times d}$. The final projection matrix $\mathbf{W}^J$ integrates these multiple perspectives into a unified representation. Each row in $\mathbf{O}$ represents the attended features for a user, capturing rich interactions while systematically accounting for hidden confounders across all attention mechanisms.

4.2.3 ***Pointwise Feed Forward Neural Network (PFFN).*** The multi-head attention mechanism effectively captures complex relationships between users in the network through learned attention weights. However, these attention-based transformations are inherently linear in nature, which may limit the model's ability to capture nonlinear patterns in user behaviors and information diffusion dynamics. To address this limitation, we incorporate a PFFN [53, 70] after the multi-head attention layer. Specifically, the PFFN introduces crucial nonlinearity through a sophisticated architecture consisting of two affine transformations bridged by a LeakyReLU activation function [72]. This design allows the model to learn more complex functional mappings while maintaining stable gradient flow during training. Mathematically, we have:

$$\mathbf{O}' = \text{LayerNorm}(\mathbf{O} + \text{Dropout}(\phi(\mathbf{O}))), \tag{11}$$

$$\mathbf{H} = \text{LeakyReLU}(\mathbf{O}'\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + b_2, \tag{12}$$

$$\mathbf{F} = \text{LayerNorm}(\mathbf{O}' + \text{Dropout}(\mathbf{H})), \tag{13}$$

where ϕ represents the PFFN operation that processes the input through multiple nonlinear layers. The LayerNorm [73] operation standardizes inputs across feature dimensions by calculating mean and variance statistics, enhancing training stability and convergence. The learnable weight matrices $\mathbf{W}_1$ and $\mathbf{W}_2$ serve as trainable parameters that enable the network to adapt and learn complex patterns through backpropagation. The intermediate representation $\mathbf{O}'$ is obtained by applying dropout regularization and layer normalization to the input $\mathbf{O}$, helping prevent overfitting and ensuring stable gradient flow. The model then generates $\mathbf{H}$ by applying the LeakyReLU activation function to an

affine transformation of $\mathbf{O}'$, chosen specifically for mitigating vanishing gradients while maintaining nonlinearity. The final output matrix $\mathbf{F} \in \mathbb{R}^{J \times d}$ is produced through an additional round of dropout and layer normalization, where each row vector $\boldsymbol{u}_j \in \mathbb{R}^d$ represents the unbiased user embedding for user u_j, capturing both linear and nonlinear relationships within the social network structure [45].

In summary, the *Causal Transformer* integrates residual embeddings as a causal adjustment into the Transformer's attention mechanism, refining comprehensive user embeddings $\boldsymbol{e}_j$ into unbiased user embeddings $\boldsymbol{u}_j$ by accounting for hidden confounders. This refinement ensures that the embeddings accurately represent users' behaviors without the distortions caused by unobserved confounding factors, enabling a more precise understanding of information diffusion dynamics. The unbiased user embeddings not only improve the accuracy of predicting diffusion patterns but also serve as a critical input for designing effective interventions. These interventions can strategically target structural and behavioral factors within the network to disrupt echo chambers, fostering diversity and enhancing the overall flow of information. By addressing hidden confounders, the *Causal Transformer* provides a robust framework for achieving both predictive accuracy and valuable insights in mitigating echo chambers.

4.3 Social Diffusion Predictor

With the unbiased user embeddings, we then introduce third component *Social Diffusion Predictor*, aiming to train a model that accurately predicts information diffusion patterns while promoting diversity. Specifically, we focus on three complementary metrics that collectively capture different aspects of information spread and diversity. First is the MAE, which quantifies the average magnitude of error in predicting diffusion probabilities between user pairs, measuring accuracy in diffusion prediction [28]. Second is the ILD, which measures the diversity of users within each diffusion cascade, measuring content variety within information flows [29, 31]. Third is the CC, which measures the diversity of content exposure in information diffusion across the network [52, 68]. By jointly optimizing these metrics, we aim to train a comprehensive model that not only predicts diffusion accurately but also fosters diversity, facilitating the design of intervention strategies to effectively break echo chambers in social networks.

First, we calculate the MAE to quantify the average absolute difference between predicted and actual diffusion probabilities for each user pair in the social network. Mathematically, MAE is defined as:

$$\mathcal{L}_m(\boldsymbol{u}_j, \boldsymbol{u}_q) = \frac{1}{|E|} \sum_{(j,q) \in E} |D_{jq} - \widehat{D}_{jq}|, \tag{14}$$

where $\mathcal{L}_m(\boldsymbol{u}_j, \boldsymbol{u}_q)$ measures the MAE between predicted diffusion probabilities $\widehat{D}_{jq}$ and the true diffusion probabilities D_{jq} for all user pairs $(j, q) \in E$, where E is the set of user pairs in the network. $\boldsymbol{u}_j$ and $\boldsymbol{u}_q$ are unbiased user embeddings for user u_j and user u_q, respectively, generated through $\mathbf{F}$ in Equation (13). The true diffusion probability D_{jq} is calculated as the ratio of successful information transmissions from u_j to u_q to the total transmissions by u_j: Mathematically, D_{jq} is calculated as:

$$D_{jq} = \frac{|I_{jq}|}{|I_j|}, \tag{15}$$

where $|I_{jq}|$ denotes the number of successful information transmissions from user u_j to user u_q. $|I_j|$ represents the total number of information transmissions initiated by user u_j. The predicted diffusion probability $\widehat{D}_{jq}$ is computed through:

$$\widehat{D}_{jq} = \sigma(\boldsymbol{M}^\top (\boldsymbol{u}_j \oplus \boldsymbol{u}_q)), \tag{16}$$

where σ is the sigmoid function and $\boldsymbol{M}$ is a learnable weight vector. $\boldsymbol{u}_j \oplus \boldsymbol{u}_q$ representing the concatenation of embeddings $\boldsymbol{u}_j$ and $\boldsymbol{u}_q$. A lower MAE indicates better prediction accuracy as it implies smaller differences between predicted and actual diffusion probabilities in the social network.

Second, we calculate the ILD to evaluate the semantic dissimilarity among recommended items within each diffusion cascade. Mathematically, ILD is defined as:

$$\mathcal{L}_d(\boldsymbol{u}_j) = \frac{1}{|C|} \sum_{c \in C} \frac{1}{|c|(|c|-1)} \sum_{j,q \in c} (1 - \cos(\boldsymbol{u}_j, \boldsymbol{u}_q)), \tag{17}$$

where $\mathcal{L}_d(\boldsymbol{u}_j)$ quantifies the ILD by measuring dissimilarity between users in diffusion sequences. C is the set of all diffusion cascades and c represents a specific cascade, with $|c|$ denoting the number of users in cascade c. The term $\cos(\boldsymbol{u}_j, \boldsymbol{u}_q)$ computes the cosine similarity between the unbiased embeddings of users u_j and u_q, measuring how similar their behavioral patterns and preferences are. A higher ILD value indicates that users are presented with content that spans different semantic dimensions, reducing the reinforcement of narrow interest areas. This diversity mitigates echo chamber effects, enabling more balanced exposure to varied content within the social network.

Third, we compute CC to measure the diversity of content categories that users engage with across the network. For each user u_j, we map his/her unbiased embedding $\boldsymbol{u}_j$ to a probability distribution over k predefined content categories (e.g., sports, technology) through a learnable transformation:

$$\text{Categories}(\boldsymbol{u}_j) = \sigma(\mathbf{W}_c \boldsymbol{u}_j + \mathbf{b}_c), \tag{18}$$

where $\mathbf{W}_c \in \mathbb{R}^{k \times d}$ is the learnable parameter. The output $\text{Categories}(\boldsymbol{u}_j) \in \mathbb{R}^k$ represents the probabilities of user u_j participating in each of the predefined k categories. After obtaining these category probabilities for all users, we then compute the diversity of content exposure across the whole network as:

$$\mathcal{L}_c(\boldsymbol{u}_j) = \frac{|\bigcup_{j=1}^{J} \text{Categories}(\boldsymbol{u}_j)|}{k}, \tag{19}$$

where $\mathcal{L}_c(\boldsymbol{u}_j)$ measures CC as the ratio of unique categories that have user participation to the total number of predefined categories k. A higher CC score reflects greater diversity in the content exposure, indicating a broader range of topics being engaged with across the network. This increase in topic diversity suggests effective disruption of echo chambers and more balanced information dissemination, aligning with the goal of reducing self-reinforcing cycles in the social network.

Finally, we integrate causal residual loss, prediction accuracy, and several diversity metrics into a composite loss function to jointly optimize our model as follows:

$$\mathcal{L} = \lambda_1 \mathcal{L}_r + \lambda_2 \mathcal{L}_m(\boldsymbol{u}_j, \boldsymbol{u}_q) + \lambda_3 (1 - \mathcal{L}_d(\boldsymbol{u}_j)) + \lambda_4 (1 - \mathcal{L}_c(\boldsymbol{u}_j)), \tag{20}$$

where $\mathcal{L}_r$ ensures the model accounts for the effects of hidden confounders through causal residual adjustments. $\mathcal{L}_m$ minimizes the MAE between predicted and actual diffusion probabilities, enhancing prediction accuracy. $(1 - \mathcal{L}_d)$ promotes user diversity within diffusion cascades, reducing homogenization and improving exposure to varied perspectives. $(1 - \mathcal{L}_c)$ encourages broader engagement across diverse content categories, fostering a more balanced network-wide information flow. The parameters λ_1, λ_2, λ_3, and λ_4 control the relative importance of each term, enabling the model to balance accuracy, causal adjustment, and diversity objectives effectively. By simultaneously optimizing these objectives, CEDA achieves robust diffusion predictions while addressing hidden confounders and promoting diversity, ultimately mitigating echo chambers in social networks.

4.4 Targeted Interventions

Leveraging our trained model, the last component *Targeted Interventions* develops systematic strategies to effectively mitigate echo chambers while preserving user engagement in social networks. The key idea lies in utilizing CEDA's robust predictive insights to identify optimal intervention points within the network structure. By analyzing both network topology and information diversity metrics, we can detect critical junctures where targeted modifications will have maximum impact in disrupting echo chambers. Specifically, our approach focuses on two complementary strategies: first, identifying low-diversity clusters where strategic edge rewiring can introduce diverse perspectives; and second, leveraging bottleneck users who can serve as bridges between isolated communities. The interventions are carefully calibrated using our model's predictions of information flow patterns and diversity metrics, ensuring that modifications enhance network diversity while maintaining user engagements. This data-driven approach enables the strategic reshaping of information diffusion across the social network, transforming isolated echo chambers into more interconnected communities.

4.4.1 Intervention Points. To systematically identify the most effective intervention points, we analyze both the network topology and information diversity through a dual-perspective approach:

— *Low Diversity Clusters*: User clusters with low ILD scores indicate the presence of potential echo chambers, as these groups primarily share and interact with similar content where diverse perspective exposure needs enhancement. Mathematically, the clusters with low ILD scores are defined as:

$$ILD_{low} = \left\{ c \in C | \frac{1}{|c|} \sum_{u_j \in c} \mathcal{L}_d(\boldsymbol{u}_j) < \theta_m \right\}, \tag{21}$$

where ILD_{low} is the set of clusters with average ILD scores below the diversity threshold θ_m. $|c|$ is the number of users in cluster c and $\mathcal{L}_d(\boldsymbol{u}_j)$ is the ILD score of user u_j. To ensure intervention quality, we further compute the Silhouette coefficient [59] for each cluster. Only those with Silhouette scores above 0.6, indicating strong internal cohesion, are selected as viable intervention targets.

— *Bottleneck Users*: These users are structurally positioned across multiple communities but demonstrate limited content diversity in what they share, reflected by low CC scores [40]. Their strategic network positions offer potential for facilitating diverse information flows across multiple communities, but are not fully utilized. Mathematically, bottleneck user is defined as:

$$U_{botnec} = \{u_j \in \mathcal{U} | N(u_j) \geq \theta_n \wedge \mathcal{L}_c(\boldsymbol{u}_j) < \theta_c\}, \tag{22}$$

where U_{botnec} represents the set of bottleneck users, $N(u_j)$ is the number of neighboring communities for user u_j, θ_n is the minimum threshold for neighbors, and θ_c is the content diversity threshold.

4.4.2 Intervention Strategies. Building upon the intervention points, we develop targeted intervention strategies to mitigate echo chambers and promote information diversity, each addressing specific network patterns that contribute to echo chamber formation:

— ***Diversity-Aware Content Injection (DA-CI)***: For user clusters with low ILD, we apply a structural rewiring strategy to inject diverse information sources [75, 82]. Specifically, we identify the highest-degree user within each low-diversity cluster—who often acts as a central information amplifier—and remove one of their incoming edges. Simultaneously, we add a new incoming edge from a user located outside the cluster but belonging to a semantically

diverse region of the network. This targeted edge replacement disrupts echo-amplifying hubs, opens new exposure channels, and increases content heterogeneity within the cluster. These modifications are expected to improve ILD within the affected clusters and mitigate structural content homogeneity that reinforces echo chamber effects.

— ***Cross-Category Bridging (CCB)***: For bottleneck users—those structurally connected across multiple communities but exhibiting limited content diversity [76], we introduce new outbound connections to users in different communities who engage with distinct content categories. These additional edges are selected to maximize the target user's potential CC, thereby enhancing the overall diversity of content shared across the network. This strategy leverages the structural position of bottleneck users to facilitate cross-community information flow. By broadening the scope of content dissemination, this intervention aims to reduce inter-community segregation and promote exposure to heterogeneous viewpoints, thus alleviating the reinforcing cycles characteristic of echo chambers.

4.4.3 Network Topology Analysis. To evaluate the structural effects of our proposed interventions on echo chamber mitigation, we analyze changes in three fundamental network metrics: modularity, clustering coefficient, and the size of the **Giant Connected Component (GCC)**. These metrics collectively capture the degree of community fragmentation, local cohesion, and global connectivity within the user network.

Modularity measures the extent to which a network is partitioned into densely connected subgroups with sparse inter-group connections [48]. High modularity values indicate strong community boundaries, which often correspond to isolated echo chambers. A decrease in modularity following intervention suggests that the structural separation between communities is reduced, implying a weakening of echo chamber boundaries and increased potential for cross-community information flow.

Clustering coefficient quantifies the density of local connections within user communities [57]. In the context of echo chambers, high clustering indicates that users are embedded in tightly-knit groups with redundant connections, reinforcing repeated exposure to similar content. A decrease in clustering coefficient reflects a reduction in localized redundancy, indicating that new connections are disrupting closed user loops and enabling access to previously unavailable information.

GCC size refers to the proportion of users belonging to the largest strongly connected subgraph [15]. An increase in GCC size signals enhanced global connectivity and the emergence of a more integrated network structure. This is critical for diluting the influence of isolated communities and ensuring that diverse content can diffuse across broader portions of the network.

In our experiments, we construct user-user interaction networks based on shared content engagement and apply the proposed interventions (DA-CI and CCB) to modify the network structure. We then compute the modularity, average clustering coefficient, and GCC size of the modified networks using standard graph analysis libraries such as NetworkX. These metrics are compared against the original (pre-intervention) network to assess the structural impact of each intervention. Specifically, we observe that successful interventions lead to reduced modularity and clustering coefficient, along with increased GCC size, thereby providing concrete evidence that our approach weakens structural isolation and enhances network-wide content diversity.

In summary, our *Targeted Interventions* component systematically identifies optimal intervention points by analyzing network topology and diversity metrics, such as low-diversity clusters and bottleneck users. It then implements multidimensional strategies, including DA-CI and CCB, to strategically reshape information diffusion. These targeted modifications disrupt self-reinforcing feedback loops, enhance interconnectivity, and promote diverse perspectives across the social network, effectively breaking echo chambers while preserving the user engagement.

4.5 Complexity Analysis

Regarding the time complexity, CEDA consists of four primary components with distinct computational requirements: The first component *User Dual Modelling* requires $O(|U| \cdot d)$ operations for generating comprehensive user embeddings, where $|U|$ represents the total number of users and d is the embedding dimension; The second component *Causal Transformer*, which forms the computational core, has a total complexity of $O(|S|^2 \cdot d + |U| \cdot d)$, where $O(|S|^2 \cdot d)$ for self-attention operations and $O(|U| \cdot d)$ for residual embedding computation. Note that the computational cost of standard transformer is dominated by self-attention, which has complexity $O(|S|^2 \cdot d)$, where $|S|$ is the sequence length and d is the embedding dimension. Since $|S|^2$ is typically much larger than $|U|$, the additional computation for residual embeddings is marginal compared to the overall cost of self-attention. However, this minor overhead significantly enhances causal adjustment, improving prediction accuracy while maintaining practical scalability. The third component *Social Diffusion Predictor* requires $O(|U| \cdot |C|)$ operations for computing diffusion probabilities and diversity metrics, where $|C|$ represents the number of content categories; The last component *Targeted Interventions* has a complexity of $O(|E|)$ for analyzing network edges, where $|E|$ represents the number of edges. The total time complexity is therefore expressed as $O(|U| \cdot d + |S|^2 \cdot d + |U| \cdot |C| + |E|)$. In practice, this simplifies to $O(|S|^2 \cdot d)$ as the sequence length term typically dominates in real-world scenarios. For space complexity, CEDA requires $O(|U| \cdot d)$ to store user embeddings and $O(|S|^2)$ for attention matrices, reflecting its efficient use of memory resources. These characteristics highlight our CEDA's scalability and practicality for large-scale social networks, enabling it to handle complex interaction patterns and deliver robust performance across diverse datasets.

5 Experiments

We conduct extensive experiments to evaluate the effectiveness of our proposed CEDA, addressing the following key research questions:

- *RQ1*: How does CEDA performs in predicting information diffusion patterns and mitigating echo chambers?
- *RQ2*: How do different components of CEDA contribute to its overall performance?
- *RQ3*: How do different hyperparameter settings affect CEDA's performance?
- *RQ4*: How effective are CEDA's intervention strategies in mitigating echo chambers on real-world social networks?

5.1 Settings

5.1.1 Datasets. We evaluate our approach using three real-world social network datasets: *Twitter*,[1] *Google+*,[2] and *Facebook*.[3] Table 2 summarizes the key statistics, including the number of networks, users, average clustering coefficient (Avg. CC), and network diameter. The *Twitter* dataset includes 973 social circles with 81,306 users, showing moderate interconnectivity reflected in an average clustering coefficient of 0.5653. The *Google+* dataset is larger, encompassing 107,614 users across 132 circles, but features sparser connectivity, with an average clustering coefficient of 0.4901. In contrast, the *Facebook* dataset consists of 4,039 users spread across 10 circles, representing a smaller but more densely connected network with a clustering coefficient of 0.6055. These datasets capture diverse social network characteristics, including variations in scale, connectivity, and community structures, providing a comprehensive evaluation context. To ensure high-quality data and meaningful interaction analysis, sequences containing fewer than five interactions are excluded.

[1] https://snap.stanford.edu/data/ego-Twitter.html.
[2] https://snap.stanford.edu/data/ego-Gplus.html.
[3] https://snap.stanford.edu/data/ego-Facebook.html.

Table 2. Statistics of the Three Selected Real-World Datasets

Statistics	Twitter	Google+	Facebook
#Networks	973	132	10
#Users	81,306	107,614	4,039
Avg. CC	0.5653	0.4901	0.6055
Diameter	7	6	8

We partition the processed data into training (70%), validation (10%), and test (20%) subsets. This partitioning strategy balances robust model training and reliable evaluation, reducing the risk of overfitting and enabling a thorough assessment of our approach across different social network environments.

5.1.2 Implementation. We implement our proposed model CEDA using the PyTorch framework, leveraging its flexibility and efficiency for deep learning tasks across our selected three real-world datasets. For the *User Dual Modelling* component, we set the dimension for user representations to 128 to capture sufficient behavioral features while maintaining computational efficiency. The *Causal Transformer* component employs 16 attention heads to effectively model multi-aspect user preferences and relationships across different social network contexts, with a dropout rate of 0.1 applied uniformly across all layers to mitigate overfitting risks. To optimize the *Social Diffusion Predictor* component, we conduct an extensive grid search to fine-tune the loss function weights λ_1, λ_2, and λ_3 within the range [0.1, 1.0] at 0.1 intervals. This can ensure the balanced optimization between diffusion accuracy and diversity objectives. The model is optimized using the Adam optimizer [34] with a learning rate of 0.001 and a batch size of 256, ensuring stable convergence and efficient training across different network scales. For fair evaluation, all baseline models undergo the same rigorous parameter tuning process within identical ranges, allowing fair performance comparisons across different social network environments.

5.1.3 Evaluation. To comprehensively evaluate CEDA's effectiveness, we implement a multi-faceted assessment framework using five established metrics that measure both predictive accuracy and echo chamber mitigation capabilities. For assessing information diffusion prediction, we employ **Root Mean Square Error (RMSE)** [28] to quantify the overall prediction accuracy across different network contexts, complemented by Precision@K and Recall@K metrics [13] that provide detailed insights into both accuracy and coverage of our Top-K predicted diffusion paths. The evaluation of echo chamber mitigation effectiveness relies on two key metrics. First is the Gini coefficient [23], which measures the equality of information distribution across the network by quantifying how evenly content is shared among users. A lower Gini value suggests that recommended content is more evenly distributed across different items, reducing popularity bias and ensuring that niche content is not overshadowed by mainstream recommendations. Second is the **Simpson's Diversity Index (SDI)** [33, 61, 80], which measures the probability that two randomly selected recommended items belong to different categories, providing a direct measure of content diversity and echo chamber reduction. A higher SDI indicates greater content diversity in recommendations, making it particularly useful for evaluating whether a model reduces content homogeneity and exposes users to a broader range of topics. With the above comprehensive evaluation approach allows us to assess both the model's predictive capabilities and its effectiveness in promoting diverse information flow. To ensure fair comparison with baseline methods that may specialize in either echo chamber detection or diversity promotion, we adapt their strategies within our SRS framework while preserving their core functionalities.

5.2 Baselines

We compare our CEDA with below state-of-the-art methods:

- *FRECH* [63]: This uses a GCN with wide architecture to learn user behaviors and implicit echo chamber representations to recommend diverse friends from outside echo chamber.
- *CECD* [46]: This employs a probabilistic generative model to detect echo chambers and characterize their influence on information propagation in social networks.
- *OCR* [66]: This proposes a quadratic optimization framework to recommend diverse content and reduce echo chambers in social media, while accounting for user preferences.
- *ECS* [2]: This uses embedding distances between users to quantify echo chamber effects in social networks, helping devise strategies to mitigate echo chamber effects.
- *ECM* [58]: This proposes an agent-based model to study how social influence and unfollowing behaviors impact the emergence of echo chambers in online social networks.
- *GRU4Rec* [27]: This uses recurrent neural networks with GRU units to model user sequential behaviors in session-based recommendation.
- *NARM* [36]: This employs an attention mechanism with RNNs to capture both sequential behavior and main purpose in session-based recommendation.
- *STAMP* [41]: This proposes a short-term attention priority model to capture both long-term and short-term user interests in session-based recommendation.
- *LLMS* [24]: This employs Large Language Models to capture semantic relations between items and enhance item representations, improving sequential recommendations.
- *ReFor* [37]: This uses a transformer-based model to learn language representations for sequential recommendation, improving cold-start performance and domain transfer.
- *DCCF* [74]: This uses counterfactual reasoning and back-door adjustment to mitigate echo chambers in recommender systems while maintaining recommendation performance.

5.3 RQ1. Performance Analysis

To comprehensively evaluate CEDA's effectiveness, we conduct extensive experiments on three real-world datasets, analyzing both quantitative performance metrics and the model's ability to address echo chambers. As shown in Table 3, CEDA consistently outperforms state-of-the-art methods across all evaluations, demonstrating significant improvements in both accuracy and ranking metrics. In terms of prediction accuracy, CEDA achieves substantial improvements in RMSE, with 13.38%, 10.05%, and 12.44% lower error rates compared to the best-performing baselines on *Twitter*, *Google+*, and *Facebook*, respectively. This improvement in accuracy indicates CEDA's superior ability to capture the underlying patterns in information diffusion. The performance gains are even more pronounced in ranking metrics, particularly at higher K values where capturing long-term dependencies becomes crucial. For instance, on the *Twitter*, CEDA demonstrates 14.22% higher Precision@40 and 11.65% higher Recall@40 compared to the strongest baseline. On the *Google+* and *Facebook*, CEDA achieves improvements of 5.99% and 7.58% in Precision@40, and 2.09% and 2.63% in Recall@40, respectively. These results highlight CEDA's robustness in modeling complex information diffusion dynamics across diverse social network environments.

Beyond prediction accuracy, we assess CEDA's ability to mitigate echo chambers using two widely used diversity metrics: the Gini coefficient [23] and SDI [61, 80]. The Gini coefficient measures the evenness of information distribution within the network, with lower values indicating a more balanced exposure to information across user groups. SDI, on the other hand, quantifies the diversity of information categories that users are exposed to, with lower values suggesting increased exposure to diverse content. Our results demonstrate that CEDA effectively reduces the Gini coefficient by 12.23%, 5.44%, and 10.36% on *Twitter*, *Google+*, and *Facebook*, respectively,

Table 3. Performance Comparison

Dataset	Metric	GRU4Rec	NARM	STAMP	LLMS	ReFor	CECD	OCR	FRECH	DCCF	ECS	CEDA	Improv.%
Twitter	RMSE	0.3621	0.3598	0.3567	0.3512	0.3489	0.3556	0.4102	0.3645	0.3599	0.3532	**0.3022**	13.38%
	Pr@5	0.5312	0.5389	0.5456	0.5634	0.5601	0.5215	0.5301	0.5589	0.5625	0.5578	**0.6015**	6.76%
	Pr@10	0.6089	0.6156	0.6223	0.6301	0.6378	0.5987	0.6056	0.6312	0.6378	0.6467	**0.6898**	6.67%
	Pr@20	0.6756	0.6823	0.6901	0.6978	0.7056	0.6654	0.6721	0.6978	0.7045	0.7156	**0.7689**	7.45%
	Pr@40	0.7412	0.7489	0.7567	0.7645	0.7723	0.7312	0.7389	0.7601	0.7878	0.7745	**0.8998**	14.22%
	Re@5	0.4956	0.5023	0.5101	0.5178	0.5356	0.4856	0.4923	0.5145	0.5201	0.5289	**0.5789**	8.08%
	Re@10	0.5678	0.5745	0.5823	0.5901	0.5978	0.5578	0.5645	0.5889	0.5956	0.6045	**0.6587**	8.97%
	Re@20	0.6345	0.6412	0.6489	0.6567	0.6745	0.6245	0.6312	0.6567	0.6634	0.6723	**0.7378**	9.38%
	Re@40	0.7301	0.7378	0.7456	0.7634	0.7612	0.7401	0.6978	0.7212	0.7289	0.7367	**0.8523**	11.65%
	Gini	0.5923	0.5856	0.5789	0.5723	0.5656	0.6123	0.5987	0.5678	0.5534	0.5456	**0.4789**	12.23%
	SDI	0.4122	0.4044	0.3966	0.3877	0.3799	0.4322	0.4211	0.3844	0.3511	0.3622	**0.2766**	23.63%
Google+	RMSE	0.3456	0.3423	0.3389	0.3356	0.3323	0.3401	0.3956	0.3489	0.3423	0.3378	**0.2989**	10.05%
	Pr@5	0.5445	0.5523	0.5601	0.5778	0.5756	0.5345	0.5432	0.5712	0.5789	0.5723	**0.6137**	6.21%
	Pr@10	0.6223	0.6301	0.6378	0.6456	0.6534	0.6123	0.6201	0.6456	0.6523	0.6578	**0.6964**	5.87%
	Pr@20	0.6889	0.6967	0.7045	0.7123	0.7201	0.6789	0.6867	0.7123	0.7189	0.7245	**0.7701**	6.29%
	Pr@40	0.7556	0.7634	0.7712	0.7789	0.7867	0.7456	0.7534	0.7745	0.8023	0.7889	**0.8504**	5.99%
	Re@5	0.5089	0.5167	0.5245	0.5323	0.5501	0.4989	0.5056	0.5289	0.5345	0.5434	**0.5751**	4.54%
	Re@10	0.5812	0.5889	0.5967	0.6045	0.6123	0.5712	0.5789	0.6023	0.6101	0.6201	**0.6559**	5.77%
	Re@20	0.6489	0.6567	0.6645	0.6823	0.6801	0.6389	0.6456	0.6712	0.6789	0.6778	**0.7274**	6.61%
	Re@40	0.7445	0.7523	0.7601	0.7678	0.7756	0.7545	0.7123	0.7356	0.7434	0.7512	**0.7918**	2.09%
	Gini	0.5789	0.5723	0.5656	0.5589	0.5523	0.5989	0.5845	0.5534	0.5389	0.5312	**0.5023**	5.44%
	SDI	0.4088	0.3999	0.3911	0.3822	0.3733	0.4188	0.4077	0.3699	0.3366	0.3477	**0.2988**	11.23%
Facebook	RMSE	0.3689	0.3656	0.3623	0.3489	0.3556	0.3589	0.4234	0.3767	0.3701	0.3534	**0.3055**	12.44%
	Pr@5	0.5201	0.5278	0.5356	0.5634	0.5512	0.5101	0.5189	0.5467	0.5534	0.5489	**0.5957**	5.73%
	Pr@10	0.5967	0.6045	0.6123	0.6201	0.6378	0.5867	0.5945	0.6201	0.6267	0.6323	**0.6834**	7.15%
	Pr@20	0.6634	0.6712	0.6789	0.6867	0.6945	0.6534	0.6601	0.6856	0.6923	0.7001	**0.7578**	8.24%
	Pr@40	0.7289	0.7367	0.7445	0.7523	0.7601	0.7189	0.7267	0.7478	0.7545	0.7612	**0.8189**	7.58%
	Re@5	0.4845	0.4923	0.5001	0.5078	0.5256	0.4745	0.4812	0.5034	0.5089	0.5167	**0.5567**	5.92%
	Re@10	0.5567	0.5645	0.5723	0.5801	0.5878	0.5467	0.5534	0.5778	0.5845	0.5912	**0.6395**	8.17%
	Re@20	0.6234	0.6312	0.6389	0.6667	0.6545	0.6134	0.6201	0.6456	0.6523	0.6589	**0.7134**	7.00%
	Re@40	0.7189	0.7267	0.7345	0.7423	0.7601	0.7289	0.6867	0.7101	0.7278	0.7256	**0.7801**	2.63%
	Gini	0.6034	0.5967	0.5901	0.5834	0.5767	0.6234	0.6101	0.5789	0.5845	0.5867	**0.5189**	10.36%
	SDI	0.4333	0.4244	0.4155	0.4066	0.3977	0.4433	0.4322	0.3955	0.3733	0.3622	**0.3155**	12.89%

The best results are bolded, while the best baselines are underlined.

indicating a more balanced information distribution among user groups. Furthermore, CEDA achieves significant improvements in SDI, with 23.63%, 11.23%, and 12.89% higher diversity scores on *Twitter*, *Google+*, and *Facebook*, respectively, compared to the best-performing baselines. These findings showcase CEDA's effectiveness in broadening users' exposure to diverse content and mitigating echo chamber effects across different social network environments.

5.4 RQ2. Ablation Study

To evaluate the contribution of each component in our proposed CEDA, we perform ablation studies by systematically modifying its elements. Specifically, we compare four variations: *DM*, where *User Dual Modelling* is replaced with embeddings based solely on user attributes, removing structural context; *CT*, where the *Causal Transformer* is excluded, using comprehensive embeddings without causal adjustments, thereby neglecting hidden confounders; *TI*, where *Targeted Interventions* employs a random forest model to predict intervention points, bypassing diversity-aware strategies; and *ALL*, the full CEDA model with all components intact.

Figure 4 reveals how each component uniquely contributes to mitigating echo chambers, improving diversity, and enhancing prediction accuracy. It is worth noting that the *Causal Transformer* has the most significant impact on overall performance, with its removal causing substantial drops in Precision@40 by 36.76% on *Twitter*, 22.66% on *Facebook*, and 7.48% on *Google+*. This highlights the critical role of addressing hidden confounders in complex social networks. On the other hand, our

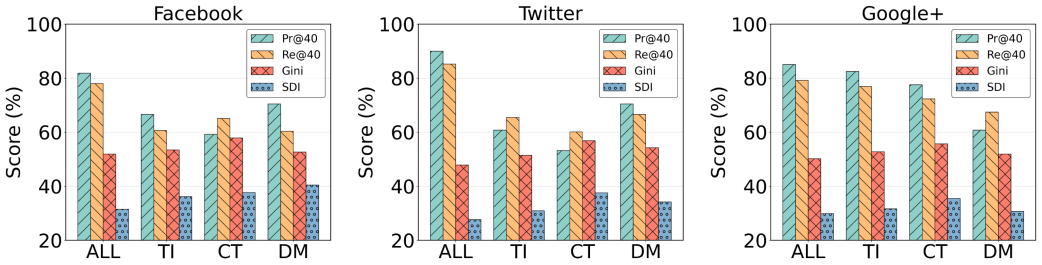

Fig. 4. Ablation study demonstrating the contribution of each CEDA's component across three datasets.

trained model in *Targeted Interventions* shows varying importance across platforms, with its removal decreasing Precision@40 by 29.20% on *Twitter*, 15.33% on *Facebook* and 2.59% on *Google+*. This highlights the effectiveness of our approach in defining the optimal intervention point. Moreover, *User Dual Modelling* demonstrates varying importance across different platforms, with its removal reducing Precision@40 by 19.53% on *Twitter*, 24.26% on *Google+*, and 8.44% on *Facebook*. Regarding diversity metrics, the *Causal Transformer* proves particularly effective in maintaining lower Gini coefficients, as its removal increases these values by 9.0% on *Twitter*, 5.44% on *Google+*, and 6.0% on *Facebook*. These findings demonstrate that while each component contributes differently across platforms, the *Causal Transformer* for addressing hidden confounders consistently plays the most crucial role in mitigating echo chambers across diverse social networks.

5.5 RQ3. Hyperparameters Analysis

To assess CEDA's sensitivity for various parameter configurations and optimize its performance, we conduct comprehensive hyperparameter analysis on three real-world datasets. Specifically, we focus on four key hyperparameters. First, the user embedding dimensions (d) determine the depth of user preference insights, with higher dimensions providing richer representations at the cost of greater computational overhead. Second, the number of attention heads (j), as defined in Equation (10), influences the model's capacity to process multiple relevance signals simultaneously, with more heads enhancing parallelism and accuracy but increasing memory requirements. Third, the batch size impacts learning stability, where larger batches stabilize gradient updates, and smaller ones offer faster but noisier convergence. Finally, the learning rate controls the speed of convergence, with higher rates accelerating training but risking overshooting, while lower rates ensure precision at the expense of longer training times. By systematically varying these parameters, we analyze their effects on prediction accuracy, diversity metrics, and computational efficiency. Thereby, providing insights into optimal configurations that ensure robust and scalable performance across diverse social environments.

As shown in Figures 5–7, CEDA demonstrates distinctive hyperparameter sensitivity patterns across different social network datasets. For attention mechanism, we find the optimal attention head number is 16 achieving Precision@40 of 89.98% on *Twitter* while 8 attention heads perform best on both *Google+* and *Facebook* reaching 85.04% and 81.89%. This suggests that social networks with denser user interactions require more sophisticated attention mechanisms to capture complex behavioral patterns. For batch size, we find a batch size of 256 achieves the best results on *Twitter* and *Facebook* while a larger batch size of 512 performs better on *Google+*. This suggests that the optimal batch size correlates with network scale as larger networks may require more samples per iteration to learn stable patterns. For embedding dimensions, we find 128-dimensional embeddings achieve the best Precision@40 of 89.98% and 81.89% on *Twitter* and *Facebook* respectively, while 256 dimensions work better on *Google+* achieving Precision@40 of 85.04%. This suggests that

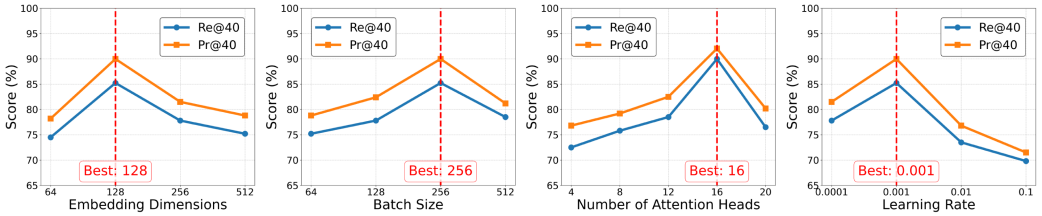

Fig. 5. Parameter sensitivity analysis on the Twitter dataset.

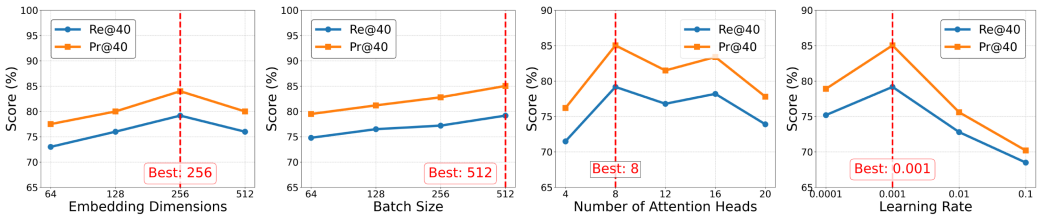

Fig. 6. Parameter sensitivity analysis on the Google+ dataset.

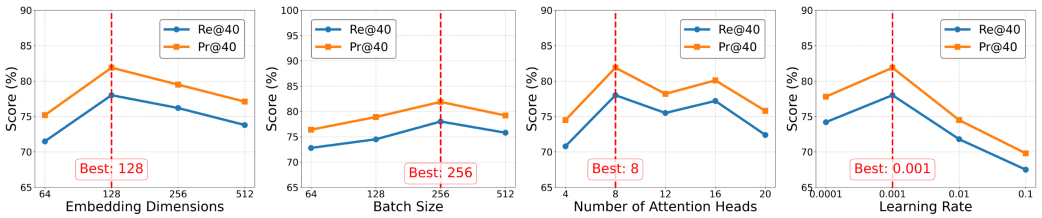

Fig. 7. Parameter sensitivity analysis on the Facebook dataset.

different social networks require varying embedding dimensions to effectively capture their unique interaction patterns and information diffusion characteristics. For learning rate, we find a learning rate of 0.001 consistently achieves peak performance on all three datasets with sharp degradation at higher or lower values. This suggests that proper learning rate calibration plays a fundamental role in model convergence regardless of network characteristics. Overall, these findings highlight that effective echo chamber mitigation requires careful platform-specific hyperparameter tuning as the optimal configuration partially depends on the network size and interaction patterns.

5.6 RQ4. Understanding Designed Intervention Strategies

To better demonstrate the effectiveness of our intervention strategies, we analyze their impacts on echo chamber mitigation across three social network platforms. Following prior works [6, 64], we randomly select 20 users from *Twitter*, *Google+*, and *Facebook* to construct three representative networks by extracting information diffusion paths originating from these users. As shown in Figures 8–10, the initial *Before* snapshots revealed distinct isolated echo chambers in these three networks, with *Twitter* showing three chambers, *Google+* showing four chambers and *Facebook* showing five chambers. These echo chambers are characterized by dense internal connections within each group but sparse connections between different groups, where users predominantly interact within their closed communities. Such isolated structures limit users' exposure to diverse perspectives and reinforce existing viewpoints, leading to increased polarization and deteriorated public discourse in social networks. To break echo chambers, we detect low diversity clusters

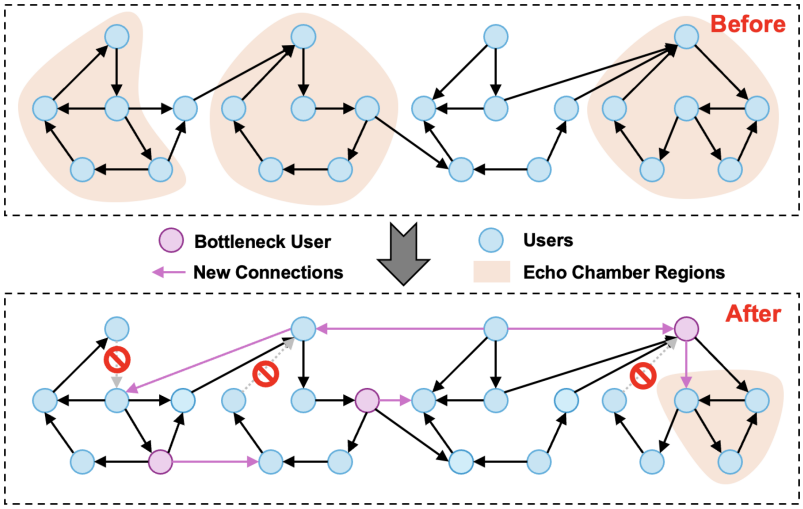

Fig. 8. Visualization of CEDA's echo chamber mitigation effect on Twitter.

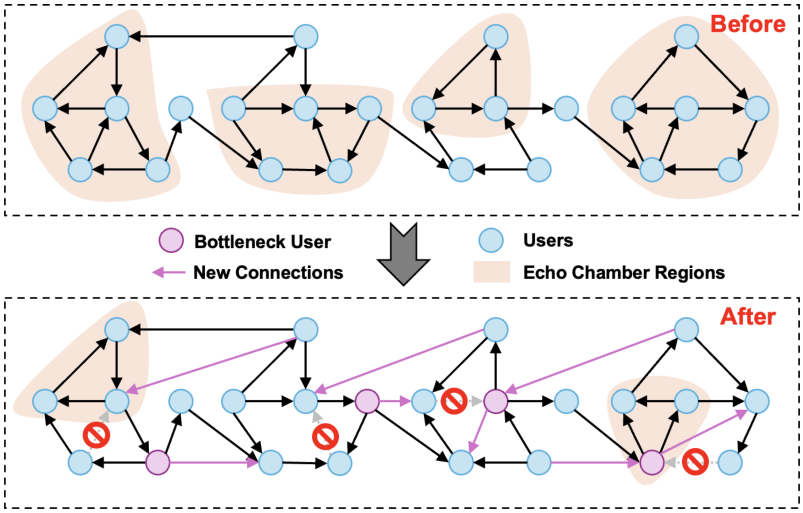

Fig. 9. Visualization of CEDA's echo chamber mitigation effect on Google+.

ILD_{low} based on Equation (21) and identify bottleneck users U_{botnec} based on Equation (22) as intervention points. Following that, we implement two targeted strategies across these networks. First, we remove input connections marked by red prohibition signs from high-degree users who acted as central information hubs within each low diversity cluster, limiting their ability to dominate information flow. This can strategically reduce the reinforcement of existing viewpoints and breaks down information monopolies. Second, we leverage bottleneck users marked as purple nodes who demonstrated interest across multiple topics to create new cross-cluster connections shown as purple lines. These new connections create pathways for diverse content and perspectives to flow between previously isolated groups. The effectiveness of these interventions is evident in the *After* snapshots where the number of echo chambers decreases significantly from three to one on *Twitter*, four to two on *Google+*, and four to zero on *Facebook*. More importantly, the new purple

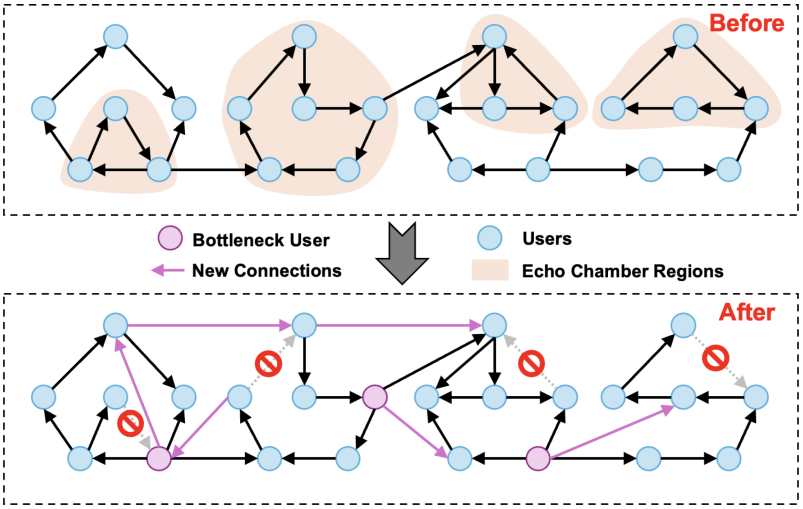

Fig. 10. Visualization of CEDA's echo chamber mitigation effect on Facebook.

bridging connections create increased interconnectivity between previously isolated communities, successfully transforming segregated echo chambers into more diverse social networks. These results validate that CEDA's causal learning through residual embedding analysis can effectively address hidden confounders in social networks, enabling targeted interventions to systematically mitigate echo chambers while maintaining the diverse information flow.

6 Conclusion and Future Work

In this article, we introduce CEDA, a novel causal-based approach that integrates causal inference with the Transformer architecture within a sequential recommendation system framework to address the persistent challenge of echo chambers in social networks. By leveraging both the sophisticated temporal modeling capabilities of Transformers and the analytical power of causal inference to uncover hidden confounders, CEDA successfully mitigates the effects of echo chambers. Our experimental evaluation across three diverse real-world datasets demonstrates CEDA's exceptional performance, consistently outperforming existing state-of-the-art methods in both predicting information diffusion patterns and effectively reducing echo chamber effects. Future work may explore applying causal techniques to process multimodal social network data, potentially uncovering additional confounders in broader social contexts.

References

[1] Luca Maria Aiello, Alain Barrat, Rossano Schifanella, Ciro Cattuto, Benjamin Markines, and Filippo Menczer. 2012. Friendship prediction and homophily in social media. *ACM Transactions on the Web* 6, 2 (2012), 1–33.

[2] Faisal Alatawi, Paras Sheth, and Huan Liu. 2023. Quantifying the echo chamber effect: An embedding distance-based approach. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, 38–45.

[3] Vladimir A. Atanasov and Bernard S. Black. 2016. Shock-based causal inference in corporate finance and accounting research. *Critical Finance Review* 5 (2016), 207–304.

[4] Christopher A. Bail, Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences* 115, 37 (2018), 9216–9221.

[5] Teodora Baluta, Shiqi Shen, S. Hitarth, Shruti Tople, and Prateek Saxena. 2022. Membership inference attacks and generalization: A causal perspective. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, 249–262.

[6] Alessandro Bozzon, Marco Brambilla, Stefano Ceri, Matteo Silvestri, and Giuliano Vesci. 2013. Choosing the right crowd: Expert finding in social networks. In *Proceedings of the 16th International Conference on Extending Database Technology*, 637–648.

[7] Qian Chen, Jianjun Li, Zhiqiang Guo, Guohui Li, and Zhiying Deng. 2023. Attribute-enhanced dual channel representation learning for session-based recommendation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 3793–3797.

[8] Qiwei Chen, Huan Zhao, Wei Li, Pipei Huang, and Wenwu Ou. 2019. Behavior sequence transformer for e-commerce recommendation in Alibaba. In *Proceedings of the 1st International Workshop on Deep Learning Practice for High-Dimensional Sparse Data*, 1–4.

[9] Wei Chen, Ruichu Cai, Kun Zhang, and Zhifeng Hao. 2021. Causal discovery in linear non-Gaussian acyclic model with multiple latent confounders. *IEEE Transactions on Neural Networks and Learning Systems* 33, 7 (2021), 2816–2827.

[10] Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini.2021. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences* 118, 9 (2021), e2023301118.

[11] Federico Cinus, Marco Minici, Corrado Monti, and Francesco Bonchi. 2022. The effect of people recommenders on echo chambers and polarization. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16, 90–101.

[12] Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. 2020. Multi-head attention: Collaborate instead of concatenate. arXiv:2006.16362. Retrieved from https://arxiv.org/abs/2006.16362

[13] Jesse Davis and Mark Goadrich. 2006. The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning*, 233–240.

[14] Sihao Ding, Peng Wu, Fuli Feng, Yitong Wang, Xiangnan He, Yong Liao, and Yongdong Zhang. 2022. Addressing unmeasured confounder for recommendation with sensitivity analysis. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 305–315.

[15] Sergey N. Dorogovtsev, José Fernando F. Mendes, and Alexander N. Samukhin. 2001. Giant strongly connected component of directed networks. *Physical Review E* 64, 2 (2001), 025101.

[16] Elizabeth Dubois and Grant Blank. 2018. The echo chamber is overstated: The moderating effect of political interest and diverse media. *Information, Communication & Society* 21, 5 (2018), 729–745.

[17] Xiao Fang, Paul Jen-Hwa Hu, Zhepeng Li, and Weiyu Tsai. 2013. Predicting adoption probabilities in social networks. *Information Systems Research* 24, 1 (2013), 128–145.

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

[19] R. Kelly Garrett. 2009. Echo chambers online? Politically motivated selective exposure among Internet news users. *Journal of Computer-Mediated Communication* 14, 2 (2009), 265–285.

[20] Yingqiang Ge, Shuya Zhao, Honglu Zhou, Changhua Pei, Fei Sun, Wenwu Ou, and Yongfeng Zhang. 2020. Understanding echo chambers in e-commerce recommender systems. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2261–2270.

[21] Andrew Gelman and Christian Hennig. 2017. Beyond subjective and objective in statistics. *Journal of the Royal Statistical Society Series A: Statistics in Society* 180, 4 (2017), 967–1033.

[22] Nabeel Gillani, Ann Yuan, Martin Saveski, Soroush Vosoughi, and Deb Roy. 2018. Me, my echo chamber, and I: Introspection on social media polarization. In *Proceedings of the 2018 World Wide Web Conference*, 823–831.

[23] Swati Goswami, C. A. Murthy, and Asit K. Das. 2018. Sparsity measure of a network graph: Gini index. *Information Sciences* 462 (2018), 16–39.

[24] Jesse Harte, Wouter Zorgdrager, Louridas Panos, Katsifodimos Asterios, Jannach Dietmar, and Fragkoulis Marios. 2023. Leveraging large language models for sequential recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*, 1096–1102.

[25] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th International Conference on World Wide Web*, 173–182.

[26] Xiangnan He, Yang Zhang, Fuli Feng, Chonggang Song, Lingling Yi, Guohui Ling, and Yongdong Zhang. 2023. Addressing confounding feature issue for causal recommendation. *ACM Transactions on Information Systems* 41, 3 (2023), 1–23.

[27] B. Hidasi. 2015. Session-based recommendations with recurrent neural networks. arXiv:1511.06939. Retrieved from https://arxiv.org/abs/1511.06939

[28] Timothy O. Hodson. 2022. Root mean square error (RMSE) or mean absolute error (MAE): When to use them or not. *Geoscientific Model Development Discussions* 2022 (2022), 1–10.

[29] Neil J. Hurley. 2013. Personalised ranking with diversity. In *Proceedings of the 7th ACM Conference on Recommender Systems*, 379–382.

[30] Youngseung Jeon, Bogoan Kim, Aiping Xiong, Dongwon Lee, and Kyungsik Han. 2021. Chamberbreaker: Mitigating the echo chamber effect and supporting information hygiene through a gamified inoculation system. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–26.

[31] Mathias Jesse, Christine Bauer, and Dietmar Jannach. 2023. Intra-list similarity and human diversity perceptions of recommendations: The details matter. *User Modeling and User-Adapted Interaction* 33, 4 (2023), 769–802.

[32] Pascal Jürgens and Birgit Stark. 2022. Mapping exposure diversity: The divergent effects of algorithmic curation on news consumption. *Journal of Communication* 72, 3 (2022), 322–344.

[33] Bo-Ra Kim, Jiwon Shin, Robin B. Guevarra, Jun Hyung Lee, Doo Wan Kim, Kuk-Hwan Seol, Ju-Hoon Lee, Hyeun Bum Kim, and Richard E. Isaacson. 2017. Deciphering diversity indices for a better understanding of microbial communities. *Journal of Microbiology and Biotechnology* 27, 12 (2017), 2089–2093.

[34] Diederik P. Kingma. 2014. Adam: A method for stochastic optimization. arXiv:1412.6980. Retrieved from https://arxiv.org/abs/1412.6980

[35] Haoxuan Li, Kunhan Wu, Chunyuan Zheng, Yanghao Xiao, Hao Wang, Zhi Geng, Fuli Feng, Xiangnan He, and Peng Wu. 2024. Removing hidden confounding in recommendation: A unified multi-task learning approach. *Advances in Neural Information Processing Systems*, Vol. 36.

[36] Jing Li, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Tao Lian, and Jun Ma. 2017. Neural attentive session-based recommendation. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 1419–1428.

[37] Jiacheng Li, Ming Wang, Jin Li, Jinmiao Fu, Xin Shen, Jingbo Shang, and Julian McAuley. 2023. Text is all you need: Learning language representations for sequential recommendation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1258–1267.

[38] Jing Li, Botong Wu, Xinwei Sun, and Yizhou Wang. 2021. Causal hidden Markov model for time series disease forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12105–12114.

[39] Qian Li, Xiangmeng Wang, Zhichao Wang, and Guandong Xu. 2023. Be causal: De-biasing social network confounding in recommendation. *ACM Transactions on Knowledge Discovery from Data* 17, 1 (2023), 1–23.

[40] Qi Liu, Zheng Dong, Chuanren Liu, Xing Xie, Enhong Chen, and Hui Xiong. 2014. Social marketing meets targeted customers: A typical user selection and coverage perspective. In *Proceedings of the 2014 IEEE International Conference on Data Mining*. IEEE, 350–359.

[41] Qiao Liu, Yifu Zeng, Refuoe Mokhosi, and Haibin Zhang. 2018. STAMP: Short-term attention/memory priority model for session-based recommendation. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1831–1839.

[42] Siyu Liu, Xuehua Song, Zhongchen Ma, Ernest Domanaanmwi Ganaa, and XiangJun Shen. 2022. MoRE: Multi-output residual embedding for multi-label classification. *Pattern Recognition* 126 (2022), 108584.

[43] Vinicius M. Marques, Celso J. Munaro, and Sirish L. Shah. 2015. Detection of causal relationships based on residual analysis. *IEEE Transactions on Automation Science and Engineering* 12, 4 (2015), 1525–1534.

[44] Antonis Matakos, Evimaria Terzi, and Panayiotis Tsaparas. 2017. Measuring and moderating opinion polarization in social networks. *Data Mining and Knowledge Discovery* 31 (2017), 1480–1505.

[45] Serena McDonnell, Omar Nada, Muhammad Rizwan Abid, and Ehsan Amjadian. 2021. CyberBERT: A deep dynamic-state session-based recommender system for cyber threat recognition. In *Proceedings of the 2021 IEEE Aerospace Conference (50100)*. IEEE, 1–12.

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

[47] Sean Munson, Stephanie Lee, and Paul Resnick. 2013. Encouraging reading of diverse political viewpoints with a browser widget. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 7, 419–428.

[48] Mark E. J. Newman. 2006. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences* 103, 23 (2006), 8577–8582.

[49] Tien T. Nguyen, Pik-Mai Hui, F. Maxwell Harper, Loren Terveen, and Joseph A. Konstan. 2014. Exploring the filter bubble: The effect of using recommender systems on content diversity. In *Proceedings of the 23rd International Conference on World Wide Web*, 677–686.

[50] Judea Pearl. 2010. An introduction to causal inference. *The International Journal of Biostatistics* 6, 2 (2010), 7.

[51] Mattia Prosperi, Yi Guo, Matt Sperrin, James S. Koopman, Jae S. Min, Xing He, Shannan Rich, Mo Wang, Iain E. Buchan, and Jiang Bian. 2020. Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nature Machine Intelligence* 2, 7 (2020), 369–375.

[52] Shameem A. Puthiya Parambath, Nicolas Usunier, and Yves Grandvalet. 2016. A coverage-based approach to recommendation diversity on similarity graph. In *Proceedings of the 10th ACM Conference on Recommender Systems*, 15–22.

[53] Ruiyang Ren, Zhaoyang Liu, Yaliang Li, Wayne Xin Zhao, Hui Wang, Bolin Ding, and Ji-Rong Wen. 2020. Sequential recommendation with self-attentive multi-adversarial network. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 89–98.

[54] Steffen Rendle. 2010. Factorization machines. In *Proceedings of the 2010 IEEE International Conference on Data Mining*. IEEE, 995–1000.

[55] Selim Reza, Marta Campos Ferreira, José J. M. Machado, and João Manuel R. S. Tavares. 2022. A multi-head attention-based transformer model for traffic flow forecasting with a comparative analysis to recurrent neural networks. *Expert Systems with Applications* 202 (2022), 117275.

[56] Shouq A. Sadah, Moloud Shahbazi, Matthew T. Wiley, and Vagelis Hristidis. 2015. A study of the demographics of web-based health-related social media users. *Journal of Medical Internet Research* 17, 8 (2015), e194.

[57] Jari Saramäki, Mikko Kivelä, Jukka-Pekka Onnela, Kimmo Kaski, and Janos Kertesz. 2007. Generalizations of the clustering coefficient to weighted complex networks. *Physical Review E, Statistical, Nonlinear, and Soft Matter Physics* 75, 2 (2007), 027105.

[58] Kazutoshi Sasahara, Wen Chen, Hao Peng, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. 2021. Social influence and unfollowing accelerate the emergence of echo chambers. *Journal of Computational Social Science* 4, 1 (2021), 381–402.

[59] Ketan Rajshekhar Shahapure and Charles Nicholas. 2020. Cluster quality analysis using silhouette score. In *Proceedings of the 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 747–748.

[60] Jesse Shore, Jiye Baek, and Chrysanthos Dellarocas. 2018. Network structure and patterns of information diversity on twitter. *MIS Quarterly* 42, 3 (2018), 849.

[61] P. J. Somerfield, K. R. Clarke, and R. M. Warwick. 2008. Simpson index. In *Encyclopedia of Ecology*. Elsevier, 3252–3255.

[62] Ludovic Terren and Rosa Borge-Bravo. 2021. Echo chambers on social media: A systematic review of the literature. *Review of Communication Research* 9 (2021), 1–39.

[63] Antonela Tommasel, Juan Manuel Rodriguez, and Daniela Godoy. 2021. I want to break free! Recommending friends from outside the echo chamber. In *Proceedings of the 15th ACM Conference on Recommender Systems*, 23–33.

[64] Michael Trusov, Anand V. Bodapati, and Randolph E. Bucklin. 2010. Determining influential users in internet social networks. *Journal of Marketing Research* 47, 4 (2010), 643–658.

[65] Mirthe Maria Van Diepen, Ioan Gabriel Bucur, Tom Heskes, and Tom Claassen. 2023. Beyond the Markov equivalence class: Extending causal discovery under latent confounding. In *Proceedings of the Conference on Causal Learning and Reasoning*. PMLR, 707–725.

[66] Antoine Vendeville, Anastasios Giovanidis, Effrosyni Papanastasiou, and Benjamin Guedj. 2022. Opening up echo chambers via optimal content recommendation. In *Proceedings of the International Conference on Complex Networks and Their Applications*. Springer, 74–85.

[67] Liwei Wu, Shuqing Li, Cho-Jui Hsieh, and James Sharpnack. 2020. SSE-PT: Sequential recommendation via personalized transformer. In *Proceedings of the 14th ACM Conference on Recommender Systems*, 328–337.

[68] Le Wu, Qi Liu, Enhong Chen, Nicholas Jing Yuan, Guangming Guo, and Xing Xie. 2016. Relevance meets coverage: A unified framework to generate diversified recommendations. *ACM Transactions on Intelligent Systems and Technology* 7, 3 (2016), 1–30.

[69] Libing Wu, Cong Quan, Chenliang Li, Qian Wang, Bolong Zheng, and Xiangyang Luo. 2019. A context-aware user-item representation learning for item recommendation. *ACM Transactions on Information Systems (TOIS)* 37, 2 (2019), 1–29.

[70] Yuxia Wu, Guoshuai Zhao, Mingdi Li, Zhuocheng Zhang, and Xueming Qian. 2023. Reason generation for point of interest recommendation via a hierarchical attention-based transformer model. *IEEE Transactions on Multimedia* 26 (2023), 5511–5522.

[71] Guandong Xu, Tri Dung Duong, Qian Li, Shaowu Liu, and Xianzhi Wang. 2020. Causality learning: A new perspective for interpretable machine learning. arXiv:2006.16789. Retrieved from https://arxiv.org/abs/2006.16789

[72] Jin Xu, Zishan Li, Bowen Du, Miaomiao Zhang, and Jing Liu. 2020. Reluplex made more practical: Leaky ReLU. In *Proceedings of the 2020 IEEE Symposium on Computers and Communications (ISCC)*. IEEE, 1–7.

[73] Jingjing Xu, Xu Sun, Zhiyuan Zhang, Guangxiang Zhao, and Junyang Lin. 2019. Understanding and improving layer normalization. *Advances in Neural Information Processing Systems*, Vol. 32.

[74] Shuyuan Xu, Juntao Tan, Zuohui Fu, Jianchao Ji, Shelby Heinecke, and Yongfeng Zhang. 2022. Dynamic causal collaborative filtering. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2301–2310.

[75] Wenjian Xu and Chi-Yin Chow. 2015. A location-and diversity-aware news feed system for mobile users. *IEEE Transactions on Services Computing* 9, 6 (2015), 846–861.

[76] Yonghui Yang, Le Wu, Zihan Wang, Zhuangzhuang He, Richang Hong, and Meng Wang. 2024. Graph bottlenecked social recommendation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 3853–3862.

[77] Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. 2021. A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data* 15, 5 (2021), 1–46.

[78] Dianer Yu, Qian Li, Hongzhi Yin, and Guandong Xu. 2023. Causality-guided graph learning for session-based recommendation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 3083–3093.

[79] Erheng Zhong, Wei Fan, and Qiang Yang. 2014. User behavior learning and transfer in composite social networks. *ACM Transactions on Knowledge Discovery from Data* 8, 1 (2014), 1–32.

[80] Jianghong Zhou, Eugene Agichtein, and Surya Kallumadi. 2020. Diversifying multi-aspect search results using Simpson's diversity index. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2345–2348.

[81] Jianming Zhu, Peikun Ni, Guangmo Tong, Guoqing Wang, and Jun Huang. 2021. Influence maximization problem with echo chamber effect in social network. *IEEE Transactions on Computational Social Systems* 8, 5 (2021), 1163–1171.

[82] Cai-Nicolas Ziegler, Sean M. McNee, Joseph A. Konstan, and Georg Lausen. 2005. Improving recommendation lists through topic diversification. In *Proceedings of the 14th International Conference on World Wide Web*, 22–32.

Received 20 December 2024; revised 5 April 2025; accepted 24 July 2025