# Nudging Towards Responsible Recommendations: a Graph-Based Approach to Mitigate Belief Filter Bubbles

Mengyan Wang, Yuxuan Hu, Shiqing Wu, Weihua Li, Quan Bai, Zihan Yuan, and Chenting Jiang

***Abstract*—Personalized recommendation systems homogenize user preferences, causing an extreme belief imbalance and aggravating user bias. This phenomenon is known as the filter bubble. This paper presents the Responsible Graph-based Recommendation (RGRec) system, designed to alleviate the filter bubble effect in personalized recommendation systems. Acting as an intermediate agency between users and existing preference-based recommendation systems, RGRec is composed of three collaborative modules: the Multi-faceted Reasoning-based Filter Bubbles Detection module (FBDetect), the Belief Nudging module, and the Generative Artificial Intelligence-based Recommendation Strategy Generation module (RecomGen). The FBDetect module identifies users with extreme belief imbalances based on their belief networks, which are represented as heterogeneous graphs. These graphs are then utilized in the Belief Nudging module, where a nudging strategy is employed to adapt prompts for the RecomGen module. Ultimately, the RecomGen module generates contextually rich items for recommendations. Experimental results demonstrate that RGRec can promote diverse content exploration based on user feedback and progressively stimulate interest in topics users initially showed less interest in, encouraging individual exploration.**

***Impact Statement*—This paper introduces an innovative approach to personalized recommendations, termed Responsible Graph-based Recommendation (RGRec). RGRec addresses the filter bubble issue associated with recommendation systems by integrating Generative Artificial Intelligence (GAI). Notably, RGRec excels in producing contextually rich recommendations through graph-based nudging strategies. By considering user preferences, it actively encourages diverse content exploration. This responsible methodology aligns with principles of libertarian paternalism and transparency, distinguishing itself from conventional recommendation systems. The proposed approach represents a noteworthy advancement in responsible AI and recommendation systems, challenging established norms by dismantling filter bubbles, fostering information diversity, and upholding user autonomy.**

***Index Terms*—Belief harmony, Filter bubble, Nudge theory, Responsible recommendation systems**

Mengyan Wang, Auckland University of Technology, Auckland, New Zealand (e-mail: cjv2124@autuni.ac.nz).

Yuxuan Hu, University of Tasmania, Hobart, Australia (yuxuan.hu@utas.edu.au).

Shiqing Wu, University of Technology Sydney, Sydney, Australia (shiqing.wu@uts.edu.au).

Corresponding author. Weihua Li, Auckland University of Technology, Auckland, New Zealand (e-mail: weihua.li@aut.ac.nz).

Quan Bai, University of Tasmania, Hobart, Australia (quan.bai@utas.edu.au).

Zihan Yuan, University of Tasmania, Hobart, Australia (zyuan0@utas.edu.au).

Chenting Jiang, University of Tasmania, Hobart, Australia (chenting.jiang@utas.edu.au).

## I. Introduction

RECOMMENDATION systems play an important role in shaping users' access to diverse information on the Internet [1]. However, most of the existing preference-based recommendation systems continuously suggest items that are similar to users' previous experiences, leading them to a homogeneous information environment. This is known as the "filter bubble" phenomenon [2]. Prolonged exposure to such filter bubbles can result in the development of extreme and imbalanced beliefs, hindering the formation of a comprehensive understanding and amplifying ideological biases [3]. In the context of responsible artificial intelligence (AI), it becomes crucial to develop AI recommendation systems that provide a diversity of content and viewpoints, rather than reinforcing existing user preferences, to mitigate filter bubbles [4].

Existing research works on mitigating the filter bubble in recommendation systems can be grouped into two main strategies: algorithm-focused and human-focused approaches [5]. Algorithm-focused strategies advocate for promoting content diversity at both the in-processing and post-processing stages [5]. These methods leverage diverse techniques, including explanation-based diversity recommendation [6], community-aware models [7], category-based diversification algorithms [8], the Diversified GNN-based Recommendation system (DGRec) [9], and graph-based user-item interaction methods [10]. Notably, graph-based approaches, relying on user preferences and category diversity insights, aim to enhance recommendation quality [11]. However, these strategies, particularly those driven by algorithms, may unintentionally ignore the essential role of human decision-making processes. In contrast, human-focused strategies emphasize individuals [5]. Techniques such as nudging-based recommendations [12], [13] indirectly influence user decisions and behaviors. Despite focusing on users, these models often encounter challenges in effectively broadening users' interests beyond their preferred topics. Unlike existing mitigation approaches, this paper integrates the strengths of both algorithmic and human-focused strategies and presents the Responsible Graph-based Recommendation framework, namely RGRec, to mitigate filter bubbles by mildly moderating users' extreme beliefs, thus exposing them to a more diverse scope of information.

RGRec is a graph-based approach that serves as an intermediary between recommendation systems and users. Its primary objective is to effectively address the filter bubbles issues by bridging the gap between user preferences and the delivery

of diverse content recommendations. The system comprises three key modules: the Multi-faceted Reasoning-based Filter Bubbles Detection module (FBDetect), the Belief Nudging module, and the Generative Artificial Intelligence-based Recommendation Strategy Generation module (RecomGen). In FBDetect, a user's belief is represented as a heterogeneous graph [14] known as a belief network. FBDetect identifies users affected by filter bubbles by evaluating the balance between a user's belief toward a specific topic of information and the recommendations received from the system. If this balance is significantly skewed, the user is flagged as being impacted by filter bubbles. The Belief Nudging module collaborates with users' belief networks to explore paths between topics that users favor and those they display less interest. These explored paths serve as prompts for RecomGen to generate items for a nudging recommendation strategy. This strategy aims to gently introduce users to content they may have shown less preference for, fostering a more balanced exposure to diverse content. The collaboration among these three modules is iterative, continuously optimizing and adjusting to mitigate filter bubbles.

The ultimate goal of this research is to gradually introduce users to a more diverse range of content, fostering belief harmony and enhancing the overall user experience. RGRec's innovative approach addresses the challenges of filter bubbles by proactively diversifying recommendations and promoting a more nuanced interaction between users and content.

- Firstly, we introduce a novel responsible approach, i.e., RGRec, designed to address the moderation of users' extreme beliefs and the mitigation of filter bubbles resulting from conventional recommendation approaches. To the best of our knowledge, RGRec stands out as one of the pioneering responsible recommendation methods explicitly focused on alleviating filter bubbles.
- Secondly, we present the Multi-faceted Reasoning-based Filter Bubbles Detection module (FBDetect), a pivotal component within RGRec. FBDetect identifies users affected by filter bubbles and scrutinizes recommendation systems relying solely on user preferences. Our approach employs diverse methodologies to comprehensively analyse filter bubbles, examining their existence and effects from various perspectives.
- Thirdly, we leverage the efficacy of nudging techniques to guide users in broadening their interests and promoting belief harmony. Our nudging process aligns with principles of libertarian paternalism, transparency, and democracy, thereby enhancing users' understanding of recommendations.
- Finally, we present the Generative Artificial Intelligence-based Recommendation Strategy Generation module (RecomGen) for crafting recommendation strategies aimed at mitigating filter bubbles. This method leverages advanced graph-based techniques to learn and analyze user beliefs, systematically exploring potential paths to alleviate users' extreme beliefs by introducing a more extensive range of information and enhancing content diversity.

The rest of this paper is organized as follows. Section II offers a comprehensive review of related literature, delving into recommendation systems, user belief bias, filter bubbles, and nudging techniques. Section III elucidates essential definitions, notations, and concepts integral to our discourse. The methodology and framework underpinning our research are expounded upon in Section IV. Section V is dedicated to the elucidation of our experimental setup, with subsequent presentation and analysis of results. In Section VI, the findings are examined and discussed. Section VII concludes the paper and outlines potential directions for future research.

## II. Related Works

Personalized recommendation systems have been criticized as inadvertently creating filter bubbles [15], which constrain users' exposure to various perspectives and information, thus potentially leading to belief biases and societal fragmentation [8]. To mitigate this concern, many researchers and practitioners have focused on dismantling filter bubbles, fostering diversity and democracy in recommendation systems, and facilitating users' belief harmony.

In this section, we review the relevant research works, deliberate on the filter bubble issue, explore the diversification of recommendation systems, and examine the prior research on nudge recommendations. Additionally, we will highlight the contributions of this study.

### A. Filter Bubbles

*1) Preference-based Recommendation Systems:* Conventional recommendation systems prioritize the generalization of user preference, implying that these systems often recommend items to users based on their specific preferences and behaviors [16]. Techniques such as Collaborative Filtering (CF) [17], Content-Based filtering (CB) [18], rule-based methods [19], or hybrid models [20] are commonly employed to analyze users' preferences and past behaviors. The recommendation system then suggests content that aligns closely with user preferences to enhance user satisfaction and engagement. However, this approach based on user preference may exacerbate the filter bubble issue, leading to ideological isolation and user bias. For example, Bryant et al. demonstrate that the YouTube algorithm, representative of a preference recommendation algorithm, exhibits a marked bias towards right-leaning political videos, including those espousing racist views propagated by the alt-right community [21]. Thus, it is important to address the limitations of current preference recommendations, boost the diversity of suggestions, and harmonize users' beliefs.

*2) Mitigating Filter Bubble Effects:* Filter bubbles emphasize the constraints of preference recommendation algorithms [22]. Dahlgren introduced the term "internet filters" to represent the phenomenon of filter bubbles, which can have various negative effects on users, including a narrowed focus on personal interests, substantial reinforcement of confirmation bias, reduced curiosity, decreased exposure to diverse ideas and people, compromised understanding of the world, and a skewed perception of reality [23].

Addressing the negative effects of filter bubbles proves to be challenging, especially when considering the notable aspect

of algorithmic bias. Chen et al. argue that the emergence of recommendation algorithm bias amplifies the experimental nature of user behavior data as opposed to observational [24]. Additionally, Dahlgren examines the recommendation algorithm bias and broadens the concept of bias into two facets, one originating from the recommendation algorithm and the other from users' behaviors [23]. Aside from algorithmic bias, another challenge in mitigating filter bubbles lies in their elusive nature [8]. Users often find themselves unaware of the filter bubble effect, which creates a homogenized view of the world. Specifically, they may not realize that their perspective differs from others in similar circumstances.

The growing influence of filter bubbles has raised increased concerns among researchers. A well-crafted recommendation system usually offers high accuracy while promoting diversity; systems oriented solely towards accuracy may inevitably lead to filter bubble effects [9]. Contemporary research proposes several strategies for breaking filter bubbles by enhancing the diversity of recommendations. The research addressing filter bubbles with graphs is reviewed as follows.

**Research in the scope of the graph.** While the above-mentioned research has significantly contributed to alleviating filter bubbles and enhancing recommendation diversity, many researchers also advocate for the critical role of graph-based recommendation algorithms. These algorithms mitigate data sparsity and cold start issues and add an essential interpretability factor to recommendation systems [11]. Yang et al. introduce the Diversified GNN-based Recommendation System (DGRec), a graph-based recommendation system built on GNN, augmenting the diversity of recommended lists by improving the embedding generation process [9]. Tang et al. propose a temporal graph-based method to learn user evolving preferences in dynamic recommendation scenarios [25]. Additionally, Li et al. adopt a graph-based methodology by constructing a user-item interaction graph for data analysis to examine the existence of a centralized recommendation phenomenon [10].

In contrast, our model surpasses traditional methods by generating more diverse items instead of marginally varied ones, based on user preferences for diversity. We prioritize incrementally stimulating users' interest in items they may initially disregard without altering existing recommendation system algorithms. The proposed novel approach aims to counteract the filter bubble effect by considering user interest and disinterest beliefs, i.e., an aspect that has received minimal attention from researchers.

**Detection of Belief Bias.** Belief bias in reasoning refers to individuals' tendency to favor conclusions that align with their pre-existing beliefs [26]. This phenomenon is connected with the formation of online filter bubbles, in which users tend to accept information that confirms their viewpoints and interests while rejecting alternative perspectives that challenge their beliefs [27], [28].

Existing methods proposed for belief bias detection include Information Source Diversity Analysis (ISDA) [8], User Interaction Pattern Analysis (UIPA) [29], Reinforcement Learning Methods (RLM) [10], and Social Network Analysis (SNA) [29], [30]. Considering the limited interpretability of RLM and the focus of SNA on alleviating echo-chamber effects rather than filter bubbles, our research concentrates on investigating selective belief bias algorithms based on ISDA and UIPA. ISDA includes various detection metrics such as topology metrics and homophily metrics [8]. Likewise, UIPA includes several established detection metrics, including the coverage algorithm and the Majority Category Domination (MCD) algorithm [29]. Drawing inspiration from these metrics, we propose the FBDetect model for dual verification of the authenticity of the filter bubble phenomenon, having the concept of "Entropy" [31] included to substantiate the existence of filter bubbles.

### *B. Nudge Techniques and Responsible Recommendations*

A *nudge* is a non-coercive intervention designed to influence behavior by modifying the context in which choices are made [32]. Such an intervention is usually transparent, optional, and responsible, enabling individuals to understand their choice consequences better and boost the likelihood of beneficial decision-making [33]. The core idea behind a nudge is to exploit individuals' beliefs and behavioral biases through various design strategies, such as providing incentives [34], [35] and utilizing social influence [36], [37], directing them towards more favorable outcomes without constraining their freedom of choice [32], [38].

Recent research in recommendation systems has begun to explore the role of nudges. However, the majority introduces nudging recommendations from an AI-deprived perspective, implying a substantial absence or lack of AI technology in their research context. For example, Jesse et al. consolidate 87 nudging mechanisms at this AI-deprived level, including alterations in font size, the reputation of the messenger, and the visibility of information [12]. Joachim et al. propose a platform empowered by AI designed to nudge, influence, and guide the behavior of individuals with diabetes [13]. Furthermore, Sitar et al. propose an automated recommendation system. This system integrates managers' priorities and user feedback and utilizes graph structures to organize items based on descending order of priority, known as nudge concepts [39].

The recommendation systems mentioned previously have revealed the importance of establishing a responsible, graph-based nudging recommendation system. However, existing models are developed solely on user preferences, neglecting the influence of filter bubbles. Unlike the existing preference-based approaches, this research aims to gently present more potential interests to users whom they may have yet to be genuinely interested in initially. Guiding user perceptions from one end of the graph (items users are highly interested in) to the other end (items users are less interested in), moderating user extreme beliefs, reducing user bias, and breaking the filter bubble effect, thereby allowing users to access a more diverse range of information.

## III. Preliminaries

In this section, we introduce definitions, notations, and concepts used in this paper. Key notations are listed in Table I.

TABLE I
TABLE OF NOTATIONS

| Notation | Description |
|---|---|
| $S$ | The recommendation environment. |
| $U$ | A set of users. |
| $A$ | AI-based algorithm for generating recommendations. |
| $C$ | A set of predefined topics. |
| $u_i$ | A user $u_i$ in the system. |
| $c_x$ | A topic $c_x$ |
| $C_x^{sub}$ | A set of aspects associated with a topic $c_x$. |
| $c_k^x$ | An aspect associated with $c_x$. |
| $I_k^x$ | A set of items whose aspect is $c_k^x$. |
| $i_m$ | An item. |
| $G_i$ | Belief network of user $u_i$. |
| $V_i$ | A set of nodes in $G_i$. |
| $\hat{C}_i$ | A set of topics that $u_i$ interacted with. |
| $\hat{C}_i^{sub}$ | A set of aspects that $u_i$ interacted with. |
| $E_i$ | A set of edges in $G_i$. |
| $E_i^b$ | A set of edges connecting from $u_i$ to topics. |
| $E_i^c$ | A set of edges connecting from topics to aspects. |
| $e_{ix}$ | An edge from $u_i$ to $c_x$. |
| $e_{xk}$ | An edge from $c_x$ to $c_k^x$. |
| $b_{ix}$ | Belief degree associated on edge $e_{ix}$. |
| $r_{ik}^x$ | Click probability associated on edge $e_{xk}$. |
| $\rho(c_x, c_y)$ | Topic similarity measure between two topics. |
| $p_{i,t}$ | Recommendation prompt path for user $u_i$ at time step $t$. |
| $p_{i,t}(k)$ | The $k^{th}$ node in the recommendation prompt path $p_{i,t}$. |
| $feed$ | A recommendation list comprising $feed_{original}$ and $GI$. |
| $C_{feed}$ | A subset of $C$ that includes only those topics appeared in the feed. |
| $feed_{original}$ | A list of items suggested by the preference-based recommendation system. |
| $GI$ | Contextually rich items generated by RGRec, based on an explored recommendation prompt path. |
| $\check{M}_{i,t}$ | A set of items within a $feed$ accepted by user $u_i$ from the initial time step to time step $t$. |
| $\bar{P}_{i,t}$ | A sequence of recommendation prompt paths shown to $u_i$, where resulting items have been declined over the same period. |

A recommendation environment is defined as $S = (U, A, C)$, where $U = \{u_1, ..., u_n\}$ represents a set of users, $A$ refers to an AI-based algorithm for recommending items to users, and $C = \{c_1, ..., c_x\}$ signifies a set of pre-defined topics. Meanwhile, each $c_x$ is associated with a set of aspects $C_x^{sub} = \{c_1^x, ..., c_k^x\}$, and each $c_k^x$ is associated with an item set $I_k^x = \{i_1, ..., i_m\}$. An item can represent a news article or a movie description in real-world applications. To simplify the problem, in this paper, we assume each item belongs to only one aspect, and each aspect is related to a single topic.

**Definition 1.** ***The belief network*** *of a user $u_i$ is represented as a directed graph $G_i = (V_i, E_i)$. Specifically, $V_i = \{u_i\} \cup \hat{C}_i \cup \hat{C}_i^{sub}$ represents a set consisting of three distinct types of nodes, where $\hat{C}_i \subset C$ denotes a set of* ***topics*** *that $u_i$ has engaged with, and $\hat{C}_i^{sub} = \{c_k^x | c_k^x \in C_x^{sub}, c_x \in \hat{C}_i\}$ represents different* ***aspects*** *related to $\hat{C}_i$. Meanwhile, $E_i = E_i^b \cup E_i^c$ represents a composite set of edges, where $E_i^b = \{e_{ix} | c_x \in \hat{C}_i\}$ comprises edges connecting from the user to each topic, and $E_i^c = \{e_{xk} | c_x \in \hat{C}_i, c_k^x \in \hat{C}_i^{sub} \cap C_x^{sub}\}$ consists of edges connecting from each topic to each of its aspects. In $G_i$, $e_{ix}$ denotes $u_i$ prefers a topic $c_x$, and the weight $b_{ix}$ associated on $e_{ix}$ represents the extent of $u_i$'s belief towards a topic $c_x$. While $e_{xk}$ indicates the affiliation relationship between each pair of topic and topic aspect, and its weight $r_{ik}^x$ reflects the probability that $u_i$ selects an item whose aspect is $c_k^x \in \hat{C}_i^{sub}$.*

The click probability $r_{ik}^x$ is calculated using Equation 1, where $|\hat{I}_k^x|$ denotes the number of $u_i$ interacted items whose aspects are $c_k^x$. In this case, given a belief graph $G_i$, the sum of all $r_{ik}^x$ is 1.

$$r_{ik}^x = \frac{|\hat{I}_k^x|}{\sum_{c_{x'} \in \hat{C}_i} \sum_{c_{k'}^{x'} \in \hat{C}_{x'}^{sub}} |\hat{I}_{k'}^{x'}|} \tag{1}$$

After calculating all click probability $r_{ik}^x$, the belief degree $b_{ix}$ of user $u_i$ towards a topic $c_x$ can be formulated as follows:

$$b_{ix} = - \sum_{c_k^x \in \hat{C}_i^{sub} \cap C_x^{sub}} r_{ix}^k \log_2(r_{ix}^k), \tag{2}$$

**Definition 2.** ***Topic similarity*** *$\rho(c_x, c_y)$ represents the similarity between two topics $c_x$ and $c_y$, with symmetry $\rho(c_x, c_y) = \rho(c_y, c_x)$. The similarity measure $\rho$ is defined within the range $[-1, 1]$, where a higher value of $\rho$ indicates greater similarity between the topics.*

As each item is only associated with one topic, we can obtain topic embedding by aggregating related item embeddings. In this paper, we obtain the topic embedding $\mathbf{c}_x$ by adopting the Hadamard product [40] to combine all corresponding item embeddings. The similarity between topic $c_x$ and $c_y$ is subsequently calculated by the cosine similarity [41]:

$$\rho(c_x, c_y) = \frac{\mathbf{c}_x \cdot \mathbf{c}_y}{\|\mathbf{c}_x\| \, \|\mathbf{c}_y\|}, \tag{3}$$

where $\mathbf{c}_x$ and $\mathbf{c}_y$ denote topic embeddings, $\mathbf{c}_x \cdot \mathbf{c}_y$ refers to the dot product of the topic embeddings, and $\|\mathbf{c}_x\|$ and $\|\mathbf{c}_y\|$ denote the corresponding Euclidean norms.

The primary emphasis of this research does not center around language embedding. In this paper, our approach involves the utilization of a pre-trained language model [42] for the purpose of item embedding.

**Definition 3.** ***A recommendation prompt path*** *$p_{i,t}$ is a sequence of topics explored by a filter bubble-affected user $u_i$ at a specific time step $t$, bridging the gap between topics $c_x^{SP_i}$ ($u_i$ strongly preferred) and $c_y^{LP_i}$ ($u_i$ less preferred) by introducing additional interacted topics. $p_{i,t}(k)$ refers to the $k^{th}$ topic in a recommendation prompt path $p_{i,t}$. Furthermore, the topics within $p_{i,t}$ can be the keywords of prompts for the RecomGen module to generate contextually rich items for recommendation to $u_i$ (see Definition 4).*

**Definition 4.** ***A recommendation list*** *$feed = \{feed_{original}, GI\}$ is a compilation of items recommended to users, where $feed_{original}$ is comprised of items suggested by the existing preference-based recommendation system, and $GI$ includes contextually rich items generated by RGRec, based on an explored recommendation prompt path (refer to Definition 3). Given the size of $feed$, we introduce a weight parameter $w$ to control the proportion of $GI$ within the $feed$, balancing the mix of original and RGRec recommendations. $\check{M}_{i,t}$ signifies the set of items in a $feed$ accepted by user $u_i$ from the initial time step to time step $t$. In contrast, $\bar{P}_{i,t}$ refers*

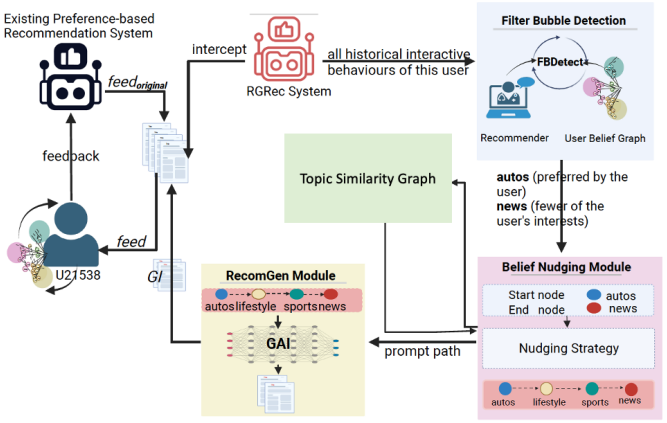



Fig. 1. Overall working process of RGRec for user "U21538".

*to the sequence of recommendation prompt paths shown to $u_i$, where the resulting items have been declined by the user over the same period.*

Incorporating $GI$ in a $feed$ bridges the gap between filter bubble-affected users' preferred and less preferred topics. It introduces more interconnected topic-related items, moderates extreme beliefs, and encourages belief harmony.

## IV. The Framework of Responsible Graph-based Recommendation

Responsible Graph-based Recommendation (RGRec) is designed to guide users gently from a state of information imbalance to one of belief harmony. RGRec stands out for its incorporation of the "nudge" concept and the utilization of GAI to produce contextually rich items. This approach encourages users to explore interests in topics they may have originally shown less interest in, providing diverse options. Figure 1 illustrates the overview of RGRec's working process, using a practical example involving the user "U21538".

As shown in Fig. 1, RGRec operates as a dynamic and interactive mediator between existing preference-based recommendation systems and the users, e.g., "U21538" in this example. It conforms to the principle of non-coercion, ensuring that the user experiences a gradual transition towards belief harmony. RGRec comprises three key modules, each integral to the elimination of filter bubbles. Firstly, the **Multi-faceted Reasoning-based Filter Bubbles Detection module (FBDetect)** plays a pivotal role in our system by comprehensively detecting filter bubbles. Consider the example in Fig. 1, where user "U21538" is identified by FBDetect as a filter bubble-affected user with extremely imbalanced beliefs. Specifically, FBDetect recognizes that this user highly favors "autos" while displaying minimal interest in "news". These precise insights about users with extremely imbalanced beliefs enable RGRec to implement targeted interventions to counteract the effects of filter bubbles effectively. Subsequently, the FBDetect module transfers these findings to the Belief Nudging module, which further contributes to achieving user belief harmony. The **Belief Nudging module** forms the core of RGRec. Its main task is to use the nudging strategy and topic similarities to provide a recommendation prompt path for the next module, RecomGen. As depicted in Figure 1, the module takes in two key inputs from the FBDetect: the user's preferred "autos" and the less preferred "news". The nudging strategy combines topic similarities to further generate the prompt path based on these two topics, which is "autos $\rightarrow$ lifestyle $\rightarrow$ sports $\rightarrow$ news". This path acts as an input for the RecomGen. Finally, the **Recommendation Strategy Generation module (RecomGen)** is the "creative" component of RGRec, tasked with generating an array of items based on the prompt path, denoted as $GI$. Drawing on the path charted by the Belief Nudging module, it constructs recommendations that are relevant and varied, enriching the user's experience. The $GI$ blends with the initial recommendation set, ensuring that the final recommendations delivered to the user are comprehensive and engaging. As users engage with these recommendations, their responses are fed back into the system to refine future recommendations.

Through these three core modules, RGRec systematically refines user beliefs and optimal paths in response to user interactions. Throughout this iterative process, the system continuously assesses the likelihood of user acceptance, adapting its approach until it aligns with the user's beliefs, thereby achieving harmony. Simultaneously, as RGRec expands the recommended items using the output of a preference-based recommendation system, it effectively upholds both responsibility and usability.

### A. The Multi-faceted Reasoning-based Filter Bubbles Detection module (FBDetect)

The Multi-faceted Reasoning-based Filter Bubbles Detection module (FBDetect) is a module designed to identify filter bubbles and users with extreme beliefs from two perspectives: system-level bias and user belief bias. Unlike conventional single-dimensional reasoning models, FBDetect operates in two complementary modes: Forward Reconnaissance (FR) and Counter Reconnaissance (CR). These components work together to identify and confirm the presence of filter bubbles. The FR component evaluates the recommendation system, assessing its potential for causing filter bubbles. It explores system-level bias and behaviors to determine if the system perpetuates a filter bubble effect. In contrast, the CR component focuses on users, pinpointing those influenced by filter bubbles with extremely imbalanced beliefs. FBDetect offers a comprehensive evaluation of filter bubbles, providing insights into biased recommendation models and users with extremely imbalanced beliefs. This implementation contributes to developing fairer recommendations and interventions to mitigate the adverse effects of filter bubbles. For a visual representation of FBDetect's structure (see Fig. 2).

*1) Forward Reconnaissance (FR) Component:* For the FR component, we quantify the potential filter bubble influence of the current recommendation model using mathematical methods. The aspect coverage score [29] is employed in this FR component as a key mathematical validation metric. The formula for calculating aspect diversity coverage $\mu$ is given below:

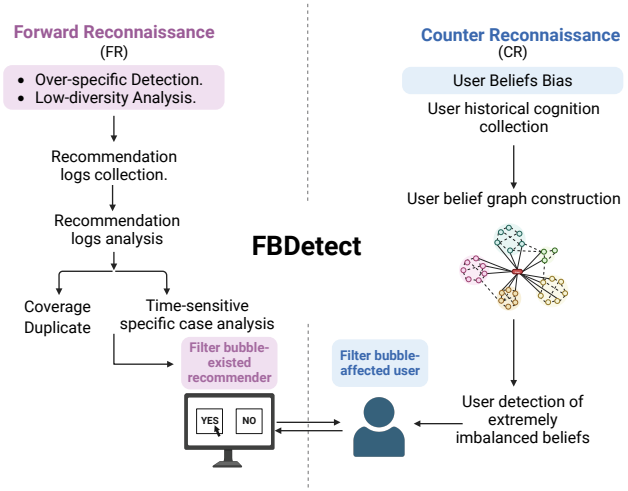



Fig. 2. FBDetect: The Multi-faceted Reasoning-based Filter Bubbles Detection module

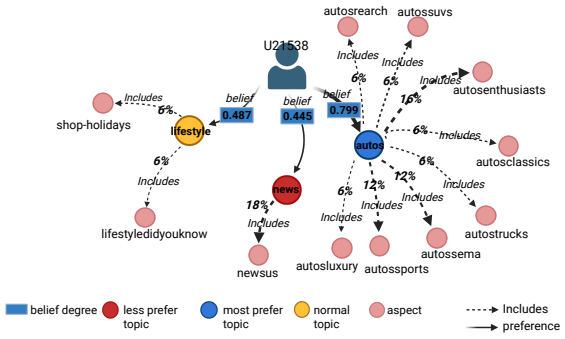



Fig. 3. The user belief network for user "U21538".

$$\mu = \frac{|\{c_x | I_k^x \cap feed \neq \emptyset, \forall x, \forall k\}|}{|C|} \times \frac{|\{c_k^x | I_k^x \cap feed \neq \emptyset, \forall x, \forall k\}|}{\sum_{c_x \in C_{feed}} |C_x^{sub}|}. \tag{4}$$

The first part measures the diversity of topics by calculating the proportion of different topics present in the feed relative to the total number of topics available. While the second part measures the diversity of the aspects within those topics, where the proportion of different aspects covered in the feed for the topics that are present in the feed, relative to the total number of aspects in those topics.

*2) Counter Reconnaissance (CR) Component:* The CR component includes two key steps: constructing the user belief network and detecting users with extremely imbalanced beliefs.

**Constructing the user belief network:** This step involves creating a specific belief network for each user, derived from their historical interaction records. This belief network is employed to analyze user preferences toward different topics and identify users with extremely imbalanced beliefs. Fig. 3 illustrates a representative user belief network for the user "U21538", showcasing an example of user belief network construction within the CR component.

In Fig. 3, the red, blue, and yellow nodes represent the topics "news", "autos", and "lifestyle", with which user "U21538" has historically interacted. The pink nodes represent the aspects related to each topic that the user has interacted with. The weights over edges labeled "Includes", linking a topic to its aspects, are calculated based on the user's click probabilities towards these aspects. Additionally, the connection between the user and a topic, "belief", indicates the user's preference towards the topic and is calculated using the entropy metric in Equation 2.

The CR component constructs a specific user belief network for each user in our dataset, enabling clear identification of users' preferences. Its primary aim is to identify users with extremely imbalanced beliefs. We combine these constructed

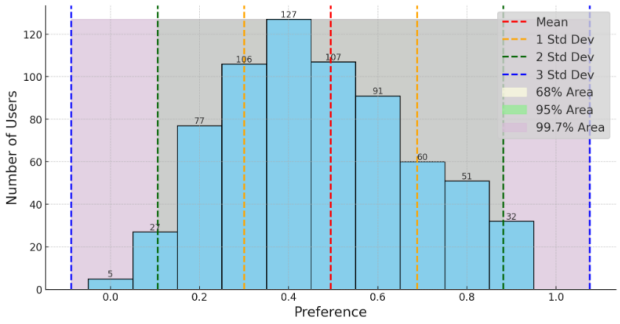



Fig. 4. A specific example of a preference distribution chart within "autos"

belief networks with the empirical rule (the "68-95-99.7" rule) [43] to explore users with extremely imbalanced beliefs.

**Detecting users with extremely imbalanced beliefs:** As previously mentioned, Fig. 3 shows the user belief network of "U21538". From this figure, we can deduce the user's preferences for "autos" and "news", which are 0.799 and 0.445, respectively. To determine whether this user has an extremely imbalanced belief, we collect statistics on all users' preferences for the "autos" and "news". As illustrated in Fig. 4, these statistics include users' preferences for "autos", with the x-axis representing the preference value and the y-axis representing the number of users. We identify "U21538" as a user with extremely imbalanced beliefs by following these steps:

- *Verify Normal Distribution:* The initial step involves verifying whether the preference distribution chart conforms to a normal distribution. To achieve this, we utilize the Kolmogorov-Smirnov (K-S) test [44]. In our scenario, the K-S test is a statistical method employed to compare the preference distribution of a specific topic with the normal distribution. The results of the K-S test typically include the value of the K-S statistic and its associated p-value. The p-value denotes the probability of observing the K-S statistic under the assumption that the distributions of the two datasets are identical. A small p-value (in this study, we set the threshold at 0.05) allows us to reject

the null hypothesis that the two datasets share the same distribution, indicating statistically significant differences. Therefore, in Fig. 4, the p-value exceeds 0.05, suggesting that the preference distribution within "autos" conforms to a normal distribution.

- *Distribution Segment:* Once it is confirmed that the preference distribution chart adheres to a normal distribution, the next step involves segmenting the distribution using the "68-95-99.7" rule [43], as depicted in Fig. 4. In the figure, it can be seen that approximately 68% of the data falls within one standard deviation of the mean, about 95% within two standard deviations, and roughly 99.7% within three standard deviations. In our study, we categorize users outside the 68% area (outside the yellow lines) as having extreme beliefs in the topic. Therefore, when user *U21538* (see Fig. 3) prefers "autos" with 0.799 degrees, falling outside the 68% area in Fig. 4, and shows less interest in "news" with 0.445 degrees, also falling outsides the 68% area in the "news" preference distribution chart. This user is identified as having extremely imbalanced beliefs and is further forwarded to the Belief Nudging module to mitigate the extremely imbalanced beliefs, i.e., filter bubbles.

### B. Belief Nudging Module

RGRec combines users' belief graphs and nudging techniques to gently stimulate users' interests in topics that were initially less preferred. The primary goal of the Belief Nudging module within the RGRec framework is to identify the most effective recommendation prompt path for the subsequent RecomGen. This is achieved by bridging the gap between the user's most favored and less preferred topics.

*1) Adaptive Path Exploration Algorithms:* The recommendation prompt path begins with topics that users highly favor (e.g., "autos" as shown in Fig. 3) and ends with those topics they are less inclined towards (such as "news" in Fig. 3). To connect these points, an adaptive path exploration algorithm discovers additional topics, forming a comprehensive recommendation prompt path and laying the groundwork for future nudge-based recommendations.

The nudge recommendation strategy in RGRec is based on this recommendation prompt path, where the adaptive path exploration algorithm automatically constructs the path based on user feedback. This forms a closed-loop feedback system that interlinks user feedback and system recommendations, dynamically adapting to deliver contextually appropriate prompts tailored to the user's current preferences.

The adaptive path exploration algorithm, inspired by the shortest path exploration algorithm known as *Dijkstra's algorithm* [45], starts from a central point, traverses neighboring points, and identifies the point with the highest weight as the starting point for the next step. We have proposed an enhanced version of this algorithm to accommodate the dynamic nature of user and topic relationships. This algorithm integrates the evolving user perceptions and topic relationships into the path discovery process. The objective is to discover a recommendation prompt path $p_{i,t}$ for an identified filter bubble-affected user $u_i$ at time step $t$. This path incorporates contextually rich items into the recommendations, thus providing more diverse content and promoting belief harmony.

In the recommendation prompt path $p_{i,t}$ for the user $u_i$, $c_x^{SP_i}$ and $c_y^{LP_i}$ represent the user's most and less preferred topics, respectively, serving as the path's start and end points. The initial node of the path is denoted as $p_{i,t}(k) = c_x^{SP_i}$ with k=1. The selection of the subsequent node, the $(k+1)^{th}$ topic $c_x$ from the set $C$, is determined by maximizing the following expression:

$$p_{i,t}(k+1) = \arg\max_{c_{x'}} \rho(c_x^{SP_i}, c_{x'}) + b_{ix'} * rej_{w.t}, \qquad (5)$$

where $\rho(c_x^{SP_i}, c_{x'})$ calculates the topic similarity between $c_x^{SP_i}$ and $c_{x'}$. The term $b_{ix'}$ is the belief degree of user $u_i$ towards topic $c_{x'}$ at the current time step, $rej_{w.t}$ is a rejection weight, typically set to 1.

We incorporate a tolerance threshold, denoted as $\theta$, to modulate the parameter $rej_{w.t}$, emphasizing the significance of user feedback. Users impacted by filter bubbles tend to have their decisions heavily influenced by these bubbles. An item congruent with a user's beliefs is more likely to be accepted, whereas items not aligned with preferred topics often face acceptance challenges. To evaluate the effectiveness of a topic within a recommendation prompt path, we monitor the frequency of rejections for topic-related items. If a user consistently rejects items from $GI$ that are generated based on the recommendation prompt path, these items are deemed ineffective, leading to an assignment of a $rej_{w.t}$ value of -1. This approach adapts effectively to evolving user preferences. The exploration process concludes once the path $p_{i,t}$ includes the user's less preferred topic $c_y^{LP_i}$.

Utilizing the start and end nodes identified as "autos" and "news" in Fig. 5, and employing the adaptive path exploration algorithm, the recommendation prompt path at time $t$ is formulated as $p_{i,t}$ = "autos $\rightarrow$ lifestyle $\rightarrow$ sports $\rightarrow$ news".

*2) Nudging Strategy:* Our nudging process incorporates incremental computing techniques [46] to enhance the efficiency of recommendation calculations. By breaking down the recommendation path into smaller segments, or sub-paths, the system recalibrates only the affected sub-path in response to a user's specific preference, instead of recalculating the entire path. This segmented approach is particularly effective for managing lengthy paths and surpasses the capabilities of traditional sequential recommendation systems. It not only reduces the number of recommendations needed but also increases overall efficiency. Once the path "autos $\rightarrow$ lifestyle $\rightarrow$ sports $\rightarrow$ news" is established, the nudging strategy is employed to refine and finalize the recommendation prompt path. For a clear illustration, both Figure 5 and Algorithm 1 demonstrate the recommendation process within the RKGRec nudging framework. This process includes generating recommendation paths through nudging, creating $GI$ by the RecomGen, and subsequently reconstructing the user belief graph.

In Fig. 5, the identified optimal path is "autos $\rightarrow$ lifestyle $\rightarrow$ sports $\rightarrow$ news". This sequence forms the basis for the nudging strategy, which is fine-tuned based on user feedback.

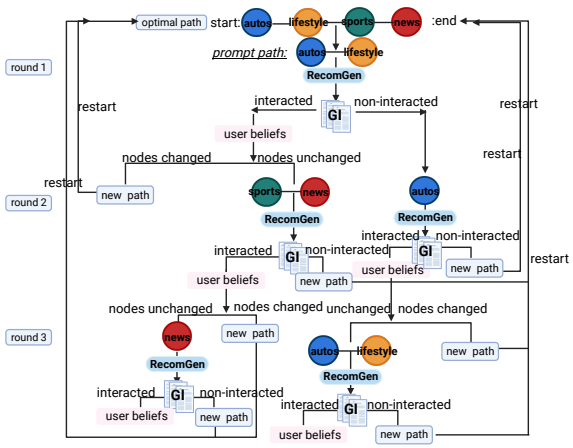



Fig. 5. An actual example of nudging recommendations process

Conforming to the principles of incremental computing, the primary focus is initially on the "autos → lifestyle" segment. RGRec dynamically updates the user's belief network as the user positively engages with the "GI" content generated from RecomGen based on this segment. If the user's most and least preferred topics remain constant, the model progresses to recommend the subsequent segment, "sports → news", instead of restarting the entire path exploration process. However, if the user's topic preferences shift, these altered topics are reintroduced to the adaptive path exploration algorithm to create a new recommendation path.

### *C. The Generative Artificial Intelligence-based Recommendation Strategy Generation module (RecomGen)*

When combined with nudge recommendation techniques, the RecomGen efficiently exploits the interconnectedness of information. This combination presents a solution to address information gaps that may arise during end-to-end recommendation processes. By employing the capabilities of the Large Language Models (LLM), e.g., GPT-3.5 Turbo, this approach offers rich semantic information at each step in the recommendation path, thereby strengthening the relationships between individual points. This strategy effectively engages users' interest in specific topics, fostering intrigue in areas they might find less attractive.

As previously mentioned, a nudging prompt $p_{i,t} = \{c_x^{SP_i}, ..., c_y^{LP_i}\}$ is generated for each time step. This prompt represents an optimal path between the starting node $c_x^{SP_i}$, and the end node $c_y^{LP_i}$. To leverage rich contextual information from point-to-point paths within the vast landscape of big data, these paths are inputted into the RecomGen as keywords or prompts. This generates contextually rich items $GI$ based on the prompt path.

For example, in Fig. 5, the responsibility of the RecomGen is to obtain the path from each $feed$ and generate $GI$ based on the path. In conclusion, integrating the RecomGen with the nudge strategy in RGRec effectively utilizes interconnections. This approach bridges information gaps, leverages the rich semantic information provided by RecomGen, and employs a nudge strategy to establish strong connections between data points. This strategy effectively stimulates user interest in topics that initially receive less attention and promotes user belief harmony.

**Algorithm 1** Nudge Recommendation Process of RGRec

**Input:** the current recommendation prompt path $p_{i,t} = \{c_x^{SP_i}, ..., c_y^{LP_i}\}$ as the current prompt $p$

Initialise an empty heap $Q = []$
Q = Binary Split Function$(p)$ ▷ Binary Split Function adopts incremental computing techniques to break down path $p$
**while** length(Q) $> 0$ **do**
  Set current prompt $p = Q[0]$
  Generate contextually rich items $GI$ with prompt $p = Q[0]$
  Recommend $GI$ to user $u_i$
  **if** $u_i$ accepts $GI$ **then**
    $\check{M}_{i,t} = \check{M}_{i,t} \cup$ GI ▷ $\check{M}_{i,t}$ represents user $u_i$'s collection of accepted items until time step $t$
    Update user belief graph $G_{u_i}$
    Q.pop(0)
  **else**
    $\bar{P}_{i,t}.append(p)$ ▷ $\bar{P}_{i,t}$ records user $u_i$'s declined recommendation prompt paths from time 0 to $t$
    **if** count$(\bar{P}_{i,t}, p) > \theta$ **then**
      Update rejection weight $rej_{w.t}$
    **end if**
    $p_t^1, p_t^2 =$ Binary Split Function $(p)$
    Q.pop(0)
    Push $p_t^1, p_t^2$ to Q
  **end if**
  **if** Binary Split Function $(p)$ is None **then**
    Reschedule path
  **end if**
**end while**
**function** BINARY SPLIT FUNCTION$(p_t)$
  **if** $length(p_t) == 2$ **then**
    **return**
  **else**
    **if** $mod(length(p_t), 2) == 1$ **then**
      $p_t^1 = p_t[0 : \frac{length(p_t)-1}{2}]$
      $p_t^2 = p_t[\frac{length(p_t)+1}{2} :]$
    **else**
      $p_t^1 = p_t[0 : \frac{length(p_t)}{2}]$
      $p_t^2 = p_t[\frac{length(p_t)}{2} :]$
    **end if**
  **end if**
  **return** $p_t^1, p_t^2$
**end function**

**Acceptance Probability Equation**

The RecomGen incorporates a list of contextual rich items $GI$ into the original recommendation list $feed_{original}$ and generates the final recommendation list $feed = \{i_1, i_2, .., i_j\}$. Once user $u_i$ receives this $feed$, the probability $AP_{u_i}^{i_j}$ that whether to accept an item $i_j$ in the recommendation list can be calculated using Equation 6:

$$AP_{u_i}^{i_j} = \frac{b_{ix}}{\sum_{c_{x'} \in \hat{C}_i} b_{ix'}}, i_j \in I_k^x \tag{6}$$

where $b_{ix}$ represents the belief degree of $u_i$ towards topic $c_x$, and $\sum_{c_{x'} \in \hat{C}_i} b_{ix'}$ calculates the total belief degrees of $u_i$.

## V. EXPERIMENTS

We conduct experiments from two general directions: system and user. The system-centered experiment primarily aims to demonstrate the effectiveness of RGRec as an intermediate agency in alleviating the system filter bubble. From the user's perspective, four user-centered experiments are conducted. These include detecting RGRec's positive effect on increasing user belief diversity, examining its effectiveness in motivating filter bubble-impacted users' interests in topics they are initially less interested in, and analyzing RGRec's ability to reduce the number of filter bubble-impacted users. Finally, we perform two parametric analyses to assess the effects of different RGRec recommendation weights and explore the impact of the threshold in RGRec on user belief diversity.

### A. Experiment Setup

*1)* ***Dataset****:* In the experiments, we utilize two real-world datasets: the Microsoft News Dataset (MIND) [1] and IMDB [2] Dataset. MIND is a public news recommendation dataset encompassing user interaction data gathered from Microsoft News. It comprises data from 5,000 users, encompassing 230,117 user reading records and 51,287 news with 17 topics. IMDB is a movie recommendation dataset consisting of 25,000 movie rating records from 333 users and a selection of 2,586 movies in 16 movie topics. We adopt a pre-trained language model, BERT [42], to represent textual features in a vector space, capturing the semantic essence of the items. Such a method has already learned much about language structures and patterns [47] and has been widely adopted in recommendation studies [48], [49], [50].

The FBDetect module in RGRec, identified 180 and 20 filter bubble-affected users from the MIND and IMDB datasets, respectively. These identified users, denoted as $u^{FB}$ in our paper, are the focus of subsequent filter bubble mitigation experiments.

*2)* ***Simulation of User Behaviors.****:* Given the impracticality and high cost associated with online testing for researchers, we have designed an offline evaluation approach: (1) Implement an "Acceptance Probability Equation" 6 to simulate user feedback, (2) generate recommendations using a "Nudge Strategy" based on the simulated user feedback, and (3) evaluate the recommendations by considering diversity and efficacy.

### B. Parameter settings and baselines

**Baselines:** We assess the performance of RGRec in comparison with several established baseline methods:

- Content-Based Filtering (CB) [18]: This strategy recommends existing items based solely on content-based filtering.
- Collaborative Filtering (CF) [17]: This approach recommends existing items using only user collaborative filtering.
- Neural Graph Collaborative Filtering (NGCF) [51]: Enhances recommendation by using user-item graphs to model collaborative signals.
- LightGCN (LGCN) [52]: A simplified, yet effective, model focusing on neighborhood aggregation; outperforms NGCF while being easier to train.
- Disentangled Graph Collaborative Filtering (DGCF) [53]: Captures user intent diversity by analyzing user-item relationships.
- RGRec: This method suggests a set of items (designated as $GI$) as recommendation feeds, derived using the RGRec approach.

Building upon the baselines described above, in our experiments, we combined RGRec with each of the baselines (except the standalone RGRec), producing six additional experimental baselines. These combinations are all named with the superscript "$*$", such as CB$^*$ and LGCN$^*$. In total, we used 11 baselines for comparative experiments, aiming to demonstrate the superior performance of the models when combined with RGRec in breaking the filter bubble and moderating user extreme beliefs.

**Evaluation Metrics:**

We assess RGRec from both system and user perspectives. From the system viewpoint, Experiment V-C1 measures recommendation diversity using aspect coverage $\mu$ as defined in Equation 4. To confirm that observed differences in experimental outcomes are not due to chance, we employ the Two-Tailed Test $t$ and corresponding $p$-value [54], attributing significant differences to model factors. The value of $t$ is shown below:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \tag{7}$$

where $\bar{X}_1$ and $\bar{X}_2$ are the means of the two sample groups, and $S_p$ is the pooled standard deviation. $n_1$ and $n_2$ represent the sample sizes of these groups. The $p$ value represents the probability of observing the $t$ statistic observed, assuming no significant difference exists between the two sample groups.

From the user perspective, our experiments focus on three areas. In Experiment V-C1, we evaluate the diversity of users' belief networks using the same $\mu$ metric in Equation 4. Experiment V-C2 analyzes the evolution of users' interests in highly favored and less favored topics over time, employing Equation 2. Lastly, Experiment V-C3 involves counting the number of filter bubble-affected users, using the CR of the FBDetect module, as detailed in Section IV-A2.

**Parameters:** The proportion $w$ of RGRec-generated items $GI$ is considered as a parameter in our experiments. We analyze the impact of varying proportions $w$ within a recommendation $feed$ on user belief and the diversity of the recommendation system. Additionally, we conduct a parameter analysis experiment of the tolerance threshold $\theta$, which involves tracking user feedback regarding the generated $GI$, described in Section IV-B.

### C. Experimental Results

*1) Experiment 1:* ***Coverage Analysis****:* The primary aim of this experiment is to assess the impact of RGRec on

[1] https://msnews.github.io/

[2] https://www.kaggle.com/datasets/meastanmay/imdb-dataset?select=tmdb_5000_movies.csv/

TABLE II
COVERAGE ANALYSIS OF RECOMMENDATION MODELS BASED ON MIND AND IMDB DATASETS. BOLDFACE DENOTES THE HIGHEST SCORE. MARKING WITH UNDERLINE DENOTES THE SIGNIFICANCE P-VALUE$<0.05$ COMPARED WITH THE BASE MODEL.

| Times | MIND | | | | | | | | | | | IMDB | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **CB** | **CB*** | **CF** | **CF*** | **DGCF** | **DGCF*** | **NGCF** | **NGCF*** | **LGCN** | **LGCN*** | **RGRec** | **CB** | **CB*** | **CF** | **CF*** | **DGCF** | **DGCF*** | **NGCF** | **NGCF*** | **LGCN** | **LGCN*** | **RGRec** |
| $feed_1$ | 0.058 | 0.294 | 0.184 | 0.335 | 0.176 | 0.294 | 0.294 | 0.470 | 0.294 | 0.412 | 0.294 | 0.062 | 0.500 | 0.212 | 0.526 | 0.188 | 0.438 | 0.250 | 0.375 | 0.125 | 0.479 | 0.500 |
| $feed_2$ | 0.059 | 0.205 | 0.186 | 0.202 | 0.176 | 0.353 | 0.235 | 0.470 | 0.235 | 0.412 | 0.195 | 0.062 | 0.345 | 0.207 | 0.386 | 0.125 | 0.500 | 0.250 | 0.563 | 0.250 | 0.354 | 0.344 |
| $feed_3$ | 0.058 | 0.154 | 0.181 | 0.207 | 0.176 | 0.176 | 0.294 | 0.294 | 0.294 | 0.294 | 0.132 | 0.062 | 0.238 | 0.209 | 0.328 | 0.125 | 0.469 | 0.250 | 0.375 | 0.125 | 0.250 | 0.216 |
| $feed_4$ | 0.059 | 0.145 | 0.179 | 0.241 | 0.117 | 0.235 | 0.294 | 0.294 | 0.294 | 0.264 | 0.123 | 0.062 | 0.180 | 0.201 | 0.274 | 0.125 | 0.313 | 0.219 | 0.281 | 0.125 | 0.188 | 0.147 |
| $feed_5$ | 0.059 | 0.151 | 0.182 | 0.215 | 0.117 | 0.117 | 0.294 | 0.294 | 0.294 | 0.294 | 0.119 | 0.062 | 0.163 | 0.207 | 0.259 | 0.281 | 0.344 | 0.219 | 0.281 | 0.188 | 0.250 | 0.127 |
| $feed_6$ | 0.059 | 0.199 | 0.184 | 0.242 | 0.059 | 0.117 | 0.294 | 0.294 | 0.235 | 0.265 | 0.171 | 0.062 | 0.163 | 0.216 | 0.254 | 0.250 | 0.281 | 0.188 | 0.313 | 0.125 | 0.188 | 0.127 |
| $feed_7$ | 0.059 | 0.205 | 0.176 | 0.212 | 0.117 | 0.117 | 0.235 | 0.412 | 0.294 | 0.294 | 0.174 | 0.062 | 0.190 | 0.209 | 0.268 | 0.188 | 0.250 | 0.219 | 0.344 | 0.125 | 0.250 | 0.149 |
| $feed_8$ | 0.059 | 0.178 | 0.186 | 0.215 | 0.176 | 0.117 | 0.235 | 0.353 | 0.235 | 0.324 | 0.154 | 0.062 | 0.184 | 0.197 | 0.269 | 0.125 | 0.156 | 0.219 | 0.219 | 0.250 | 0.250 | 0.158 |
| $feed_9$ | 0.059 | 0.164 | 0.185 | 0.224 | 0.117 | 0.117 | 0.294 | 0.471 | 0.294 | 0.382 | 0.139 | 0.062 | 0.170 | 0.214 | 0.247 | 0.062 | 0.188 | 0.188 | 0.250 | 0.125 | 0.188 | 0.139 |
| $feed_{10}$ | 0.059 | 0.156 | 0.191 | 0.221 | 0.117 | 0.117 | 0.294 | 0.412 | 0.235 | 0.352 | 0.136 | 0.062 | 0.164 | 0.218 | 0.243 | 0.125 | 0.125 | 0.219 | 0.219 | 0.125 | 0.1875 | 0.135 |
| *sum.* | 0.588 | **1.851** | 1.832 | **2.315** | 1.352 | **1.941** | 2.765 | **3.765** | 2.706 | **3.294** | 1.501 | 0.625 | **2.297** | 2.089 | **3.053** | 1.594 | **3.063** | 2.219 | **3.219** | 1.563 | **2.583** | 2.041 |
| *Improv.* | - | 214.79% | - | 26.31% | - | 43.48% | - | 36.17% | - | 21.74% | - | - | 267.6% | - | 46.15% | - | 92.16% | - | 45.07% | - | 65.33% | - |

content diversification compared to the baseline model, and its effect on user belief networks. To accomplish this, we identified users affected by filter bubbles using specific criteria, including users such as "U25354" from the MIND dataset and "U128" from the IMDB dataset. For each of these users, we conducted a series of experimental rounds where they received 10 recommendations in each round, corresponding to a unique "feed". This process was repeated for 100 rounds. The repetition of these rounds aimed to ensure the reliability of our results by reducing the influence of randomness. Finally, the data from these 100 rounds were averaged to obtain the final experimental results, offering a comprehensive evaluation of RGRec's effectiveness in enhancing diversity in recommendations.

**Coverage Analysis for Systems**: Initially, we analyzed the evolution of content diversity across seven different models, focusing on selected users (user "U25354" from MIND and user "U128" from IMDB). The detailed results are presented in Table II, showcasing each model's diversity coverage. Note that "*sum.*" represents each model's total diversity coverage degree throughout the recommendation process, and "*Improv.*" indicates the growth rate in diversity.

In the MIND dataset, the models with $wR$ suffix, indicating RGRec integration, generally exhibit significant improvements in diversity coverage compared to their standard counterparts. Notably, models such as DGCF*, NGCF*, and LGCN* show substantial enhancements, as reflected by their high *sum.* values and growth rates. This indicates that incorporating RGRec markedly enhances recommendation diversity.

Similarly, in the IMDB dataset, models integrated with RGRec demonstrate considerable improvements in diversity. The impact of RGRec is particularly pronounced in models like NGCF* and LGCN*, which display the highest growth rates in diversity.

The statistical significance values for all models integrated with RGRec on both datasets are all below the 0.05 threshold compared with the base model, suggesting that the improvements in diversity coverage by RGRec-integrated models are statistically significant compared to their counterparts. This underscores the vital role of RGRec in augmenting recommendation diversity.

Overall, the analysis demonstrates that integrating RGRec into recommendation models significantly enhances the diversity of content recommended to users, which is crucial for mitigating filter bubble effects.

**Coverage Analysis for User Belief Networks.** Table III offers a detailed examination of the impact of different recommendation models on the diversity of user beliefs across the MIND and IMDB datasets.

In both datasets, models integrated with RGRec demonstrate a significant improvement in total diversity coverage (*sum.*) compared to their respective baseline models. This indicates that integrating RGRec effectively broadens the range of user beliefs. The most notable diversity gains are seen in models like DGCF*, NGCF*, and LGCN*, especially in the MIND dataset. This highlights the effectiveness of these RGRec-enhanced models in diversifying user recommendations.

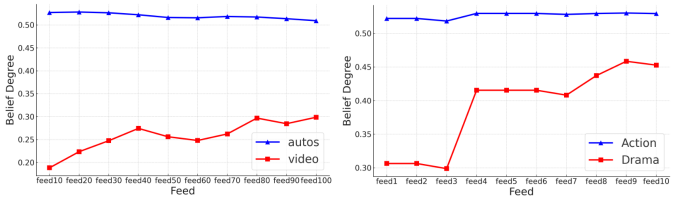
The *Improv.* metric shows a marked percentage increase in user belief diversity for the RGRec integrated models, particularly in the MIND dataset. For instance, the DGCF* model shows improvements exceeding 600%. The consistent enhancements across various models and both datasets reinforce RGRec's effectiveness in enhancing user belief diversity. The statistical significance values for all models integrated with RGRec on both datasets are all below the 0.05 threshold compared with the base model, confirming that the improvements in user belief diversity are statistically significant and attributable to the model variations, particularly due to RGRec integration. This statistical robustness emphasizes the reliability of the trends observed.

Overall, the analysis confirms that the inclusion of RGRec markedly enhances the diversity of user beliefs, which is crucial in mitigating filter bubble effects and enriching user belief diversity.

*2) Experiment 2:* ***User Beliefs Analysis****:* RGRec is designed to present diverse information to users based on existing preference-based recommendation systems. Our goal is to demonstrate that users may come to accept information they initially preferred less through RGRec recommendations over a relatively short timeframe, thereby broadening their perspectives. To this end, we selected user "U276" from the IMDB dataset, who presents a shorter recommendation path $p_{shorter.u}$, indicating closer feature distances between most and less interested topics. This approach allows us to observe changes in user beliefs within a short period. Additionally,

TABLE III
COVERAGE ANALYSIS OF USER BELIEFS ON THE MIND AND IMDB DATASETS. BOLDFACE DENOTES THE HIGHEST SCORE. MARKING WITH UNDERLINE DENOTES THE SIGNIFICANCE P-VALUE$<0.05$ COMPARED WITH THE BASE MODEL.

| Times | MIND | | | | | | | | | | | IMDB | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **CB** | **CB*** | **CF** | **CF*** | **DGCF** | **DGCF*** | **NGCF** | **NGCF*** | **LGCN** | **LGCN*** | **RGRec** | **CB** | **CB*** | **CF** | **CF*** | **DGCF** | **DGCF*** | **NGCF** | **NGCF*** | **LGCN** | **LGCN*** | **RGRec** |
| $feed_1$ | 0.056 | 0.090 | 0.056 | 0.104 | 0.059 | 0.294 | 0.059 | 0.294 | 0.059 | 0.294 | 0.047 | 0.038 | 0.097 | 0.049 | 0.090 | 0.062 | 0.375 | 0.031 | 0.375 | 0.062 | 0.062 | 0.085 |
| $feed_2$ | 0.059 | 0.116 | 0.075 | 0.133 | 0.059 | 0.471 | 0.059 | 0.294 | 0.059 | 0.353 | 0.074 | 0.054 | 0.147 | 0.082 | 0.159 | 0.062 | 0.375 | 0.094 | 0.375 | 0.062 | 0.062 | 0.104 |
| $feed_3$ | 0.059 | 0.126 | 0.089 | 0.161 | 0.059 | 0.471 | 0.059 | 0.294 | 0.059 | 0.353 | 0.091 | 0.059 | 0.172 | 0.111 | 0.198 | 0.062 | 0.375 | 0.094 | 0.375 | 0.062 | 0.062 | 0.126 |
| $feed_4$ | 0.059 | 0.135 | 0.099 | 0.186 | 0.059 | 0.471 | 0.059 | 0.294 | 0.059 | 0.353 | 0.106 | 0.062 | 0.183 | 0.133 | 0.223 | 0.125 | 0.375 | 0.094 | 0.375 | 0.062 | 0.125 | 0.145 |
| $feed_5$ | 0.059 | 0.144 | 0.110 | 0.204 | 0.059 | 0.471 | 0.059 | 0.294 | 0.059 | 0.353 | 0.116 | 0.062 | 0.191 | 0.151 | 0.243 | 0.125 | 0.375 | 0.125 | 0.375 | 0.062 | 0.125 | 0.159 |
| $feed_6$ | 0.059 | 0.158 | 0.122 | 0.221 | 0.059 | 0.471 | 0.059 | 0.294 | 0.059 | 0.353 | 0.132 | 0.062 | 0.197 | 0.172 | 0.257 | 0.125 | 0.375 | 0.125 | 0.375 | 0.062 | 0.125 | 0.164 |
| $feed_7$ | 0.059 | 0.165 | 0.131 | 0.228 | 0.059 | 0.471 | 0.059 | 0.294 | 0.059 | 0.353 | 0.142 | 0.062 | 0.208 | 0.189 | 0.279 | 0.125 | 0.375 | 0.125 | 0.375 | 0.062 | 0.125 | 0.174 |
| $feed_8$ | 0.059 | 0.171 | 0.138 | 0.238 | 0.059 | 0.471 | 0.059 | 0.353 | 0.059 | 0.393 | 0.148 | 0.062 | 0.219 | 0.203 | 0.286 | 0.125 | 0.375 | 0.125 | 0.375 | 0.062 | 0.125 | 0.181 |
| $feed_9$ | 0.059 | 0.178 | 0.149 | 0.245 | 0.059 | 0.471 | 0.059 | 0.353 | 0.059 | 0.393 | 0.155 | 0.062 | 0.228 | 0.213 | 0.300 | 0.125 | 0.375 | 0.125 | 0.375 | 0.062 | 0.125 | 0.188 |
| $feed_{10}$ | 0.059 | 0.184 | 0.160 | 0.252 | 0.059 | 0.471 | 0.059 | 0.353 | 0.059 | 0.393 | 0.167 | 0.062 | 0.233 | 0.220 | 0.310 | 0.125 | 0.375 | 0.125 | 0.375 | 0.062 | 0.125 | 0.192 |
| *sum.* | 0.585 | **1.467** | 1.128 | **1.970** | 0.588 | **4.529** | 0.588 | **3.112** | 0.590 | **3.588** | 1.178 | 0.587 | **1.873** | 1.522 | **2.344** | 1.063 | **3.750** | 1.063 | **3.750** | 0.625 | **1.062** | 1.517 |
| *Improv.* | - | 150.75% | - | 74.56% | - | 669.99% | - | 430.00% | - | 510.00% | - | - | 218.93% | - | 53.98% | - | 252.94% | - | 252.94% | - | 70.00% | - |




(a) Temporal Variation in User Beliefs about Topics of Most Interest and Less Interest on MIND

(b) Temporal Variation in User Beliefs about Topics of Most Interest and Less Interest on IMDB

Fig. 6. Temporal Variation in User Beliefs about Topics of Most Interest and Less Interest

user "U18469", with a longer recommendation path $p_{longer.u}$, is included in the experiment to represent cases with longer recommendation paths.

This experiment employs the CF* model to investigate whether users, in the context of RGRec, will shift their focus from their highly preferred topic to a topic they are less interested in. Fig. 6 illustrates the temporal evolution of interest levels in topics initially less interesting and highly interesting to the users, as indicated by their belief networks.

For the user with longer paths, "$p_{longer}$", we have adjusted the timescale in Fig. 6 so that a single time interval now represents 10 recommendation feeds. This allows us to measure the user's interest level in the "autos" and "video" every 10 recommendations. In contrast, for the user with shorter paths, "$p_{shorter}$", we track belief changes after each recommendation. In these figures, "autos" and "Action" refer to the topics of high interest, while "travel" and "Drama" represent the less favored topics for each user.

From both figures, it is observed that the users' preferences for "autos" and "Action" remain relatively stable or exhibit a slight decline over time. On the other hand, continuous recommendations progressively increase their interest in the initially less favored topics, such as "video" and "Drama". This trend is particularly pronounced for the IMDB user, who shows a significant increase in interest in "Drama". These observations robustly validate the effectiveness of the RGRec-based model in promoting belief harmony among users by diversifying their interests.

*3) Experiment 3:* ***Filter Bubble Users Detection****:* Table IV provides a comprehensive overview of the filter bubble effect across different recommendation models in both the MIND and IMDB datasets. This experiment focuses on tracking the changes over time in the number of users influenced by the filter bubble effect.

In the MIND dataset, a consistent pattern is observed where models integrated with RGRec consistently affect fewer users with the filter bubble than their original counterparts. This trend is evident across all recommendation feeds, showcasing RGRec's effectiveness in reducing the filter bubble impact.

Particularly, the original DGCF and NGCF models, which initially show a high number of users affected, see a significant decrease following RGRec integration. This highlights the model's capability to diversify user recommendations.

Similarly, in the IMDB dataset, RGRec-enhanced models also demonstrate a reduction in the number of users affected by the filter bubble. For example, the CF* model shows a decrease from 6 to 4 users affected and then exhibits fluctuation, reflecting the dynamic nature of user preferences and the system's adaptability. RGRec's ability to adjust the recommendation path based on user feedback likely contributes to these changes, allowing for a more responsive recommendation system that aligns with evolving user interests.

Overall, the analysis from the table confirms the effectiveness of the RGRec strategy in mitigating the filter bubble effect across different recommendation models and datasets. The observed temporal fluctuations further highlight the dynamic interaction between users and the recommendation system.

*4) Experiment 4:* ***Parameter Analysis-1****:* As mentioned in Subsection V-B, RGRec utilizes two key parameters: the proportion $w$ of RGRec-generated contextually rich items $GI$, and a tolerance threshold $\theta$ that tracks user feedback on these items. This experiment examines the impact of varying the weights $w$ of RGRec-generated items within a recommendation feed on user belief diversity. We focus on how changing the weight of RGRec-generated items influences the diversity of user perspectives. By adjusting the $w$ parameter, we assess the extent of RGRec's influence on the recommendation outcomes. The experiment employs user "U1629" from the MIND dataset as a case study and uses the CF* model.

As shown in Fig. 7, the weight assigned to RGRec recommendations has a significant effect on user belief diversity.

TABLE IV
FILTER BUBBLE USERS DETECTION ON MIND AND IMDB DATASETS

| Times | MIND | | | | | | | | | | | IMDB | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **CB** | **CB*** | **CF** | **CF*** | **DGCF** | **DGCF*** | **NGCF** | **NGCF*** | **LGCN** | **LGCN*** | **RGRec** | **CB** | **CB*** | **CF** | **CF*** | **DGCF** | **DGCF*** | **NGCF** | **NGCF*** | **LGCN** | **LGCN*** | **RGRec** |
| $feed_1$ | 2 | 2 | 28 | 24 ↓ | 26 | 18 ↓ | 26 | 16↓ | 20 | 16 ↓ | 26 | 0 | 0 | 6 | 4 ↓ | 6 | 5 ↓ | 6 | 5 ↓ | 6 | 5 ↓ | 4 |
| $feed_2$ | 2 | 2 | 28 | 24 ↓ | 26 | 16 ↓ | 26 | 16↓ | 20 | 16 ↓ | 26 | 0 | 0 | 6 | 4 ↓ | 6 | 5 ↓ | 6 | 5 ↓ | 6 | 5 ↓ | 4 |
| $feed_3$ | 2 | 2 | 28 | 24 ↓ | 26 | 14 ↓ | 26 | 18↓ | 20 | 18 ↓ | 26 | 1 | 0 ↓ | 6 | 4 ↓ | 6 | 5 ↓ | 8 | 5 ↓ | 7 | 5 ↓ | 4 |
| $feed_4$ | 4 | 4 | 28 | 24 ↓ | 26 | 14 ↓ | 28 | 18↓ | 22 | 18 ↓ | 26 | 2 | 0 ↓ | 6 | 6 | 8 | 5 ↓ | 8 | 5 ↓ | 7 | 5 ↓ | 4 |
| $feed_5$ | 4 | 4 | 28 | 24 ↓ | 26 | 10 ↓ | 28 | 18↓ | 24 | 20 ↓ | 26 | 3 | 1 ↓ | 6 | 6 | 8 | 5 ↓ | 8 | 5 ↓ | 7 | 5 ↓ | 4 |
| $feed_6$ | 4 | 4 | 28 | 24 ↓ | 26 | 10 ↓ | 28 | 18↓ | 26 | 18 ↓ | 26 | 3 | 2 ↓ | 6 | 6 | 8 | 5 ↓ | 8 | 5 ↓ | 7 | 5 ↓ | 4 |
| $feed_7$ | 4 | 4 | 28 | 24 ↓ | 28 | 10 ↓ | 28 | 18↓ | 26 | 18 ↓ | 26 | 3 | 3 | 6 | 6 | 8 | 5 ↓ | 8 | 5 ↓ | 8 | 5 ↓ | 4 |
| $feed_8$ | 6 | 4 ↓ | 28 | 24 ↓ | 28 | 10 ↓ | 28 | 16↓ | 28 | 18 ↓ | 24 | 3 | 2 ↓ | 6 | 6 | 8 | 5 ↓ | 8 | 5 ↓ | 8 | 5 ↓ | 4 |
| $feed_9$ | 6 | 4 ↓ | 28 | 24 ↓ | 28 | 10 ↓ | 28 | 16↓ | 28 | 18 ↓ | 24 | 3 | 2 ↓ | 6 | 6 | 8 | 5 ↓ | 8 | 5 ↓ | 8 | 5 ↓ | 4 |
| $feed_{10}$ | 10 | 4 ↓ | 28 | 24 ↓ | 28 | 10 ↓ | 28 | 16↓ | 28 | 16 ↓ | 24 | 4 | 2 ↓ | 6 | 5 ↓ | 8 | 5 ↓ | 8 | 5 ↓ | 8 | 5 ↓ | 4 |

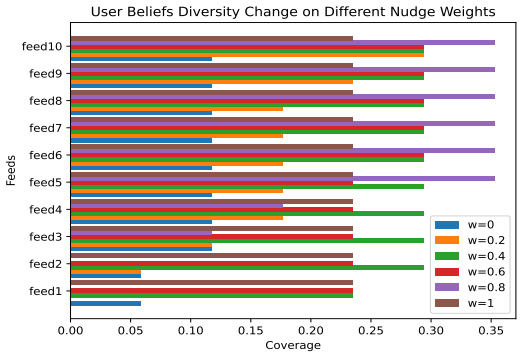



Fig. 7. User Beliefs Diversity Change on Different Nudge Weights

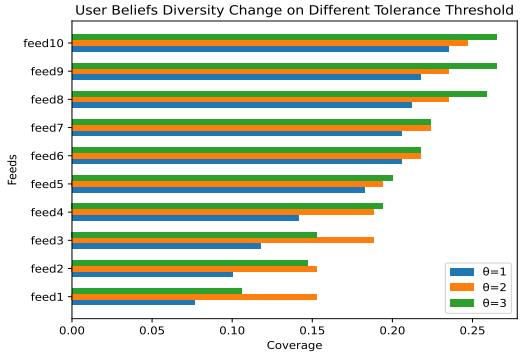



Fig. 8. User Beliefs Diversity Change on Different Tolerance Threshold

It becomes clear that as the number of recommendations increases over time, the impact of the recommendation weight grows more pronounced. Notably, when the weight is set at 0.4, users are more likely to accept the recommendations, and this influence remains steady. However, after a certain number of recommendations, the impact of RGRec-recommended information on users reaches its peak. The more significant the fraction of items recommended by RGRec in the total recommendation list, the more significant the impact on users in the later stages of recommendations.

In conclusion, the analysis shows that the weight of RGRec recommendations considerably affects user belief diversity. The influence of this weight amplifies as the number of recommendations increases. However, there is a saturation point at which the impact of RGRec-recommended information on users reaches its maximum. In the experiment, we select $w$=0.6 as the RGRec recommendation weight owing to its consistently improving performance.

*5) Experiment 4:* ***Parameter Analysis-2****:* In this experiment, we investigate the impact of the tolerance threshold in RGRec on user belief diversity. The threshold setting reflects the model's consideration of user feedback, determining how many times a user must reject recommendations before RGRec alters its nudge recommendation strategy. We simulate various scenarios where users reject recommendations to mimic different feedback situations.

Analyzing Fig. 8, it is clear that user belief diversity gradually increases as the number of recommendations grows, with different threshold values playing a pivotal role. Higher thresholds result in greater user belief diversity, showcasing the system's enhanced adaptability to user feedback. Notably, when the threshold value is 3, the user belief diversity achieves the largest value, surpassing 0.25. This implies that a moderate threshold encourages the system to adjust its recommendations based on user feedback while still maintaining substantial diversity. The experiment intentionally avoids setting thresholds higher than 3 to prevent persistently recommending items that users dislike, which could lead to disengagement from the recommendation system. Across all thresholds, initial acceptance of recommendations is typically low. However, user acceptance of RGRec gradually increases as it adapts to feedback from multiple rejections, refining its nudge strategy and enhancing the overall effectiveness of the approach.

In conclusion, this experiment highlights the critical role of threshold settings in influencing user belief diversity. The findings suggest that a threshold value of 3 maximizes user belief diversity while appropriately respecting user preferences.

## VI. DISCUSSION

The experiments illustrate that models integrating RGRec as an intermediary significantly outperform those without RGRec. RGRec is particularly potent in reducing filter bubbles and achieving a more balanced user belief network. It proves to be more effective and efficient in diversifying recommendations and fostering increased user belief diversity as the recommendation count grows.

Furthermore, the experiments assessed the impact of RGRec on users with varying weights of RGRec recommendations.

The findings reveal its effectiveness in increasing users' acceptance of initially less preferred information, particularly for those with shorter recommendation paths. For users with longer paths, the model aids in gradually adapting to less favored information, underscoring its capacity to improve information acceptance and balance user beliefs.

In summary, experimental results demonstrate that RGRec excels in mitigating filter bubble effects, underscoring its substantial contribution to the evolution of recommendation systems. Beyond just enhancing recommendation diversity, RGRec fosters belief harmony, carrying meaningful implications for the elevation of user satisfaction. These insights play a pivotal role in guiding the development and refinement of future personalized recommendation models and strategies.

## VII. Conclusion and Future Work

In this paper, we focus on addressing the issue of filter bubbles in recommendation systems and propose a Responsible Graph-based Recommendation, RGRec, as a solution to mitigate the negative effects of filter bubbles by promoting belief harmony among users.

RGRec is an intermediary between existing preference-based recommendation systems and users, aiming to facilitate democratic and transparent recommendations. It incorporates several key features. First, it utilizes FBDetect for filter bubble identification. Second, it applies nudging techniques to broaden users' interests and balance their beliefs incrementally. Finally, it incorporates a user feedback loop based on RecomGen to capture evolving user beliefs over time and increase recommendation diversity.

The experimental results explicitly reveal the effectiveness of RGRec in mitigating filter bubbles and balancing user beliefs. For future research, we plan to explore additional techniques to augment the model's performance further, undertake user studies to assess the long-term impacts, and investigate the influence of RGRec on critical thinking abilities and inclusivity. Persistent refinement and evaluation of the model could lead to improved recommendation systems that cater more effectively to the diverse needs of users.

## Acknowledgments

The authors would like to acknowledge the financial support from Callaghan Innovation (CSITR1901, 2021), New Zealand, without which this research would not have been possible. We are grateful for their contributions to the advancement of science and technology in New Zealand. The authors would also like to thank CAITO.ai for their invaluable partnership and their contributions to the project.

## References

[1] H. Ko, S. Lee, Y. Park, and A. Choi, "A survey of recommendation systems: recommendation models, techniques, and application fields," *Electronics*, vol. 11, no. 1, p. 141, 2022.

[2] L. Michiels, J. Leysen, A. Smets, and B. Goethals, "What are filter bubbles really? a review of the conceptual and empirical work," in *Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, 2022, pp. 274–279.

[3] A. G. Ekström, D. C. Niehorster, and E. J. Olsson, "Self-imposed filter bubbles: Selective attention and exposure in online search," *Computers in Human Behavior Reports*, vol. 7, p. 100226, 2022.

[4] M. Elahi, D. Jannach, L. Skjærven, E. Knudsen, H. Sjøvaag, K. Tolonen, Ø. Holmstad, I. Pipkin, E. Throndsen, A. Stenbom *et al.*, "Towards responsible media recommendation," *AI and Ethics*, pp. 1–12, 2022.

[5] F. Alatawi, L. Cheng, A. Tahir, M. Karami, B. Jiang, T. Black, and H. Liu, "A survey on echo chambers on social media: Description, detection and mitigation," 2021.

[6] C. Yu, L. Lakshmanan, and S. Amer-Yahia, "It takes variety to make a world: diversification in recommender systems," in *Proceedings of the 12th international conference on extending database technology: Advances in database technology*, 2009, pp. 368–378.

[7] Q. Grossetti, C. Du Mouza, and N. Travers, "Community-based recommendations on twitter: avoiding the filter bubble," in *Web Information Systems Engineering–WISE 2019*, 2019, pp. 212–227.

[8] G. M. Lunardi, G. M. Machado, V. Maran, and J. P. M. de Oliveira, "A metric for filter bubble measurement in recommender algorithms considering the news domain," *Applied Soft Computing*, vol. 97, p. 106771, 2020.

[9] L. Yang, S. Wang, Y. Tao, J. Sun, X. Liu, P. S. Yu, and T. Wang, "Dgrec: Graph neural network for recommendation with diversified embedding generation," in *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, 2023, pp. 661–669.

[10] Z. Li, Y. Dong, C. Gao, Y. Zhao, D. Li, J. Hao, K. Zhang, Y. Li, and Z. Wang, "Breaking filter bubble: A reinforcement learning framework of controllable recommender system," in *Proceedings of the ACM Web Conference 2023*, 2023, pp. 4041–4049.

[11] Q. Guo, F. Zhuang, C. Qin, H. Zhu, X. Xie, H. Xiong, and Q. He, "A survey on knowledge graph-based recommender systems," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 8, pp. 3549–3568, 2020.

[12] M. Jesse and D. Jannach, "Digital nudging with recommender systems: Survey and future directions," *Computers in Human Behavior Reports*, vol. 3, p. 100052, 2021.

[13] S. Joachim, A. R. M. Forkan, P. P. Jayaraman, A. Morshed, and N. Wickramasinghe, "A nudge-inspired ai-driven health platform for self-management of diabetes," *Sensors*, vol. 22, no. 12, p. 4620, 2022.

[14] W. Li, Q. Bai, and M. Zhang, "Siminer: A stigmergy-based model for mining influential nodes in dynamic social networks," *IEEE Transactions on Big Data*, vol. 5, no. 2, pp. 223–237, 2019.

[15] G. K. Patro, A. Biswas, N. Ganguly, K. P. Gummadi, and A. Chakraborty, "Fairrec: Two-sided fairness for personalized recommendations in two-sided platforms," in *Proceedings of the web conference 2020*, 2020, pp. 1194–1204.

[16] Q. Guo, Z. Sun, J. Zhang, and Y.-L. Theng, "An attentional recurrent neural network for personalized next location recommendation," in *Proceedings of the AAAI Conference on artificial intelligence*, vol. 34, no. 01, 2020, pp. 83–90.

[17] F. Wang, H. Zhu, G. Srivastava, S. Li, M. R. Khosravi, and L. Qi, "Robust collaborative filtering recommendation with user-item-trust records," *IEEE Transactions on Computational Social Systems*, vol. 9, no. 4, pp. 986–996, 2021.

[18] S. Reddy, S. Nalluri, S. Kunisetti, S. Ashok, and B. Venkatesh, "Content-based movie recommendation system using genre correlation," in *Smart Intelligent Computing and Applications: Proceedings of the Second International Conference on SCI 2018, Volume 2*, 2019, pp. 391–397.

[19] Z. Wu, C. Li, J. Cao, and Y. Ge, "On scalability of association-rule-based recommendation: A unified distributed-computing framework," *ACM Transactions on the Web (TWEB)*, vol. 14, no. 3, pp. 1–21, 2020.

[20] Y. Afoudi, M. Lazaar, and M. Al Achhab, "Hybrid recommendation system combined content-based filtering and collaborative prediction using artificial neural network," *Simulation Modelling Practice and Theory*, vol. 113, p. 102375, 2021.

[21] L. V. Bryant, "The youtube algorithm and the alt-right filter bubble," *Open Information Science*, vol. 4, no. 1, pp. 85–90, 2020.

[22] B. Kitchens, S. L. Johnson, and P. Gray, "Understanding echo chambers and filter bubbles: The impact of social media on diversification and partisan shifts in news consumption." *MIS quarterly*, vol. 44, no. 4, 2020.

[23] P. M. Dahlgren, "A critical review of filter bubbles and a comparison with selective exposure," *Nordicom Review*, vol. 42, no. 1, pp. 15–33, 2021.

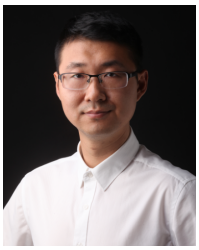
[24] J. Chen, H. Dong, X. Wang, F. Feng, M. Wang, and X. He, "Bias and debias in recommender system: A survey and future directions," *ACM Transactions on Information Systems*, vol. 41, no. 3, pp. 1–39, 2023.

[25] H. Tang, S. Wu, G. Xu, and Q. Li, "Dynamic graph evolution learning for recommendation," in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023, pp. 1589–1598.
[26] D. P. Calvillo, A. B. Swan, and A. M. Rutchick, "Ideological belief bias with political syllogisms," *Thinking & Reasoning*, vol. 26, no. 2, pp. 291–310, 2020.
[27] P. Resnick, R. K. Garrett, T. Kriplean, S. A. Munson, and N. J. Stroud, "Bursting your (filter) bubble: strategies for promoting diverse exposure," in *Proceedings of the 2013 conference on Computer supported cooperative work companion*, 2013, pp. 95–100.
[28] Y. Hu, S. Wu, C. Jiang, W. Li, Q. Bai, and E. Roehrer, "Ai facilitated isolations? the impact of recommendation-based influence diffusion in human society," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, 2022, pp. 5080–5086.
[29] W. Wang, F. Feng, L. Nie, and T.-S. Chua, "User-controllable recommendation against filter bubbles," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022, pp. 1251–1261.
[30] U. Chitra and C. Musco, "Analyzing the impact of filter bubbles on social network polarization," in *Proceedings of the 13th International Conference on Web Search and Data Mining*, 2020, pp. 115–123.
[31] F. Xiao, "Multi-sensor data fusion based on the belief divergence measure of evidences and the belief entropy," *Information Fusion*, vol. 46, pp. 23–32, 2019.
[32] J. Blumenthal-Barby and D. J. Opel, "Nudge or grudge? choice architecture and parental decision-making," *Hastings Center Report*, vol. 48, no. 2, pp. 33–39, 2018.
[33] J. Beshears and H. Kosowsky, "Nudging: Progress to date and future directions," *Organizational behavior and human decision processes*, vol. 161, pp. 3–19, 2020.
[34] S. Wu, W. Li, H. Shen, and Q. Bai, "Identifying influential users in unknown social networks for adaptive incentive allocation under budget restriction," *Information Sciences*, vol. 624, pp. 128–146, 2023.
[35] S. Wu, W. Li, and Q. Bai, "Gac: A deep reinforcement learning model toward user incentivization in unknown social networks," *Knowledge-Based Systems*, vol. 259, p. 110060, 2023.
[36] W. Li, Q. Bai, M. Zhang, and T. D. Nguyen, "Automated Influence Maintenance in Social Networks: an Agent-based Approach," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 10, pp. 1884–1897, 2019.
[37] G. Wang, W. Li, Q. Bai, and E. M.-K. Lai, "Maximizing social influence with minimum information alteration," *IEEE Transactions on Emerging Topics in Computing*, pp. 1–13, 2023.
[38] S. Wu, Q. Bai, and S. Sengvong, "Greencommute: An influence-aware persuasive recommendation approach for public-friendly commute options," *Journal of Systems Science and Systems Engineering*, vol. 27, no. 2, pp. 250–264, 2018.
[39] D.-A. Sitar-Tăut, D. Mican, and R. A. Buchmann, "A knowledge-driven digital nudging approach to recommender systems built on a modified onicescu method," *Expert Systems with Applications*, vol. 181, p. 115170, 2021.
[40] E. Million, "The hadamard product," *Course Notes*, vol. 3, no. 6, pp. 1–7, 2007.
[41] M. Behrendt and S. Harmeling, "Arguebert: How to improve bert embeddings for measuring the similarity of arguments," in *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, 2021, pp. 28–36.
[42] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, 2019, p. 2.
[43] A. Wooditch, N. J. Johnson, R. Solymosi, J. Medina Ariza, and S. Langton, "The normal distribution and single-sample significance tests," *A Beginner's Guide to Statistics for Criminology and Criminal Justice Using R*, pp. 155–168, 2021.
[44] D. J. Steinskog, D. B. Tjøstheim, and N. G. Kvamstø, "A cautionary note on the use of the kolmogorov–smirnov test for normality," *Monthly Weather Review*, vol. 135, no. 3, pp. 1151–1157, 2007.
[45] M. Barbehenn, "A note on the complexity of dijkstra's algorithm for graphs with weighted vertices," *IEEE transactions on computers*, vol. 47, no. 2, p. 263, 1998.
[46] G. Ramalingam and T. Reps, "A categorized bibliography on incremental computation," in *Proceedings of the 20th ACM SIGPLAN-SIGACT symposium on Principles of programming languages*, 1993, pp. 502–510.
[47] Y. Hao, L. Dong, F. Wei, and K. Xu, "Visualizing and understanding the effectiveness of BERT," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 4143–4152.
[48] Q. Zhang, J. Li, Q. Jia, C. Wang, J. Zhu, Z. Wang, and X. He, "Unbert: User-news matching bert for news recommendation," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, 2021, pp. 3356–3362.
[49] C. Jeong, S. Jang, E. Park, and S. Choi, "A context-aware citation recommendation model with bert and graph convolutional networks," *Scientometrics*, vol. 124, pp. 1907–1922, 2020.
[50] Z. Qiu, X. Wu, J. Gao, and W. Fan, "U-bert: Pre-training user representations for improved recommendation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 5, 2021, pp. 4320–4327.
[51] X. Wang, X. He, M. Wang, F. Feng, and T.-S. Chua, "Neural graph collaborative filtering," in *Proceedings of the 42nd international ACM SIGIR conference on Research and development in Information Retrieval*, 2019, pp. 165–174.
[52] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, and M. Wang, "Lightgcn: Simplifying and powering graph convolution network for recommendation," in *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, 2020, pp. 639–648.
[53] X. Wang, H. Jin, A. Zhang, X. He, T. Xu, and T.-S. Chua, "Disentangled graph collaborative filtering," in *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, 2020, pp. 1001–1010.
[54] J. P. Shaver, "What statistical significance testing is, and what it is not," *The Journal of Experimental Education*, vol. 61, no. 4, pp. 293–316, 1993.

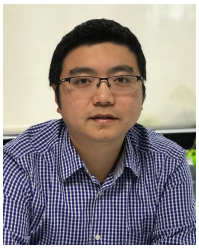
**Mengyan Wang** received her master's degree in Computer and Information Sciences from Auckland University of Technology, New Zealand, in 2019. Currently, she is pursuing a Ph.D. degree at Auckland University of Technology, and her research mainly focuses on recommendation systems and knowledge graphs.

**Yuxuan Hu** received her Ph.D. from the University of Tasmania, Australia, in 2023, the M.E degree from the University of Auckland, New Zealand, in 2017, and the bachelor's degree from Hunan Normal University, China, in 2016. Her research interests include agent-based modeling, social network analysis, and recommendation systems.

**Shiqing Wu** is a Postdoctoral Research Associate in the School of Computer Science at the University of Technology Sydney. He received his Ph.D. from the University of Tasmania in 2022 and a joint B.Sc. from Auckland University of Technology and China Jiliang University in 2016. His research interests involve recommendation systems, reinforcement learning, social influence analysis, and agent-based modeling. He has published several research papers in prestigious venues, such as SIGIR, ICDM, IJCAI, AAMAS, etc.



**Weihua Li** received his Ph.D. from Auckland University of Technology, New Zealand in 2018 and his M.Tech from the National University of Singapore in 2014. He is currently a senior lecturer in the School of Engineering, Computer & Mathematical Sciences. Weihua is an active researcher, and his research interest mainly focuses on Artificial Intelligence, including agent-based modeling and simulation in complex systems and natural language processing. He has over 40 publications and is involved in the organization of a number of international conferences in the field of AI.



**Quan Bai** is a distinguished researcher with a primary focus on machine learning, agent-based modeling and knowledge representation. He earned his Ph.D. in Computer Science from the University of Wollongong in 2007. Over the years, Dr. Bai's career has encompassed a range of pivotal roles in the field. He began as a postdoctoral research fellow at the Commonwealth Scientific and Industrial Research Organisation (CSIRO) in Australia. Subsequently, he expanded his horizons as a lecturer and later as a senior lecturer at Auckland University of Technology in New Zealand. Presently, Dr. Bai serves as an associate professor at the University of Tasmania and leads an active AI Research Group at the university. Dr. Bai has published over 160 research publications, underscoring his commitment to advancing in his field.

**Zihan Yuan** received her Bachelor's degrees in 2019 and Master's degree in 2022 from Monash University, Australia. She is currently a Master of Information Technology and Systems student at the University of Tasmania, pursuing the research pathway in responsible recommendation systems.

**Chenting Jiang** holds a Bachelor's degree in Geography and Economics (BE in Geography and BEc in Economics) from South China Normal University, earned in 2011. In 2021, she achieved a Master's degree in Information and Communication Technology from the University of Tasmania. Her research focuses on diverse areas, including machine learning, neural networks, data assimilation, predicting soil water retention function, and hybrid hydraulic modeling. Additionally, she explores topics in online social networks, influence diffusion, and behavioral economics.