

Understanding Echo Chambers in E-commerce Recommender Systems

| | | | |
|--|--|--|--|
| Yingqiang Ge [*] [†] Rutgers University yingqiang.ge@rutgers.edu | Shuya Zhao [*] New York University sz2257@nyu.edu | Honglu Zhou Rutgers University honglu.zhou@rutgers.edu | Changhua Pei Alibaba Group changhuapei@gmail.com |
| Fei Sun Alibaba Group ofey.sunfei@gmail.com | Wenwu Ou Alibaba Group santong.oww@taobao.com | Yongfeng Zhang Rutgers University yongfeng.zhang@rutgers.edu | |

ABSTRACT

Personalized recommendation benefits users in accessing contents of interests effectively. Current research on recommender systems mostly focuses on matching users with proper items based on user interests. However, significant efforts are missing to understand how the recommendations influence user preferences and behaviors, e.g., if and how recommendations result in *echo chambers*. Extensive efforts have been made in examining the phenomenon in online media and social network systems. Meanwhile, there are growing concerns that recommender systems might lead to the self-reinforcing of user's interests due to narrowed exposure of items, which may be the potential cause of echo chamber. In this paper, we aim to analyze the echo chamber phenomenon in Alibaba Taobao – one of the largest e-commerce platforms in the world.

Echo chamber means the effect of user interests being reinforced through repeated exposure to similar contents. Based on the definition, we examine the presence of echo chamber in two steps. First, we explore whether user interests have been reinforced. Second, we check whether the reinforcement results from the exposure of similar contents. Our evaluations are enhanced with robust metrics, including cluster validity and statistical significance. Experiments are performed on extensive collections of real-world data consisting of user clicks, purchases, and browse logs from Alibaba Taobao. Evidence suggests the tendency of echo chamber in user click behaviors, while it is relatively mitigated in user purchase behaviors. Insights from the results guide the refinement of recommendation algorithms in real-world e-commerce systems.

CCS CONCEPTS

- Information systems → Recommender systems; Web log analysis; Test collections.

^{*}Co-first authors with equal contributions.

[†]This work was done when Yingqiang Ge worked as an intern in Alibaba.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SIGIR '20, July 25–30, 2020, Virtual Event, China

© 2020 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-8016-4/20/07...\$15.00

<https://doi.org/10.1145/3397271.3401431>

KEYWORDS

E-commerce; Recommender Systems; Echo Chamber; Filter Bubble

ACM Reference Format:

Yingqiang Ge, Shuya Zhao, Honglu Zhou, Changhua Pei, Fei Sun, Wenwu Ou, and Yongfeng Zhang. 2020. Understanding Echo Chambers in E-commerce Recommender Systems. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20), July 25–30, 2020, Virtual Event, China*. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3397271.3401431>

1 INTRODUCTION

Recommender systems (RS) comes into play with the rise of online platforms, e.g., social networking sites, online media, and e-commerce[16, 18, 19]. Intelligent algorithms with the ability to offer personalized recommendations are increasingly used to help consumers seek contents that best match their needs and preferences in forms of products, news, services, and even friends[1, 49, 50]. Despite the significant convenience that RS has brought, the outcome of the personalized recommendations, especially how it reforms social mentality and public recognition – which could potentially reconfigure the society, politics, labor, and ethics – remains unclear. Extensive attention has been drawn at this front, thus arriving at the two coined terms, *echo chamber* and *filter bubble*. Both effects might occur after the use of personalized recommenders and entail far-reaching implications. Echo chamber describes the rising up of social communities who share similar opinions within the group [41], while filter bubble [36], as the phenomenon of an overly narrow set of recommenders, was blamed for isolating users in information echo chambers [1].

Owing to the irreversible and striking impact that the internet has brought on the mass communication, echo chamber and filter bubble are appearing in online media and social networking sites, such as MovieLens [33], Pandora [1], YouTube [23], Facebook [37], and Instagram [39]. Significant research efforts have been put forward in examining the two phenomena in online media and social networks [4, 6, 7, 14, 20, 30]. Recently, researchers have concluded that the decisions made by RS can influence user beliefs and preferences, which in turn affect the user feedback, e.g., the behavior of click and purchase received by the learning system, and this kind of user feedback loop might lead to echo chamber and filter bubbles [26]. On the other hand, the two concepts are not isolated, since filter bubble is a potential cause of echo chamber [1, 12].

In this work, we are primarily concerned with the existence and the characteristics of echo chamber in real-world e-commerce systems. We define echo chamber as the effect of users' interests being reinforced due to repeated exposure to similar items or categories of items, thereby generalizing the definition in [26]. This is because users' consuming preferences are so versatile and diverse that cannot simply be classified into positive or negative directions as what it looks like in political opinions [40]. Based on the above definition of echo chamber, we formulate the research in two steps by answering the following two related research questions:

- RQ1: Does the recommender system, to some extent, reinforce user click/purchase interests?
- RQ2: If user interests are indeed strengthened, is it caused by RS narrowing down the scope of items exposed to users?

To measure the effect of recommender systems on users, we first follow the idea introduced in [33] and separate all users into categories based on how often they actually "take" the recommended items. This separation helps us to compare recommendation followers against a controlled group, namely, the recommendation ignorers. The remaining problem is how to measure the effect of echo chamber on each group. Users in social network platforms have direct ways to interact with other users, potentially through actions of friending, following, commenting, etc [38]. A similar analogy is that users in the recommender system could interact with other users *indirectly* through the recommendations offered by the platform, since recommendation lists are usually generated as a result of considering the user's previous preferences and the preferences of similar users (*i.e.*, collaborative filtering). Due to the absence of an explicit network of user-user interaction, which is naturally and commonly provided in social networks, we decide to measure echo chamber in e-commerce at the population level. This is because users who share similar interaction records (*e.g.*, clicking the same products) will be closely located in a latent space, and the cluster of these users in that space, along with its temporal changes, could serve as signals to detect echo chamber. Finally, we measure the content diversity in recommendation lists for each group to see whether recommender system narrows down the scope of items exposed to users, so as to answer RQ2.

The key contributions of our paper can be summarized as follows:

- We study echo chamber effect at a population level by implementing clustering on different user groups, and measure the shifts in user interests with cluster validity indexes.
- We design a set of controlled trials between recommendation followers and ignorers, and employ a wide range of technical metrics to measure the echo chamber effect to provide a broader picture.
- We conduct our experiments based on real-world data from Alibaba Taobao — one of the largest e-commerce platforms in the world. Our analytical results, grounded with reliable validity metrics, suggest the tendency of echo chamber in terms of the user click behaviors, and relatively mitigated effect in user purchase behaviors.

2 RELATED WORK

Today's recommender systems are criticized for bringing dangerous byproducts of echo chamber and filter bubble. Sunstein argued that

personalized recommenders would fragment users, making like-minded users aggregate [41]. The existing views or interests of these users would be reinforced and amplified since "group polarization often occurs because people are telling one another what they know" [41, 42]. Pariser later described filter bubble, as the effect of recommenders making users isolated from diverse content, and trapping them in an unchanging environment [36]. Though both are concerned with the malicious effect that recommenders would pose, echo chamber emphasizes the polarized environment, while filter bubble lays stress on the undiversified environment.

Researchers are expressing their concerns of the two effects, and attempting to formulate a richer understanding of the potential characteristics [4, 10, 23, 32, 39]. Considering echo chamber as a significant threat to modern society as they might lead to polarization and radicalization [11], Risius et al. analyzed news "likes" on Facebook, and distinguished different types of echo chambers [37]. Mohseni et al. reviewed news feed algorithms as well as methods for fake news detection and focused on the unwanted outcomes of echo chamber and filter bubble after using personalized content selection algorithms [30]. They argued that personalized newsfeed might cause polarized social media and the spread of fake content.

Another genre of research aims to clear up strategies to mitigate the potential issues of echo chamber and filter bubble, or design new recommenders to alleviate such effects [2, 3, 13, 17, 22, 35]. Badami et al. proposed a new recommendation model for combating over-specialization in polarized environments after finding that matrix factorization models are easier to learn in polarized environments, and in turn, encourage filter bubbles that reinforce polarization. Tintarev et al. attempted to use visual explanations, *i.e.*, chord diagrams and bar charts, to address the problems [43].

There is a certain amount of work focusing on the detection or measuring of echo chamber and filter bubble, questioning whether they do exist [1, 4, 15, 27, 31, 33]. For example, Hosanagar et al. used data from an online music service, trying to find out whether personalization is, in fact, fragmenting the population, and concluded that it does not [24]. They claimed personalization is a tool that helps users widen their interests, which in turn creates commonality with others. Sasahara et al. suggested echo chambers are somewhat inevitable given the mechanisms at play in social media, specifically, the basic influence and unfriending [38]. Their simulation dynamics showed that the social network rapidly devolves into segregated, homogeneous, and polarized communities, even with the minimal amount of influence and unfriending.

Despite the reasonableness of prior works, severe limitations do exist, making the claims only plausible. One major aspect is that most of the existing works draw conclusions by means of simulation, or relying on some self-defined networks and measurements with simplified dynamics [6, 9, 15, 20, 26, 31]. Building upon the subjective assumptions, whether the modeling and analysis have the capability to reflect the truth seems to be dubious [37]. On the other hand, many of the prior works confound the meaning of the two effects or solely examine one of them without considerations of the other [17, 23, 27, 28, 37]. Exceptions such as [26], disentangles echo chamber from filter bubble, but suffers from the previously-mentioned deficiency, *i.e.*, reliance on simulation and simplified artificial settings. With the desire to address the limitations, we aim to explore the existence of echo chamber in real-world e-commerce

systems while investigating filter bubble via differentiating it as the potential cause of echo chamber. To the best of our knowledge, this is the first work that utilizes real-world data of recommendation and user-item interaction from a large-scale real-world e-commerce platform, with solid validity metrics, instead of from the artificial well-controlled experimental settings. We do not induce any prior assumptions that might be unreliable. We draw analysis from a latent space without relying on any explicit pre-built networks.

3 DATA COLLECTION AND ANALYSIS

3.1 Data Collection

We first collected 86,192 users' five months accessing logs spanning Jan. 1, 2019 to May 31, 2019, from Taobao. There exist three types of accessing logs: browse log, click log, and purchase log.

Browse Log records the recommendations that have been browsed for each user, including the timestamp, page view id, user id, recommended item id, item position in the recommendation list, and whether it was clicked or not. The recommended items per page are known as a page view (PV) in e-commerce.

Click Log is used to record each user's click behaviors in the whole electronic market, which means that user click behaviors outside the recommender list will also be recorded (i.e., the user may initiate a search and click an item). It includes the timestamp, PV id, user id, user profile, clicked item id, and the item price.

Purchase Log is in the same format as the click log, except it is used to record users' purchase behaviors. Besides, since purchase is a more expensive action than click for users, the sparsity of the purchase log is much higher than that of the click log.

We extract dataset with the above components (browse log, click log, and purchase log) for the following three reasons:

- E-commerce is a complicated scenario, where users have multiple types of interactions with items, such as browsing, clicking, purchasing, and rating. The effect of RS might have a different influence on them, because different interactions induce different costs for users, e.g., browsing and clicking merely cost time, while purchasing costs time and money.
- Clicking and purchasing are used to represent consumer's implicit feedback in RS, which contains rich information about consumer preferences. However, browsing contains much more noise as most of the browsed items may be indifferent to a user, which means the user neither likes nor dislikes the item.
- In RS, ratings are user's explicit feedback to items. However, since the platform automatically fulfills the rating score to 5 stars (the highest score) if a user purchased an item without leaving a rating, so the ratings may not reflect users' true preference, thus, we do not use ratings in this work.

We further remove each user's logs in the first two months to make sure that all users have enough time to get familiar with the e-commerce platform, and that the RS receives sufficient interactions from consumers to understand their preferences well.

3.2 Identifying Recommendation Takers

The purpose of our work is to explore the temporal effect of recommendation systems on users, especially to investigate the existence and the characteristics of the echo chamber effect. In order to study

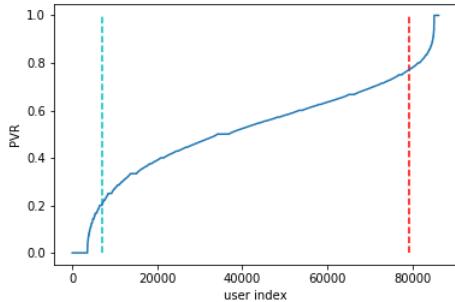


Figure 1: The users sorted by PVR from the lowest to the highest score. Each x-axis index refers to a unique user. The two split points represent 20% (blue) and 80% (red) users.

the effect of “taking” recommendations, we need to classify the users in the dataset into users who “follow” recommendations and users who do not. For consistency, in this paper, we call the ones who “follow” recommendations as the *Following Group* and those who do not as the *Ignoring Group*. Inspired by the experiment setting of [33], we draw comparisons between two groups of consumers – the *Following Group* and the *Ignoring Group* – to explore the effect of recommendation systems on users.

The first step is to classify users into the two groups, and there are several different approaches to this classification. One straightforward approach is to use the ratio between the number of clicked items and the number of browsed items. We can calculate this ratio for each user i based on his or her browsing history. However, one extreme case is that user i only viewed each item once, and the user clicked most of them, but never came back to the platform to use the recommendation system again. This will classify the user into the *Following Group* because the ratio is close to one but he/she is misclassified because he/she actually abandoned the recommender system. We can see that this approach cannot help to investigate the long-term influence of RS on consumers.

In order to fulfill the need for long-term observation, we adopt the “block” idea [33] into our classification task. We first identify clicked PV, which is a recommendation list on a single page where the consumer clicked at least one item displayed in it. Then, we compute the number of clicked PVs over the total number of all PVs for a given consumer, and we define this ratio as PVR (namely, page view ratio). The intuition behind this design is that we believe the effect that RS imposes on consumers depends on the frequency that consumers are exposed to it, as well as consumers' responses to the recommendations (e.g., clicks). Once we have calculated the PVRs for all users, we sort the users from the lowest to the highest PVR scores, and the result is shown in Figure 1. Based on this figure, we define users who took recommendations in at most 20% of their PVs as the *Ignoring Group*, and users who took recommendations in at least 80% of their PVs as the *Following Group*, which gives us 6,183 followers and 6,979 ignorers in total.

3.3 User Interaction Blocks

In order to examine the temporal effect of recommender systems on users, we divide the user-item interaction history (could be *browse*, *click*, or *purchase*) into discrete intervals for each user. We follow the similar “blocks of interaction” idea in [33] to divide the interaction history of a user into intervals, which is used to make

sure that all users have the same amount of interactions with the recommender system throughout an interaction block.

We define an interval as a block consisting of n consecutive interactions, where n is a constant decided by experimental studies. Moreover, different interactions occur at different frequencies, for example, the number of browsed items is much higher than the number of clicked items, and the number of clicked items is again higher than the purchased items, indicating that the length of the block (i.e., n) may vary based on the corresponding interactions. We primarily set the length of the interval as $n = 200$ for browsed items (named as a browsing block), $n = 100$ for clicked items (named as a clicking block), and $n = 10$ for purchased items (named as a purchasing block). If there are not enough interactions to constitute the last block, we will drop it to make sure that all blocks have the same number of interactions. Meanwhile, to ensure the temporal effect of RS, we only keep those users who have at least three intervals in the three months, for each type of user-item interaction. Finally, as shown in Table 1, we have 7,477 users to examine the echo chamber effect on click behaviors, 3,557 users for purchase behaviors, and 7,417 users for browsing behaviors.

Considering the browse log is potentially noisy (i.e., indifferent to a user, see Section 3.1), as well as the fact that clicking and purchasing are commonly used to represent users' implicit feedback to RS, in the following, we use click and purchase behaviors to represent users' preferences on the items (while we detect the echo chamber effect), and examine the temporal changes in content diversity of recommended items via browsing behaviors (which may be the potential cause of echo chamber).

3.4 User Embeddings

As the user interests are closely related to the user interactions with items, we argue that the items that the user clicked can reflect his/her click interests, and the items that the user purchased can represent his/her purchase interests. However, only using discrete indexes to denote the interacted items is not sufficient to represent user interests since we cannot know the collaborative relations between different items only based on the indexes. Following the basic idea of collaborative filtering, we use the user-item interaction information to train an embedding model.

Items are encoded into item embeddings based on one of the state-of-the-art models [46]. To cluster and compare the items for different users at different times, we need to guarantee that the item embeddings are stable across the period of time that is under investigation. For this purpose, the embeddings are trained on all of the collected data until May 31, 2019, which is the last day that our dataset contains. This is for two reasons: (1) since the training data contains all of the user-item interactions, it helps to learn more accurate embeddings; and (2) since the training procedure includes all items under consideration, we can guarantee that all embeddings are learned in the same space. After that, we use the average pooling on the item embeddings to compute the user embeddings. Specifically, we use the average of the item embeddings of items that the user clicked (or purchased) within a user-item interaction block to represent the user's click preferences (or purchase preferences) during a certain period of time. In this way, user embeddings and item embeddings are in the same representation space.

| | Click | Purchase | Browse |
|-----------------|--------|----------|--------|
| Following group | 5, 025 | 2, 099 | 5, 507 |
| Ignoring group | 2, 452 | 1, 458 | 1, 910 |
| All users | 7, 477 | 3, 557 | 7, 417 |

Table 1: Statistics of each user group.

4 MEASURES FOR ECHO CHAMBERS

To answer RQ1, we propose to study the reinforcement in user interests at a population level. The pattern of reinforcement could happen in a simple scenario, where some members highly support one opinion, while others believe in a competing opinion. The phenomenon is reflected as the dense distribution on the two sides of the opinion axis. However, user's interest in e-commerce is much more complicated, such that it cannot be simply classified into positive and negative. What we observe is that users can congregate into multiple groups in terms of distinct preferences. As a result, we implement clustering on user embeddings and measure the change in user interests with cluster validity indexes (more details can be found in Section 4.2). We measure the changes in terms of clustering on the embeddings at the beginning and at the end of the user interaction record, i.e., we compute the user embeddings respectively for the first and the last interaction block and measure the changes. To be clear, we refer to these two blocks as the "first block" and the "last block".

To answer RQ2, we propose to measure the content diversity in recommendation lists at the beginning and the end, and more details can be found in Section 4.3. We examine whether there exists a trend that recommendation systems narrow down the contents provided to the users. Before we cluster the *Following Group* and the *Ignoring Group*, respectively, we need to know whether the two groups are clusterable and what is the appropriate number of clusters for each of the group. Thus, we first examine the clustering tendency and select the proper clustering settings, which will be introduced in Section 4.1.

4.1 Measuring Clusters

4.1.1 Clustering Tendency.

Assessing clustering tendency is employed to evaluate whether there exist meaningful clusters in the dataset before applying clustering methods. We use **Hopkins statistic** (H) [5, 29] to measure the tendency since it can examine the spatial randomness of the data by testing the given dataset with a uniformly random-distributed dataset. The value of H is from 0 to 1. A result close to 1 indicates a highly clustered dataset, while a result around 0.5 indicates that the data is random.

Let $X \in R^D$ be the given dataset of N elements, and $Y \in R^D$ is the uniformly random dataset of M ($M \ll N$) elements with the same variation as X . Then we get a random sample $\{x_1^D, x_2^D, \dots, x_M^D\}$ from X . And s_i^D and t_i^D are the distances from x_i^D and y_i^D to their nearest neighbor in X , respectively. Hopkins statistic is computed as the following:

$$H = \frac{\sum_{i=1}^M t_i^D}{\sum_{i=1}^M s_i^D + \sum_{i=1}^M t_i^D} \quad (1)$$

The results are shown in Table 3. Hopkins statistic examines the datasets before applying further measurement to them. The characteristics of a clusterable dataset ($H > 0.5$) can be observed under its optimal setting in K-means clustering. Then we can select the proper number of clusters for each user group.

4.1.2 Clustering Settings.

We use the **Bayesian Information Criterion** (BIC) to determine the number of clusters for each group. Due to the high-dimensional characteristics of user embeddings, it is hard to choose the optimal K (i.e., the number of clusters) via some common k-selection techniques, like elbow method or average silhouette method. To deal with it, we use model selection technique to compare the clustering results under different K s and we choose BIC, which aims to select the model with maximum likelihood. Its revised formula for partition-based clustering suits our tasks well and there is also a penalty term avoiding overfitting in the formula.

$$BIC = \sum_{i=1}^K n_i \left(\log \frac{n_i}{N} - \frac{n_i D \log 2\pi \Sigma}{2} - \frac{D(n_i - 1)}{2} \right) - \frac{K(D+1) \log N}{2} \quad (2)$$

where the variance is defined as $\Sigma = \frac{1}{N-K} \sum_{i=1}^K \sum_{j=1}^{n_i} \|x_j - c_i\|_2$.

The K -class clustering has N points $x_j \in X^D$, c_i is the center of the i -th cluster with the size of n_i , $i = 1, \dots, K$. BIC evaluates the likelihood of different clustering settings. In our case, we use BIC to determine the number of clusters (i.e., K). The K of the maximum BIC is the optimal number of clusters. We pick the corresponding K (i.e., K^*) of the first decisive local maximum (i.e., BIC^*).

4.2 Measuring Reinforcement of User Interests

We use cluster validity [44] to compare the user embeddings of two user groups, and observe the changes in clustering through different months. Originally, this technique is known as the procedure to evaluate how the clustering algorithm performs on the given datasets. The process evaluates the results on different parameter settings via a set of cluster validity indexes. These indexes for cluster validation can be grouped into two types, internal indexes (Section 4.2.1), and external indexes (Section 4.2.2). The external indexes are based on the ground-truth clustering information, which is not always available for a dataset. On the contrary, the internal index can evaluate the clustering without knowing the optimal classification. We choose each of them to measure the temporal changes in clustering in both user groups.

4.2.1 Internal Validity Indexes.

A good clustering algorithm is required to satisfy several valid properties, such as compactness, connectedness, and spatial separation [21]. One type of internal indexes is to evaluate to what extent the clusters satisfy these properties, and a prominent example is the Calinski-Harabasz index [8]. Another type is applied to crisp clustering or fuzzy clustering [34, 47]. Since we want to explore how user interests shift at the population level, we apply the former type of internal indexes on the clustering results of the user embeddings, in order to detect the polarization tendency in user preferences by tracking how the index changes over time.

Calinski-Harabasz (CH_K) index scores the clustering considering the variation ratio between the sum-of-squares between clusters

(SSB_K) and the sum-of-squares within clusters (SSW_K) under K -class clustering. Based on this, we can compare the clustering of the same group at different times under the same setting, and a higher score indicates a better clustering. Let N denote the size of the dataset $\{x_1, \dots, x_N\}$, and K denote the number of clusters. The centroids of clusters are denoted as C_i , $i = 1, 2, \dots, K$. For a data point x_j , it belongs to a cluster p_j and we have the corresponding cluster centroid C_{p_j} , where $j = 1, 2, \dots, N$, $p_j = 1, 2, \dots, K$. The Calinski-Harabasz index is thus calculated as follows:

$$CH_K = \frac{SSB_K}{SSW_K} \cdot \frac{(N - K)}{(K - 1)} \quad (3)$$

where $SSW_K = \sum_{j=1}^N \|x_j - c_{p_j}\|^2$, $SSB_K = \sum_{i=1}^K \|c_i - \bar{X}\|^2$, and \bar{X} represents the mean of the whole dataset. Based on this definition, an ideal clustering result means that elements within a cluster congregate and elements in-between clusters disperse, leading to a high CH_K value. Intuitively, we can assign users into clusters, and calculate the CH index for this clustering result based on the user embeddings. After a certain period of time, the user embeddings would change due to the user's new interactions during the time, and we can use the new embeddings to calculate the CH index without changing the users' cluster assignment. By comparing the CH index before and after the user embeddings change, we will be able to evaluate to what extent the user preferences have changed (see Figure 3, details to be introduced later). Furthermore, based on the user IDs, we can track how each user's preference changed in the latent space.

4.2.2 External Validity Indexes.

We use external validity indexes to measure the similarity between the “first block” and the “last block” embeddings in terms of clustering. This kind of indexes utilizes the ground truth class information to evaluate the clustering results. The clustering result close to the optimal clustering has high index scores. In other words, the external indexes compute the similarity between the given clustering and the optimal clustering.

External validity indexes are constructed on the basis of contingency tables [48]. It is built as a matrix containing the interrelation between two partitions on a set of N points. Partitions, $\mathbf{P} = \{P_1, P_2, \dots, P_{K_1}\}$ of K_1 clusters and $\mathbf{Q} = \{Q_1, Q_2, \dots, Q_{K_2}\}$ of K_2 clusters, give us a $K_1 \times K_2$ contingency table (C_{PQ}) consisting of the number of common points (n_{ij}) in P_i and Q_j as follows:

$$C_{PQ} = \begin{bmatrix} n_{11} & n_{12} & \cdots & n_{1K_2} \\ n_{21} & n_{22} & \cdots & n_{2K_2} \\ \vdots & \vdots & \ddots & \vdots \\ n_{K_1 1} & n_{K_1 2} & \cdots & n_{K_1 K_2} \end{bmatrix} \quad (4)$$

where P_i and Q_j are the clusters in \mathbf{P} and \mathbf{Q} with the size of p_i and q_j , $i = 1, 2, \dots, K_1$, $j = 1, 2, \dots, K_2$. Therefore, we have $p_i = \sum_{j=1}^{K_2} n_{ij}$, $q_j = \sum_{i=1}^{K_1} n_{ij}$, and $\sum_{i=1}^{K_1} \sum_{j=1}^{K_2} n_{ij} = N$.

The techniques of comparing clusters for external validation are divided into three groups, pair-counting, set-matching, and information-theoretic [45]. We use **Adjusted Rand Index** (ARI) [25], which is a pair-counting index that counts the pairs of data points

on which two clusters are identical or dissimilar. In this way, we can evaluate the portion of users shifting to another cluster on K-means clustering over time. As a representative pair-counting based measure, ARI satisfies the three crucial properties of a clustering comparison measure [45]: metric property, normalization, and constant baseline property. It can be computed as follows:

$$ARI = \frac{\sum_{ij} \binom{n_{ij}}{2} - [\sum_i \binom{p_i}{2} \sum_j \binom{q_j}{2}] / \binom{N}{2}}{\frac{1}{2} [\sum_i \binom{p_i}{2} + \sum_j \binom{q_j}{2}] - [\sum_i \binom{p_i}{2} \sum_j \binom{q_j}{2}] / \binom{N}{2}} \quad (5)$$

ARI measures the similarity between two different clusterings (on two datasets) with the same clustering setting (K) for the same user group in our experiment. Traditionally, it evaluates the different clusterings under different clustering settings on the same dataset. The similarity between embeddings of the “first block” and the “last block” shows how much the clustering changes under the same K . The higher the ARI score, the higher the similarity between the clusterings at the beginning and the end, and fewer changes in the latent space. The difference in ARI helps us to understand how the *Following Group* acts under the influence of RS. Since the distribution of user interest embeddings would change under the influence of external conditions, such as sales campaigns and the release of new products, the group that has fewer changes implies stable interests in certain types of items, showing the tendency of reinforcement. Unlike the comparison in CH_K , a higher ARI implies that the preferences of the *Following Group* is relatively reinforced.

4.3 Measuring the Changes of Content Diversity

To answer the second question, we measure the content diversity in recommendation lists. To detect how diverse a list of recommendations is, we compute the pairwise distance of item embeddings and use the average of the item distance to represent the content diversity of the list [33]. We collect the first N items recommended to users as the first recommendation list, and the last N items as the last recommendation list. The Euclidean distance is computed between two item embeddings with the dimension of D . Let v_i be the vector of item embedding, $v_i \in R^D$, $i = 1, \dots, N$, and v_i^d is the value of v_i on the d -th dimension. Then we have the distance:

$$distance_{v_i, v_j} = \sqrt{\sum_{d=1}^D (v_i^d - v_j^d)^2} \quad (6)$$

The smaller the distance, the smaller the difference between the two items. We take the average of the pairwise Euclidean distance within the block to measure its content diversity [51], and then utilize the temporal changes in content diversity to examine the effect of the recommender system on different user groups.

5 ANALYZING ECHO CHAMBER

In this section, we first present the description of data after pre-processing in Section 5.1. Then, we present our clustering settings (i.e., the number of clusters selected) in Section 5.2. We measure the cluster validity to examine the reinforcement in the *Following Group* to answer RQ1 in Section 5.3. Moreover, we evaluate the shifts in content diversity of recommendation lists so as to answer RQ2. In the following, we refer to the user embedding based on

| Dataset | Statistic | Numerical Values |
|--------------|----------------------|------------------|
| Click Log | Num of click logs | 7, 386, 783 |
| | Num of users | 7, 477 |
| | Num of items | 2, 686, 591 |
| Purchase Log | Num of purchase logs | 98, 135 |
| | Num of users | 3, 557 |
| | Num of items | 71, 973 |
| Browse Log | Num of browse logs | 6, 225, 301 |
| | Num of users | 7, 417 |
| | Num of items | 5, 077, 268 |

Table 2: Statistics of experiment data.

clicked and purchased items as “click embedding” and “purchase embedding” respectively for simplicity.

5.1 Data Description

As what we have introduced in Section 3, experiments are performed on an extensive collection of real-world data with user click, purchase, and browse logs. The code of our entire experiment is released in our GitHub repository ¹. Meanwhile, after completing the data pre-processing in Section 3, e.g., identifying recommendation followers and creating user-item interaction blocks, the detailed statistics for the final dataset regarding each type of user-item interaction are shown in Table 2. Note that the *Following Group* (i.e., recommendation takers) and the *Ignoring Group* differ in size after the above pre-processing procedure. Since the size of the dataset would affect cluster analysis results, we need to resize the larger user group – *Following Group* – in all types of user actions, to avoid the influence. As a result, we sample from the *Following Group* to make sure the *Following Group* and the *Ignoring Group* have an equal amount of users. We apply this operation (denoted as resize-sampling) to each kind of user interaction logs and generate three pairs of equal-sized user groups for our experiments.

Additionally, we take a random sample of $p\%$ users (p -sampling) from the dataset in each computation, and repeat every experiment 50 times to calculate the statistical significance for each measure. We use 10% for p -sample in Hopkins statistic, which is large enough to represent the distribution of a dataset. We use the ratio of 80% for other indexes. We calculate the average of the 50 experiments as the final result. As a result, for the *Following Group*, it takes two steps of samples – resize-sampling and p -sampling – before each computation. For instance, we need to compute cluster indexes for the *Following Group* on the click logs 50 times. For each time, we resize the *Following Group* into 2452 users and then take another sample of 80%. The final size would be 1962 users, which is the same as the size of the *Ignoring Group* after p -sample.

5.2 Clustering Settings

5.2.1 Measuring the Cluster Tendency. We examine the clustering tendency based on Hopkins statistic for the user embeddings of both *Following Group* and *Ignoring Group*, and the results are shown in Table 3. Our results show that both groups of user embeddings are clusterable ($H > 0.5$), so cluster analysis could generate meaningful results in the following experiments. Meanwhile, we also observe that the clustering tendency for each group is not stable, and we

¹<https://github.com/szhaofelicia/EchoChamberInEcommerce>

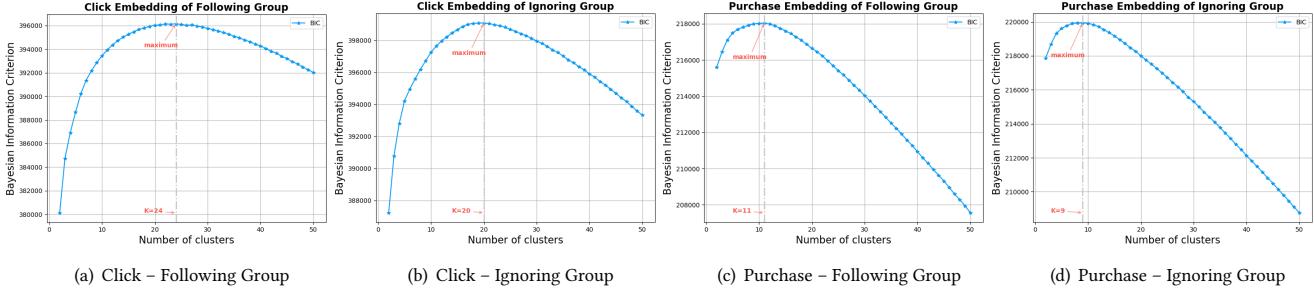


Figure 2: Bayesian Information Criterion (BIC). The K of the maximum BIC is the optimal number of clusters.

| Action | User type | Amount | First Block | Last Block | P-value |
|----------|-----------------------|--------|-------------|------------|------------|
| Click | All users | 4904 | 0.7742 | 0.7713 | $4.33e-9$ |
| | Following | 2452 | 0.7756 | 0.7746 | $3.05e-2$ |
| | Ignoring | 2452 | 0.7728 | 0.7680 | $1.53e-21$ |
| | Between-group p-value | 4904 | $1.58e-8$ | $2.88e-28$ | |
| Purchase | All users | 2916 | 0.7264 | 0.7279 | $1.41e-2$ |
| | Following | 1458 | 0.7223 | 0.7248 | $1.25e-6$ |
| | Ignoring | 1458 | 0.7305 | 0.7310 | 0.27 |
| | Between-group p-value | 2916 | $2.65e-28$ | $1.02e-24$ | |

Table 3: Hopkins statistic.

utilize p -value of the t -test to examine its statistical significance since the temporal change of H score is quite small.

Comparing the changes of H score in the first and last block, the clustering tendency decreases in click embeddings but increases in purchase embeddings. Furthermore, the clustering tendency of purchase embedding is less changeable and even slightly increased because there are fewer local shifts in the latent space of purchase embedding compared with the temporal changes in click embedding. This might attribute to the fact that users' tastes reflected in purchased items are relatively stable since users cannot choose to buy whatever they want as they need to pay the price for it.

5.2.2 Select the Number of Clusters.

Since we have shown that the *Following Group* and *Ignoring Groups* are clusterable in the previous section, we now use BIC to detect the optimal number of clusters (K^*) for each group of embeddings. The average BIC curves are plotted in Figure 2. We do not force the clustering on each dataset to use the same number of clusters in consideration of underestimation caused by inappropriate K settings. Clustering settings, such as K , have to fit the datasets well to guarantee the optimal clustering results. The inaccurate results might lead to overestimating or underestimating of the echo chamber effect.

We set the corresponding K of the maximum BIC as the optimal number of clusters K^* . The K^* 's are 24, shown in Figure 2(a), for the *Following Group*, and 20, shown in Figure 2(b), for the *Ignoring Group*, with the click embeddings. Meanwhile, K^* is 11, shown in Figure 2(c), for the *Followings Group*, and 9, shown in Figure 2(d), for the *Ignoring Group*, with the purchase embeddings. Intuitively, we could directly compare the user groups with their own K^* , but the local areas around maximum in the curves seem to be quite flat. As a result, we believe the measurements around K^* would show more reliable and plausible results than the measurement exactly at K^* . Thus, we execute the experiments in the range of

$[K^*-5, K^*+5]$. The average of results is used to examine the changes of clustering via cluster validity indexes introduced next. Finally, the *Following Group* uses K with the range of [19, 29] and [6, 16] for the click embedding and the purchase embedding respectively, and the *Ignoring Group* uses K with the range of [15, 25] and [4, 14].

5.3 Results Analysis

RQ1: Does RS reinforce user click or purchase interests?

5.3.1 Internal Validity Indexes.

As introduced in Section 4.2.1, CH can measure the extent of variation of the within-group clustering at different times. We first use CH to examine the tendency of reinforcement in user interests. The average scores of the CH index are plotted in Figure 3, we can find that both groups have a drop in CH after three months at all K s, and CH also decreases as K becomes larger both in the first blocks and the last blocks. The common decreasing trend over time in two groups might attribute to how we compute CH in the last blocks.

As we mentioned in Section 4.2.1, we assign the clustering partition results of the first blocks to the last blocks, which means that the same user will have the same cluster label both at the beginning and at the end. However, the decrease in CH suggests that the temporal shifts in the user embeddings might have made the assignment of clusters unsuitable, i.e., the ideal clustering partition of the first blocks, can no longer serve as the ideal clustering partition of the last blocks due to the temporal shifts in the user embeddings. Besides the effect of RS, the changes in user embeddings can also result from other factors. For instance, user interest can vary a lot, along with changes in external conditions in e-commerce, such as sales campaigns. As a result, these temporal shifts of user embeddings in the latent space are reflected as the decrease in CH.

In practice, we can hardly avoid this "natural" reduction caused by the e-commerce platform. As a result, we evaluate the difference between the two user groups to find the effect that comes from the RS. We compute the temporal decreases in CH for each group with K in $[K^*-5, K^*+5]$ (see Table 4). As is shown in the table, the drops of CH at K^* are 48.22 and 50.73 for the *Following Group* and *Ignoring Group* in click embedding, and the average drops of CH in $[K^*-5, K^*+5]$ are 48.41 and 50.95 respectively. Moreover, purchase embeddings have similar results that the decreases of CH for the *Following Group* are 32.74 at K^* and 33.26 for average, and the reductions for the *Ignoring Group* are 35.45 and 37.29. We further check the statistical significance of the difference between two groups, finding that all differences are at 95% confidence interval (i.e., p -value is less than 0.05). Overall, CH drops slower in *Following Groups*, showing a more stable tendency than that in *Ignoring Group*.

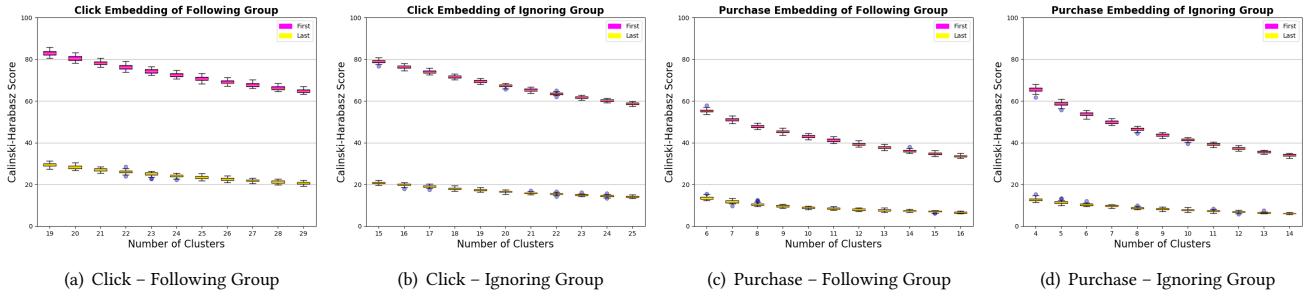


Figure 3: CH score under different K selected using BIC.

| | Click | | | Purchase | | |
|-----------|-----------|----------|----------|-----------|----------|----------|
| | Following | Ignoring | P-value | Following | Ignoring | P-value |
| $k^* - 5$ | 53.50 | 58.14 | 3.62e-41 | 41.76 | 52.72 | 6.94e-60 |
| $k^* - 4$ | 52.21 | 56.51 | 2.10e-39 | 39.30 | 47.35 | 3.54e-56 |
| $k^* - 3$ | 51.12 | 54.95 | 1.58e-40 | 37.30 | 43.41 | 1.44e-50 |
| $k^* - 2$ | 50.10 | 53.54 | 4.79e-35 | 35.60 | 40.21 | 3.50e-48 |
| $k^* - 1$ | 49.19 | 52.02 | 7.64e-31 | 34.13 | 37.70 | 7.47e-39 |
| k^* | 48.22 | 50.73 | 4.29e-28 | 32.74 | 35.45 | 3.55e-29 |
| $k^* + 1$ | 47.17 | 49.38 | 9.78e-22 | 31.34 | 33.62 | 3.96e-24 |
| $k^* + 2$ | 46.45 | 48.00 | 3.92e-15 | 30.09 | 32.06 | 8.70e-23 |
| $k^* + 3$ | 45.74 | 46.89 | 8.67e-11 | 28.88 | 30.54 | 4.99e-19 |
| $k^* + 4$ | 44.78 | 45.69 | 1.36e-7 | 27.78 | 29.14 | 2.69e-18 |
| $k^* + 5$ | 44.05 | 44.62 | 2.61e-4 | 26.97 | 28.00 | 4.78e-14 |
| AVE | 48.41 | 50.95 | 5.97e-32 | 33.26 | 37.29 | 1.06e-46 |

Table 4: Decreases in Calinski-Harabasz score. The corresponding K s for each groups are introduced in Section 5.2.2

Accordingly, the *Ignoring Group*, which falls faster in CH, disperses to a wide range on the latent space, reveals that it may receive a milder influence of reinforcement on user preference.

The less dispersion of *Following Group* in the latent space is possibly due to multiple factors. One the one hand, the *Following Group* might have more users who hold on to the previous preference in items; on the other hand, *Following Group* has fewer changes in their interests than the *Ignoring Group* does. Either of the reasons could give us the conclusion that user interest in the *Following Group* has a strengthening trend over time, resulting in that the dispersion in latent space is suppressed to some extent.

5.3.2 External Validity Indexes.

Also, we examine the temporal changes in clustering via the external validity index, ARI. Unlike CH using the same labels on two datasets, ARI compares the different clusterings of the first and the last block and measures the similarity between clusterings. We plot the similarities (ARI) at different K s in Figure 4 and list the average results of ARI in Table 5. We find that in the click embedding, the *Following Group* has a higher ARI than the *Ignoring Group* (average ARI of 0.0986 and 0.0765 respectively with the p -value of 2.28e-51). In the purchase embedding, the difference between the two groups is not statistically significant; the p -value for the difference of the average ARI is 0.53. Similar observation is also shown in the curves in Figure 4, the curve in Figure 4(a) is higher than the curve in Figure 4(b), but curves in Figure 4(c) and Figure 4(d) almost overlap. Let us take a look at each pair of ARI of different user groups

| | Click | | | Purchase | | |
|-----------|-----------|----------|----------|-----------|----------|----------|
| | Following | Ignoring | P-value | Following | Ignoring | P-value |
| $k^* - 5$ | 0.1136 | 0.0877 | 1.34e-33 | 0.0969 | 0.0829 | 3.70e-8 |
| $k^* - 4$ | 0.1099 | 0.0856 | 1.85e-32 | 0.0825 | 0.0677 | 1.67e-10 |
| $k^* - 3$ | 0.1060 | 0.0829 | 1.23e-33 | 0.0725 | 0.0686 | 2.77e-2 |
| $k^* - 2$ | 0.1040 | 0.0811 | 4.19e-34 | 0.0697 | 0.0712 | 0.35 |
| $k^* - 1$ | 0.0996 | 0.0780 | 2.88e-32 | 0.0639 | 0.0650 | 0.52 |
| k^* | 0.0974 | 0.0756 | 5.52e-32 | 0.0615 | 0.0635 | 0.23 |
| $k^* + 1$ | 0.0949 | 0.0734 | 4.49e-41 | 0.0585 | 0.0634 | 9.17e-4 |
| $k^* + 2$ | 0.0929 | 0.0717 | 1.62e-33 | 0.0555 | 0.0598 | 4.12e-3 |
| $k^* + 3$ | 0.0900 | 0.0698 | 1.13e-35 | 0.0524 | 0.0583 | 1.88e-5 |
| $k^* + 4$ | 0.0890 | 0.0687 | 1.10e-39 | 0.0511 | 0.0552 | 1.48e-3 |
| $k^* + 5$ | 0.0871 | 0.0670 | 4.20e-34 | 0.0489 | 0.0510 | 0.13 |
| AVE | 0.0986 | 0.0765 | 2.28e-51 | 0.0648 | 0.0642 | 0.53 |

Table 5: ARI scores. Same as CH_K , the corresponding K s for each groups are introduced in Section 5.2.2

in purchase embedding, the differences among half of the K s in $[K^* - 5, K^* + 5]$ are not significant.

To sum up, the *Following Group* has fewer temporal shifts in click embedding but no evident difference in purchase embedding. In other words, in terms of click interests, partitions at the beginning and the end in the *Following Group* are more similar, indicating more connections. While in purchase interests, clustering at the end in both groups does not show the trend of sticking to the previous clustering. More changes appearing in both groups in purchase embedding could be caused by the fact that users have fewer choices to purchase because of objective constraints, such as their incomes and the item prices. The effect of RS cannot “force” users to buy some items they cannot afford, thus, even if user interest has been reinforced, the shifts in preferences might not appear in purchase behaviors. However, in click behaviors, the *Following Group* seems to strengthen their preference under the effect of RS, since users are free to click items they are interested in, and their interests presented in click embedding do not have any other constraints. The group with higher ARI has fewer changes, which means they stick to the items they interacted with before and intensify their preferences. The evidence confirms the conclusion in Section 5.3.1 that there exists the tendency of reinforcement in user interests in the *Following Group*.

RQ2: If user interests are strengthened, is it caused by RS narrowing down the scope of items exposed to users?

After an affirmative answer to RQ1, we examine RQ2 to explore the potential cause of the reinforcement in user interests: narrowed

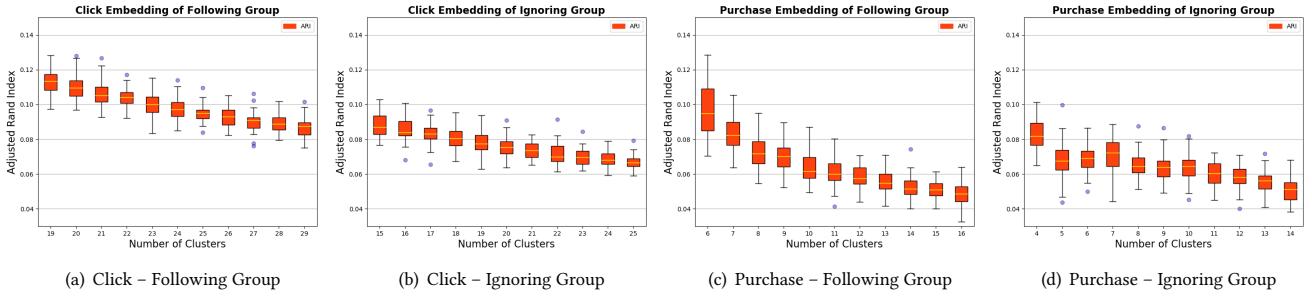


Figure 4: ARI under different K s, which are selected by BIC.

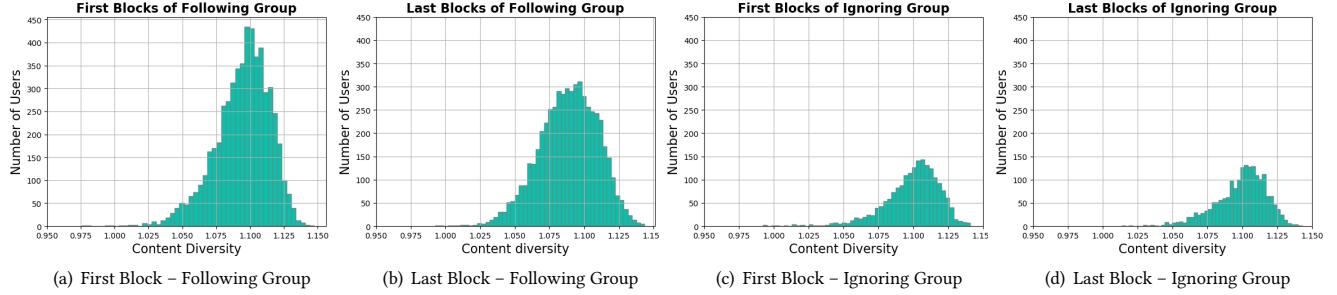


Figure 5: Recommendation content diversity (pairwise distance between user embeddings) in two user groups.

contents offered to users. To do so, we measure the content diversity of recommendation lists at the beginning and the end. The distributions of content diversity (the average pairwise distance of item embeddings) in the first and last blocks for display are plotted in Figure 5, and the corresponding average content diversities are listed in Table 6. These distributions are approximately normal, and the first blocks have a larger density around higher content diversity than the last blocks do. Also, the distribution becomes dispersing over time, lowering the average of the whole group. Furthermore, the average content diversity gives us consistent observation. When paying attention to the overall temporal changes, we find that the content diversity among all users falls from 1.0960 to 1.0937 with p -value of 6.10e-11. Even though the drop is tiny, it indicates that both groups go through the trend of narrowing down the scope of the content displayed to users. Additionally, the content diversities in the *Following Group* have a larger reduction from 1.0945 to 1.0882 than the reduction in *Ignoring Group*. On the contrary, the decrease in the *Ignoring Group* can even be ignored because of the high p -value of 0.67. This is because RS learns more about followers from their interactions, such as click, purchase. Thus, it is more likely for RS to provide items similar to what users have previously interacted with.

In e-commerce, user affects recommendations exposed to them through user actions, and their actions get influenced in return, these procedures form a feedback loop, which strengthens the personalized recommendation and shrinks the scope of the content offered to users. As a result, the filter bubble effect occurs in *Following Group*. Conversely, *Ignoring Group* only provide minimal information about their tastes in items, since they do not interact with RS much. Hence, RS recommends items from a broad scope to explore the users' preferences. The difference between the two groups is also statistically significant at all times. The *Ignoring Group* has a

| | Amount | First | last | Within-group p-value |
|-----------------------|--------|----------|----------|----------------------|
| All users | 3820 | 1.0969 | 1.0937 | 6.10e-11 |
| Following group | 1910 | 1.0945 | 1.0882 | 1.95e-20 |
| Ignoring group | 1910 | 1.0992 | 1.0989 | 0.67 |
| Between-group p-value | 3820 | 2.13e-12 | 2.16e-56 | |

Table 6: The content diversity of recommended items.

higher diversity of 1.0992 in the beginning, and the difference is further enlarged in the end since the content diversity in *Following Group* drops a lot. Then, the scope of items recommended to the *Following Group* has been repeatedly narrowed down, strengthening user interests in this group as a consequence.

As we claimed in answer to RQ1, the reinforcement in preferences — echo chamber effect — is reflected in the temporal shifts of user embeddings in clustering. Particularly, echo chamber appears in both user click interests and user purchase interests, but the effect in the latter is sort of slight. However, in other RS platforms, such as movie recommendations [33], opposite observations appear in the *Following Group*, indicating that RS helps users explore more items and mitigate the reduction in content diversity. One possible reason is that unlike products in e-commerce, promotional campaigns for movies mostly focus on those commercial films. Therefore, movie recommendation platforms could still fill the recommendation list with niche movies and slow down the reduction in content diversity.

6 CONCLUSIONS AND FUTURE WORK

In this paper, we examine and analyze echo chamber effect in a real-world e-commerce platform. We found that the tendency of echo chamber exists in personalized e-commerce RS in terms of user click behaviors, while on user purchase behaviors, this tendency is mitigated. We further analyzed the underlying reason for the

observations and found that the feedback loop exists between users and RS, which means that the continuous narrowed exposure of items raised by personalized recommendation algorithms brings consistent content to the *Following Group*, resulting in the echo chamber effect as a reinforcement of user interests. This is one of our first steps towards socially responsible AI in online e-commerce environments. Based on our observations and findings, in the future, we will develop refined e-commerce recommendation algorithms to mitigate the echo chamber effects, so as to benefit online users for more informed, effective, and friendly recommendations.

REFERENCES

- [1] David Paul Allen, Henry Jacob Wheeler-Mackta, and Jeremy R Campo. 2017. The Effects of Music Recommendation Engines on the Filter Bubble Phenomenon. (2017).
- [2] Arda Antikacioglu and Ravi. 2017. Post processing recommender systems for diversity. In *KDD*. ACM, 707–716.
- [3] Mahsa Badami, Olfa Nasraoui, and Patrick Shafto. 2018. PrCP: Pre-recommendation Counter-Polarization. In *KDIR*. 280–287.
- [4] E. Bakshy, S. Messing, and L. A. Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348, 6239 (may 2015), 1130–1132.
- [5] Amit Banerjee and Raiesh N. Dave. 2004. Validating clusters using the Hopkins statistic. *IEEE International Conference on Fuzzy Systems* 1 (2004), 149–153.
- [6] Dimitrios Bountouridis, Jaron Harambam, Mykola Makhortykh, Mónica Marrero, Nava Tintarev, and Claudia Hauff. 2019. SIREN: A Simulation Framework for Understanding the Effects of Recommender Systems in Online News Environments. In *ACM Conference on Fairness, Accountability, and Transparency*. ACM, 150–159.
- [7] Laura Burbach, Patrick Halbach, Martina Ziefle, and André Calero Valdez. 2019. Bubble Trouble: Strategies Against Filter Bubbles in Online Social Networks. In *International Conference on Human-Computer Interaction*. Springer, 441–456.
- [8] Tadeusz Caliński and Harabasz JA. 1974. A Dendrite Method for Cluster Analysis. *Communications in Statistics - Theory and Methods* 3 (01 1974), 1–27.
- [9] Allison J.B. Chaney, Brandon M. Stewart, and Barbara E. Engelhardt. 2018. How Algorithmic Confounding in Recommendation Systems Increases Homogeneity and Decreases Utility. *RecSys* (2018).
- [10] James N Cohen. 2018. Exploring Echo-Systems: How Algorithms Shape Immersive Media Environments. *Journal of Media Literacy Education* 10, 2 (2018), 139–151.
- [11] Pranav Dandekar, Ashish Goel, and David T Lee. 2013. Biased assimilation, homophily, and the dynamics of polarization. *Proceedings of the National Academy of Sciences* 110, 15 (2013), 5791–5796.
- [12] Abhishek Dash, Animesh Mukherjee, and Saptarshi Ghosh. 2019. A Network-centric Framework for Auditing Recommendation Systems. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE, 1990–1998.
- [13] Carlos Manuel Moita de Figueiredo. 2014. Emotions and recommender systems: A social network approach. (2014).
- [14] Seth Flaxman, Sharad Goel, and Justin M Rao. 2016. Filter bubbles, echo chambers, and online news consumption. *Public opinion quarterly* 80, S1 (2016), 298–320.
- [15] Daniel M Fleder and Kartik Hosanagar. 2007. Recommender systems and their impact on sales diversity. In *ACM conference on Electronic commerce*. ACM.
- [16] Zuohui Fu, Yikun Xian, Ruoyuan Gao, Jieyu Zhao, Qiaoying Huang, Yingqiang Ge, Shuyuan Xu, Shijie Geng, Chirag Shah, Yongfeng Zhang, and Gerard de Melo. 2020. Fairness-Aware Explainable Recommendation over Knowledge Graphs. In *Proceedings of SIGIR 2020*. ACM, New York, NY, USA.
- [17] Mingkun Gao, Hyo Jin Do, and Wai-Tat Fu. 2018. Burst Your Bubble! An Intelligent System for Improving Awareness of Diverse Social Opinions. In *23rd International Conference on Intelligent User Interfaces*. ACM, 371–383.
- [18] Yingqiang Ge, Shuyuan Xu, Shuchang Liu, Zuohui Fu, Fei Sun, and Yongfeng Zhang. 2020. Learning Personalized Risk Preferences for Recommendation. *SIGIR* (2020).
- [19] Yingqiang Ge, Shuyuan Xu, Shuchang Liu, Shijie Geng, Zuohui Fu, and Yongfeng Zhang. 2019. Maximizing marginal utility per dollar for economic recommendation. In *The World Wide Web Conference*. 2757–2763.
- [20] Daniel Geschke, Jan Lorenz, and Peter Holtz. 2019. The triple-filter bubble: Using agent-based modelling to test a meta-theoretical framework for the emergence of filter bubbles and echo chambers. *British Journal of Social Psychology* 58, 1 (2019), 129–149.
- [21] Julia Handl, Joshua Knowles, and Douglas B. Kell. 2005. Computational cluster validation in post-genomic data analysis. *Bioinformatics* 21, 15 (08 2005), 3201–3212.
- [22] Natali Helberger, Kari Karppinen, and Lucia Dá ÁZacunto. 2018. Exposure diversity as a design principle for recommender systems. *Information, Communication & Society* 21, 2 (2018), 191–207.
- [23] Martin Hilbert, Saifuddin Ahmed, Jaeho Cho, Billy Liu, and Jonathan Luu. 2018. Communicating with algorithms: a transfer entropy analysis of emotions-based escapes from online echo chambers. *Communication Methods and Measures* 12, 4 (2018), 260–275.
- [24] Kartik Hosanagar, Daniel Fleder, Dokyun Lee, and Andreas Buja. 2013. Will the global village fracture into tribes? Recommender systems and their effects on consumer fragmentation. *Management Science* 60, 4 (2013), 805–823.
- [25] Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of Classification* 2, 1 (01 Dec. 1985), 193–218.
- [26] Ray Jiang, Silvia Chiappa, Tor Lattimore, András György, and Pushmeet Kohli. 2019. Degenerate feedback loops in recommender systems. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 383–390.
- [27] Xuehan Jiang. 2018. A community-evolution based approach for detecting the echo chamber effect in recommender systems. (2018).
- [28] Lanu Kim, JD West, and Katherine Stovel. 2017. Echo chambers in science?. In *American Sociological Association (ASA) Annual Meeting*.
- [29] Richard G. Lawoski and Peter C. Jurs. 1990. New index for clustering tendency and its application to chemical problems. *Journal of Chemical Information and Computer Sciences* 30, 1 (1990), 36–41.
- [30] Sina Mohseni and Eric Ragan. 2018. Combating Fake News with Interpretable News Feed Algorithm. *arXiv preprint arXiv:1811.12349* (2018).
- [31] Judith Möller, Damian Trilling, Natali Helberger, and Bram van Es. 2018. Do not blame it on the algorithm: an empirical assessment of multiple recommender systems and their impact on content diversity. *Information, Communication & Society* 21, 7 (2018), 959–977.
- [32] CHA Namjun, CHO Hosoo, LEE Sangman, and Junseok HWANG. 2019. Effect of AI Recommendation System on the Consumer Preference Structure in e-Commerce: Based on Two types of Preference. In *2019 21st International Conference on Advanced Communication Technology (ICACT)*. IEEE, 77–80.
- [33] Tien T Nguyen, Pik-Mai Hui, F Maxwell Harper, Loren Terveen, and Joseph A Konstan. 2014. Exploring the filter bubble: the effect of using recommender systems on content diversity. In *Proceedings of the 23rd WWW*. ACM, 677–686.
- [34] Malay Kumar Pakhira, Sanghamitra Bandyopadhyay, and Ujjwal Maulik. 2004. Validity index for crisp and fuzzy clusters. *Pattern Recognit.* 37 (2004), 487–501.
- [35] Zachary A. Pardos and Weijie Jiang. 2019. Combating the Filter Bubble: Designing for Serendipity in a University Course Recommendation System. *IntRS’19: Joint Workshop on Interfaces and Human Decision Making for Recommender Systems* (2019).
- [36] Eli Pariser. 2011. *The filter bubble: What the Internet is hiding from you*. Penguin UK.
- [37] Marten Risiis, Okan Aydinguel, and Maximilian Haug. 2019. TOWARDS AN UNDERSTANDING OF CONSPIRACY ECHO CHAMBERS ON FACEBOOK. (2019).
- [38] Kazutoshi Sasahara, Wen Chen, Hao Peng, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. 2019. On the Inevitability of Online Echo Chambers. *arXiv preprint arXiv:1905.03919* (2019).
- [39] Ana-Andrea Stoica and Augustin Chaintreau. 2019. Hegemony in Social Media and the effect of recommendations. In *Companion Proceedings of The 2019 World Wide Web Conference*. ACM, 575–580.
- [40] Cass R. Sunstein. 2007. *Republic.com 2.0*. Princeton University Press. <http://www.jstor.org/stable/j.ctt7bsw>
- [41] Cass R Sunstein. 2009. *Going to extremes: How like minds unite and divide*. Oxford University Press.
- [42] Cass R Sunstein. 2018. *Republic: Divided democracy in the age of social media*. Princeton University Press.
- [43] Nava Tintarev, Shahin Rostami, and Barry Smyth. 2018. Knowing the unknown: visualising consumption blind-spots in recommender systems. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*. ACM, 1396–1399.
- [44] Michalis Vazirgiannis. 2009. *Clustering Validity*. Springer US, Boston, MA, 388–393. https://doi.org/10.1007/978-0-387-39940-9_616
- [45] Nguyen Xuan Vinh, Julien Epps, and James Bailey. 2009. Information Theoretic Measures for Clusterings Comparison: Is a Correction for Chance Necessary?. In *Proceedings of the 26th ICML (ICML ’09)*. ACM, New York, NY, USA, 1073–1080.
- [46] Jizhe Wang, Pipei Huang, Huan Zhao, Zhibo Zhang, Binqiang Zhao, and Dik Lun Lee. 2018. Billion-scale Commodity Embedding for E-commerce Recommendation in Alibaba. *Proceedings of the 24th ACM SIGKDD* (2018).
- [47] Weinan Wang and Yunjie Zhang. 2007. On fuzzy cluster validity indices. *Fuzzy Sets and Systems* 158 (10 2007), 2095–2117. <https://doi.org/10.1016/j.fss.2007.03.004>
- [48] Junjie Wu, Jian Chen, Hui Xiong, and Ming Xie. 2009. External validation measures for K-means clustering: A data distribution perspective. 36 (04 2009), 6050–6061.
- [49] Yongfeng Zhang, Min Zhang, Yiqun Liu, Shaoping Ma, and Shi Feng. 2013. Localized matrix factorization for recommendation based on matrix block diagonal forms. In *Proceedings of the 22nd international conference on World Wide Web*. ACM, 1511–1520.
- [50] Guanjie Zheng, Fuzheng Zhang, Zihan Zheng, Yang Xiang, Nicholas Jing Yuan, Xing Xie, and Zhenhui Li. 2018. DRN: A deep reinforcement learning framework for news recommendation. In *Proceedings of the 2018 World Wide Web Conference*. 167–176.

- [51] C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen. 2005. Improving recommendation lists through topic diversification.. In *Proceedings of the 14th international conference on World Wide Web*. ACM, 23–32.