Quantifying User Agency vs. Algorithmic Influence in Movie Recommender Systems

Marco Conti
University of Illinois
Chicago, Illinois, USA

Sai Shridhar Balamurali
University of Illinois
Chicago, Illinois, USA

Nitheesh Mannava
University of Illinois
Chicago, Illinois, USA

ABSTRACT

This study examines the impact of user choices and algorithmic recommendations on the formation of filter bubbles in movie recommendation systems. We conducted an extensive literature review, presenting a formal definition of filter bubbles and the most critical studies on the topic. We then performed a simulation study using a collaborative filtering recommendation system trained on the MovieLens dataset to evaluate the impact of algorithmic recommendations and user choices on narrowing content exposure. In the second part of the paper, we propose a method to compute the average treatment effect of following recommendations on the homogeneity of future recommendations and watch history. Our results show that the algorithm plays an 80% role in the formation of inter-user filter bubbles, while the user plays a limited but still significant 20% role.

1 INTRODUCTION

Recommended systems have become integral to our digital experiences. These systems leverage user data and algorithms to provide personalized suggestions to enhance the user experience. While they offer convenience, concerns have been raised about "filter bubbles" where users' choices are isolated from diverse perspectives, potentially reinforcing biases and limiting exposure to new ideas [12]. As the influence of recommender systems grows, a critical question emerges: Are the recommendation algorithms truly responsible for these filter bubbles, or are the people to blame?

Our research focuses on collaborative filtering-based recommendation systems for movie streaming platforms. In this field, a filter bubble has a deep ethical impact because it reduces exposure to new content, leading to radicalization, lack of creativity, narrowing of user taste, and a reduction in common discussion ground. Furthermore, as many movies are now developed to match users' tastes, this limits innovation in the entire field of film production, also impacting users who do not use recommendation systems. There are also economic downsides to the "lack of diversity" generated by a filter bubble:

- The user may get bored and leave the platform.
- The collected data, which may be used for advertisements, are not accurate.

So, if the important ethical reasons are not enough to convince platforms to introduce systems that limit the filter bubble effect, at least the economic reasons should.

Our research aims to quantify user agency versus algorithmic influence in filter bubble formation using a causal approach. While many studies attempt to detect the presence of filter bubbles, to the best of our knowledge, none have addressed this issue using a causal approach.

In this research, we have analyzed the collaborative filtering recommendation algorithm, one of the most common recommendation algorithm, used in MovieLens. In the first part, we perform a simulation to analyze the effect of the algorithm on both random users and a "realistic" users. The random user represents a user with no impact on the recommendations, while ChatGPT is used to simulate the realistic user. The idea is to observe whether different types of users are impacted differently by the algorithm.

In the second approach, we analyze the MovieLens dataset to collect data about two subsets of users: those who tend to follow algorithmic recommendations and those who do not. We use ATE estimation to discern the role of "following the recommendation" in filter bubble formation. We found that this behavior does not reduce, but rather increases the diversity of recommendations and watch history, though it makes them more similar to those of other users. This leads us to extend the definition of a filter bubble, which can also occur when the same diverse contents are served to all users.

2 RELATED WORK

Eli Pariser first introduced the term "filter bubble" in his book *The Filter Bubble: What the Internet Is Hiding from You* [13]. He described how popular social media platforms and search engines employ algorithms that display content based on users' existing behaviours and beliefs, effectively creating a "bubble" around each individual. These filter bubbles have a feedback loop that reinforces preexisting views, leading to isolation, radicalization, and a loss of creativity. One major problem with filter bubbles is their invisibility, which makes users unaware of their isolated condition.

One of the consequences of these filter bubbles is their potential to give rise to or sustain an echo chamber. According to the literature review by Argueda et al. [3], an echo chamber is a situation of limited exposure or confinement to new ideas or content, influenced either by media supply or by the user's own preferences. By contrast, a filter bubble is an echo chamber created by a ranking algorithm without any active choice by the user. People themselves may be blamed for filter bubble formation: they seek information that aligns with their views — whether or not it is factually valid — getting trapped in an echo chamber where opposing views are excluded or minimized.

Filter bubbles are a complex phenomenon explored by numerous researchers, sometimes with contradictory results. Nguyen et al. [11] were among the first to study the presence and effects of filter bubbles in movie recommendation systems. By analyzing the MovieLens dataset, they found that the narrowing effect of the recommendation system is statistically significant but small in value and that the population that follows the recommendations reports an overall more positive experience. They developed a metric to

assess content diversity in the movie domain based on the Tag Genome, a system of tags and relevance scores, designed by Vig et al. [14] to facilitate user exploration and integrated into MovieLens. The Tag Genome is a set of 1,028 of the most commonly used attributes for movies descriptions. Specifically, a relevance score for each attribute is assigned to each movie. The idea of Nguyen et al. is to use the difference between the ratings of each attribute to define a concept of distance between two films.

This of Ngugyen et al is one of several studies suggesting that the effect of filter bubbles is not as severe as initially thought. In "Eli Pariser is wrong," Linden et al. [8] argue that recommender systems are essential for improving serendipitous encounters with new items, as users cannot discover an item they are unaware of. Another study by Bakshy et al. [2] demonstrated that individual choice plays a more vital role than Facebook's ranking algorithm in limiting exposure to cross-cutting content. However, these studies have limitations. For instance, Bakshy's study examines an already polarized audience, typically less susceptible to algorithmic influences. This is why further research is necessary to validate or challenge these findings across different scenarios and with varying assumptions.

As an example of an opposing result, we reference the research by Chaney et al. [1]. They ran simulations and measured, through the Jaccard index, the amount of homogeneity in collaborative filtering recommendation systems, showing that they homogenize user behaviour beyond what is necessary to increase utility. Additionally, they draw parallels with the explore/exploit paradigm and warn against using confounded data when selecting recommender systems. For example, the MovieLens dataset is biased toward collaborative filtering algorithms.

Another study is by Lunardi et al. [10], who introduce a metric for homogeneity, the homogeneity level (HL), study filter bubble formation, and propose diversification to reduce this effect. However, their study claims that the reduction of homogeneity due to diversification is not statistically significant. Liu et al. [9] developed an innovative news recommendation system with a transparent approach to making users aware of the presence of a filter bubble. Jiang et al. [7] introduced a novel approach to distinguish between an echo chamber and filter bubble effects. They approximate user interests, showing they tend to degenerate under mild conditions. They also conducted simulation experiments on degeneration in various recommender systems and proposed continuous exploration and an expanding candidate pool as the only solutions to reduce this effect.

In their work, "Causal Inference in Recommender Systems: A Survey and Future Directions," Gao et al. [4] provide a survey of causal investigations within recommender systems. Notable related applications of causal inference to filter bubbles include the work of Wang et al. [15], who apply the backdoor adjustment method to alleviate the filter bubble effect, and Xu et al. [16], who employ counterfactual reasoning to mitigate the echo chamber effect. To the best of our knowledge, no significant studies have analyzed the role of algorithms in filter bubble formation using causal inference.

3 PROBLEM DESCRIPTION

A filter bubble occurs when algorithms automatically narrow the range of recommended content, creating self-reinforcing patterns or feedback loops. The homogeneity of algorithm-recommended items can be used to detect filter bubbles. Following Lunardi et al. [10], a filter bubble is characterized by a low level of homogeneity in the recommended set, reflecting the narrowing of diversity in proposed items.

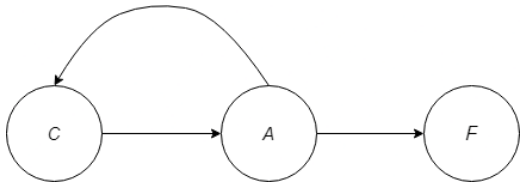

Figure 1: Structural Causal Model For Movie Recommender Systems. C represents the User choice, A represents the recommendation algorithm and F represents the presence of a filter bubble. We do not represent U and R as these are just resultants of the choice function and the algorithm which can also be considered a function.

Let U_t represent user's movie history at time step t, containing the set of movies $M_t = \{m_1, \ldots, m_k\}$ that have been viewed and rated. In collaborative filtering systems, the set of all user histories $\mathcal{H}_t = \{U_t^1, \ldots, U_t^n\}$ forms the foundational input for the algorithm, denoted as $\mathcal{A}$, which generates the recommendation list $R_t = \mathcal{A}(U_t)$.

To quantify the formation of filter bubbles, we introduce the heterogeneity score $He(R_t)$ (detailed in Section 5.1), where lower values of $He(R_t)$ indicate bubble formation, and the homogeneity score $Ho(R_t)$ (detailed in Section **??**) to detect inter-users similarity. The emergence of a filter bubble is functionally dependent on R_t, which is in turn determined by $\mathcal{A}$. Thus, U_t influences the filter bubble formation exclusively through $\mathcal{A}$.

The temporal dynamics of this system can be expressed as follows:

$$U_{t+1} = f(U_t, C_t(R_t)) \tag{1}$$

where C_t represents the user's choice function at time t, influenced by R_t. This creates a temporal feedback loop where recommendations at time $t-1$ influence user choices C_{t-1}, which subsequently affect U_t and, consequently, R_t. This recursive process is visualized in the causal diagram presented in Fig. 2.

To quantify the causal contributions of the user choice C and the algorithm $\mathcal{A}$ in filter bubble formation, we conduct two experimental simulation analyses:

1. *User-Algorithm Interaction in a Controlled Environment*: We implement a controlled simulation environment for the interaction between the user and the recommender system, analyzing the causal effects of C and $\mathcal{A}$ using both direct and indirect effects.
2. *The Effect of Following Recommendations*: We recover the users' ratings from MovieLens. By simulating the recommendations at the time of the ratings, we split the users into two groups: those who follow the recommendations

and those who do not. We then analyze the consequences of these behaviors on future recommendations and the movies watched.

4 DATA DESCRIPTION

This project is based on the "MovieLens-small-latest" dataset, gathered from MovieLens, a movie recommendation and rating website that implements a collaborative filtering-based algorithm [5]. MovieLens-small-latest is a widely used benchmark dataset for recommender systems research and testing. The `"ratings"` table contains 100,836 movie ratings, including `"movieId"`, `"userId"`, `"timestamp"`, and `"rating"` values ranging from 0 to 5, while the `"movies"` table maps `"movieId"` to movie names. The dataset also includes the tag genome, consisting of 1,128 attributes used for movie descriptions. The `"genome_score"` table provides the relevance of each tag for each movie, including `movieId`, `tagId`, and a relevance score from 0 to 1. The association between `"tagId"` and `"tag name"` is stored in the `"genome_tag"` table.

Along with this observational data, we performed several simulations employing a collaborative filtering algorithm trained on the MovieLens dataset to gather additional data under intervention for causal effect estimation.

5 METRIC

A filter bubble reduces the diversity of recommendations. To identify filter bubbles, we have defined two metrics to compute the amount of diversity.

The first metric is the heterogeneity score, which computes the amount of diversity of movies within a recommendation. The second is the homogeneity score, which is used to compute the amount of homogeneity of a single recommendation with respect to the other recommendations.

5.1 Heterogeneity Score

The heterogeneity score $He(R_k)$ is the average of the tag genome distance, as defined by Nguyen et al. [11], between all pairs of movies in the recommendation R_k:

$$He(R_k) = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=i+1}^{N} d_{(m_i,m_j)} \tag{2}$$

Here N is the length of the recommendation list $|R_k|$ and $d_{(m_i,m_j)}$ is the tag genome distance between movies m_i and m_j.

The tag genome, included in the MovieLens dataset, is a set of 1028 attributes. These attributes represent the most common keywords or short phrases used to describe the films. Specifically, based on the movies' descriptions, a score is assigned at each attribute for each movie.

The tag genome distance $d_{(m_i,m_j)}$ between two movies is sum of the squared differences between the rating of their attribute:

$$d_{(m_i,m_j)} = \sqrt{\sum_{k=1}^{|t|} \left[\text{rel}(t_k, m_i) - \text{rel}(t_k, m_j)\right]^2} \tag{3}$$

Here, $\text{rel}(t_k, m_i)$ represents the relevance score of tag t_k for movie m_i, and $|t|$ is the total number of tags.

5.2 Homogeneity score

The heterogeneity score focuses on the internal diversity within a recommendation list. On the contrary, the homogeneity score measures the extent to which a recommendation resembles those of other users.

Given that recommendations have length K, the idea is to compute the $topK$ movies, i.e., the movies that appear with higher frequency in the recommendation list. To do so, we count the occurrences of each film in the final recommendation list of all the users, selecting the K movies with the highest count.

The homogeneity score $Ho(R_i)$ for the recommendation R_i is the percentage of the number of films in common with the $topK$ list:

$$Ho(R_i) = \frac{|R_i \cap \text{topK}|}{|R_i|} \tag{4}$$

6 USER-ALGORITHM INTERACTION IN A CONTROLLED ENVIRONMENT

Let B_t denote the $H(R_t)$ at time t. We decompose the effect on B_t into:

(1) Direct Effect: $DE(\mathcal{A}) = \mathbb{E}[B_t|do(\mathcal{A} = a)] - \mathbb{E}[B_t|do(\mathcal{A} = a_0)]$
(2) Indirect Effect: $IE(C) = \mathbb{E}[B_t|C = c, \mathcal{A}] - \mathbb{E}[B_t|C = c_0, \mathcal{A}]$

where c_0 represents a baseline choice function, a_0 represents a baseline recommender algorithm and $do(.)$ denotes the causal intervention operator.

The direct effect quantifies the immediate impact of a system component on recommendation diversity while maintaining other factors constant. The indirect effect, conversely, captures how components influence bubble formation through their interactions with other system elements.

6.1 Setup

Experiment 1 investigated the relative impact of user choice function C and recommendation algorithm $\mathcal{A}$ on filter bubble emergence. We simulated $n = 5$ virtual users interacting with a movie recommendation system over $T = 10$ timesteps. At each timestep t, the system generated recommendation lists R_t^i of size $|R_t^i| = 10$ for each user i. We recorded user movie choices C_t^i, ratings, and recommendations R_t^i. The heterogeneity score $H(R_t^i)$ quantified potential filter bubble formation by measuring recommendation diversity.

The experiment comprised three conditions:

(1) Random recommendation algorithm (a_0) with preference-based user choice (c_p) [NC]
(2) Collaborative filtering algorithm (a_c) with preference-based user choice (c_p) [AC]
(3) Collaborative filtering algorithm (a_c) with random user choice (c_0) [AR]

To simulate C_p, we employed GPT-4o-mini language model $\mathcal{G}$. For each user i, we initialized their history U_0^i with viewing and rating data from 30 movies in the MovieLens dataset: $U_0^i = \{(m_k, r_k)\}_{k=1}^{30}$, where m_k represents a movie and r_k its rating. The

first instance of $\mathcal{G}$ generated a user profile $P^i = \mathcal{G}(U_0^i)$. A second instance used this profile to implement C_p:

$$c_p(R_t^i, P^i) = (m^*, r^*) \tag{5}$$

where (m^*, r^*) represents the chosen movie and rating. For c_0, users selected randomly from R_t^i with uniformly distributed ratings.

The random recommendation algorithm a_0 generated R_t^i by sampling uniformly from the movie database. By comparing $H(R_t^i)$ across conditions over T timesteps, we aimed to decompose the direct effect of $\mathcal{A}$ from the indirect effect of C on R_t^i, illuminating their relative contributions to filter bubble formation in a controlled environment.

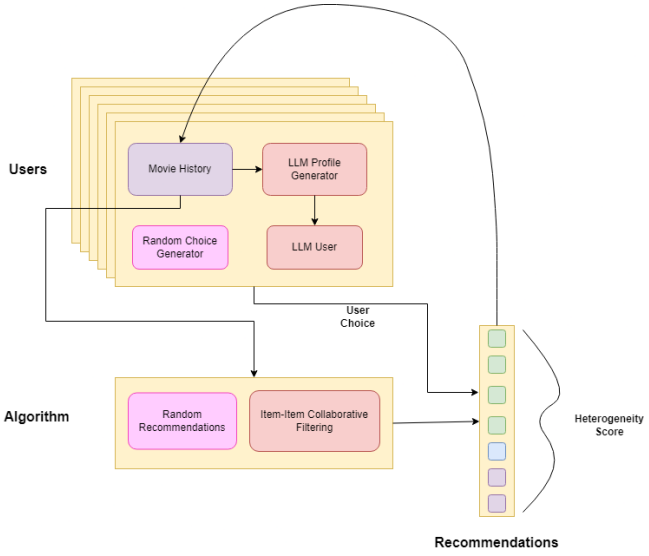

Figure 2: System Diagram

6.2 Result

We observed that the simulation combining the recommendation algorithm and the user simulated with ChatGPT (referred to in our simulation as AC) resulted in the lowest diversity of the watched movies. In contrast, randomizing user behavior while still using the algorithm (referred to as AR) increased diversity to some extent. This makes sense, as real users will choose films based on their own tastes, typically selecting only certain types of movies. We observed the highest diversity when no recommendation algorithm was used for movie suggestion, and users had complete freedom of choice. However, there are not enough differences in the diversity of the watched history across the three methods to provide significant insights.

Further analysis of homogeneity scores across the different methods revealed that AC exhibited the widest extremes and the lowest mean score. In contrast, AR and NC (no recommendation algorithm, with users simulated using ChatGPT) had the highest homogeneity scores. While this might seem counterintuitive, the lack of active choice in AR and the absence of a recommendation algorithm in

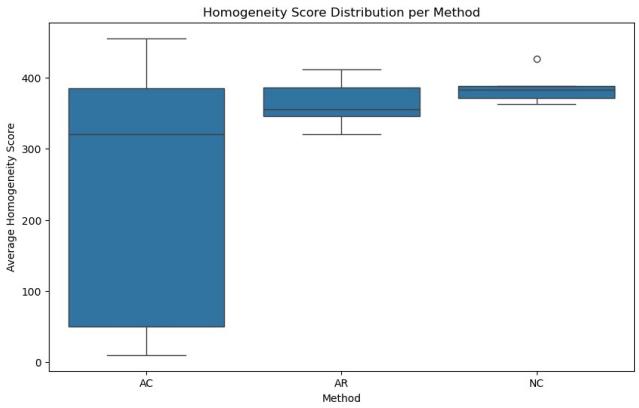

Figure 3: Homogeneity score for the three methods

NC makes users more susceptible to the intrinsic biases present in the MovieLens dataset, as noted by Chaney et al. [1].

Genre diversity and entropy analysis reinforced this observation. AR had moderate genre diversity and entropy scores, while user agency in AR introduced more entropy and diversity, although the algorithm's decisions in AR dominated. NC exhibited the highest combined genre diversity and entropy scores. The high homogeneity score indicated that, without an optimization system, users tended to stay within their natural boundaries, raising questions about whether other factors influenced this behavior.

A more detailed comparison using ANOVA, HSD scores, and backdoor estimation demonstrated that the type of algorithm played a more significant role in determining recommendation diversity than user agency. Without a recommendation algorithm, the homogeneity score dropped by an average of 82 points.

To validate these results, we conducted refutation and placebo analyses on the detected backdoor expressed model. The refutation analysis confirmed that adding random values to the data did not significantly alter the results, demonstrating that the analysis was not vulnerable to outliers. However, the model collapsed under the placebo analysis (adding a random variable to test robustness), further proving its validity.

A pair plot analysis of the methods revealed that the homogeneity of users in the AC method formed a bimodal distribution. This structure indicated that while AC produced better scores overall, the results were skewed, with users either knowingly or unknowingly ending up in feedback loops. The width between the two peaks created a misleading appearance of optimal values. NC had the highest homogeneity and coverage, but its scattered distribution suggested that outcomes were highly random and vulnerable to underlying biases in both the data and user preferences. AR, with its combination of an optimization algorithm and randomized user behavior, balanced the regressive clustering tendencies of the recommender system, resulting in intermediate values. In summary, AC has an overall negative effect, creating a feedback loop, AR achieves a middle ground, and NC maximizes diversity but lacks alignment with user preferences and real-life scenarios.

In conclusion, the influence of the algorithm shows an overall negative effect, limiting exposure to different types of films. Of

course, the limited number of users involved is one of the main limitations of this experiment. Furthermore, the employed algorithm is collaborative filtering, chosen because it is the same as the one used by MovieLens, from which we took our training data. If a content-based algorithm were employed, which suggests items similar to previously consumed ones, we might observe a higher level of homogeneity in recommendations and watch history.

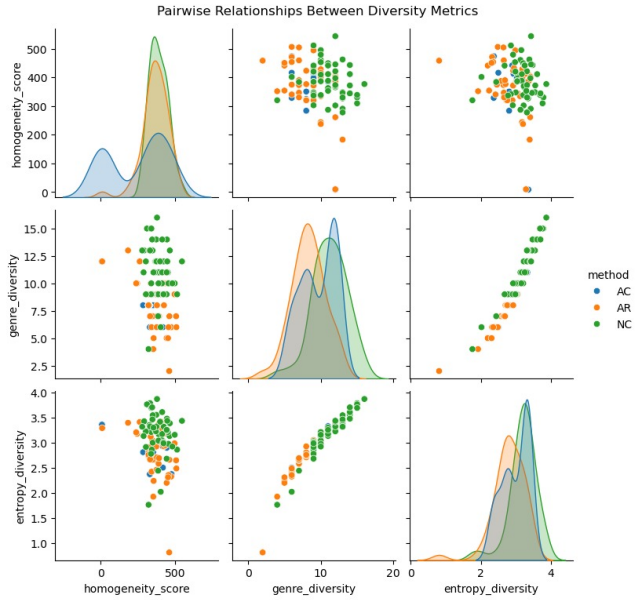

Figure 4: Pair plot analysis of methods

7 THE EFFECT OF FOLLOWING RECOMMENDATIONS

In the previous section we have analyzed a "closed scenario", in which the user was able to choose only among the recommended items. In this section, we consider a "more open scenario", where there is a set of movies and a subset of recommendations, and the user can choose either from the recommendations or directly from the movies set. This new scenario is far more realistic. For instance, Netflix homepage displays recommended movies and series, but users can also search for a specific genre or look directly for a particular title.

Our proposal is account for the events "the user chose a movies from the recommendation list" and "the user chose a movies outside recommendation list," analyzing their impact both on the homogeneity of the future recommendation set and of the watching history. This approach aims to quantify the effect of respectively user choices and algorithm on filter bubble formation.

When examining the homogeneity of the recommendation set, we cannot causally identify the role of the algorithm, as it is the sole direct cause of the recommendation set and we can't "disable it". What we can do is analyze the user's role in terms of choosing or not from the proposed items.

On the opposite, when we analyze the history of movies watched by the user, we cannot quantify the user's role but we can identify the effect of the algorithm on the homogeneity of the consume content. The idea is that if the user does not follow the recommendation list, the algorithm plays no role in determining the watched history.

7.1 Framework

We combined MovieLens data with a simulation to collect observational data about users following or ignoring recommendations. Using a simulation alone would require developing criteria to determine whether choosing among the recommendations. This is challenging because it requires a model for human like decision-making, to be trained on a dataset of user choices.

MovieLens contains a chronological sequence of ratings, along with the IDs of the user and the movie involved. With approximation, we can say this contains the watch history of each user. The idea is to use directly the MovieLens data to simulate user choices, without going through training. We split the ratings into training data (before 2018) and simulation data (after 2018). The training data are used to train a collaborative filtering algorithm with recommendation size of 15, which is used to simulate the MovieLens recommendations. Among the simulation data, we select a treatment group consisting of the 29 "new users," who have no ratings before 2018.

The simulation process involves examining the ratings in the simulation data in chronological order and updating the collaborative filtering matrix at each step. The main point is that, for each rating belonging to a user in the "treatment group," we compute its recommendation and check if it includes the rated movies. If the rated movie is included in the recommendation, we can say that the user followed the recommendation.

At the end of the simulation, for each user in the treatment group, we store the number of movies watched, the number of times they followed the recommendation, the sequence of chosen movies, and the last recommendation. This allows us to define a "recommendation-following" percentage, which is used as the treatment in our approach.

The subset of the results of the simulation is shown in Figure 5.

userId	follow	total	last recommendation	watch history
209	11	35	[Dark Knight, The (2008), Schindler's List (19...	[Matrix, The (1999), 127 Hours (2010), Watchme...
338	7	39	[Inception (2010), Matrix, The (1999), America...	[Schindler's List (1993), Fight Club (1999), B...
596	31	411	[Shawshank Redemption, The (1994), Seven (a.k....	[Pitch Black (2000), Godzilla (1998), Romy and...
89	5	518	[Up (2009), Dark Knight, The (2008), Pirates o...	[Dead Hate the Living!, The (2000), RoboGeisha...
98	5	92	[Fight Club (1999), Raiders of the Lost Ark (I...	[Star Wars: The Last Jedi (2017), Get Out (201...
362	21	109	[Dark Knight, The (2008), Saving Private Ryan ...	[Dangerous Minds (1995), Thing, The (1982), In...
491	6	64	[Harry Potter and the Prisoner of Azkaban (200...	[Ted (2012), Marley & Me (2008), Santa Clause,...
111	23	646	[Star Wars: Episode V - The Empire Strikes Bac...	[O.J.: Made in America (2016), Honest Liar, An...
248	10	51	[Lord of the Rings: The Two Towers, The (2002)...	[Pirates of the Caribbean: On Stranger Tides (...
382	19	291	[Pirates of the Caribbean: The Curse of the Bl...	[Zoolander 2 (2016), Sucker Punch (2011), Anna...

Figure 5: A screenshot from the result of the simulation

7.2 Evaluation of the User's Role

To evaluate the user's role, we examine the effect of following recommendations at least once out of five on filter bubble formation.

The first metric we used is the heterogeneity score, based on the tag genome, which produces a content diversity score ranging from 1 to 100 (Section 5.1). Since Nguyen et al.'s version [11] used

a relevance score from 1 to 5, we perform scaling and then normalization to obtain a value from 0 to 100%. From the same paper, we know that the minimum distance is 5.1 and the maximum distance is 44.24.

Setting a threshold for assessing a 'diverse' set of movies is challenging because the majority of scores fall between 44% and 47%, even though some recommendations are more similar than others. By examining the names of the films, we chose a threshold of 46% for a 'diverse' set of movies. By defining the treatment (following recommendations at 20%) and the effect (diversity over 46%), we can use the collected data to compute the average treatment effect of following the recommendation on diversity:

$$P(\text{user = following}) = P(F) = 0.276$$
$$P(\text{user = following, recommendation = diverse}) = P(F, D) = 0.103$$
$$P(\text{user = not following, recommendation = diverse}) = P(\neg F, D) = 0.172$$
$$\text{ATE} = P(D|\text{do}(F)) - P(D|\text{do}(\neg F)) = P(D|F) - P(D|\neg F) = \frac{P(D,F)}{P(F)} - \frac{P(D,\neg F)}{1-P(F)} = 0.137$$

This means that user choice plays a limited role in the diversity of the recommendations, and not in the way we might expect: choosing to follow the recommendations, the diversity of the proposed content seems to increase (rather than decrease) .

This contradicts common sense, as we often assume that recommendation algorithm reduce diversity, leading to a filter bubble. This result aligns with other non-causal research, which suggests that recommendations do not play such a negative a role as commonly believed and instead help users discover content they were previously unaware of [2, 8, 11].

The reason behind this may lie in the nature of collaborative filtering algorithms. Rather than strictly suggesting films that align closely with a user's history, these algorithms recommend films watched by other users with similar histories. For example, if a user frequently watches horror and fantasy films, a new user who primarily watches horror films will likely be recommended fantasy films. By choosing among these, the user may become more similar to a "pure fantasy" user or one who watches fantasy-adventure films, getting suggestion for these type of movies. Similarly, in the context of an e-shop, clicking on recommended items often leads users to explore new items, further enhancing content diversity.

Collaborative filtering algorithm tends to cluster users into distinct groups, each associated with specific recommendations. Rather than promoting homogeneity within recommendation sets, this may lead to a kind of global homogeneity, where users across the system become increasingly similar to one another. This phenomenon is not detectable using the tag genome metric but still constitutes a form of filter bubble. This is why we introduced our second metric, the homogeneity score, which measures the similarity to the rest of the users (Section 5.2).

After some observations, we defined a recommendation as "homogeneous" if the homogeneity score is greater than or equal to 20%, meaning that at least 3 films must be common with the top-watched movies. We can see the homogeneity score and its relationship with the following percentage in picture 6.

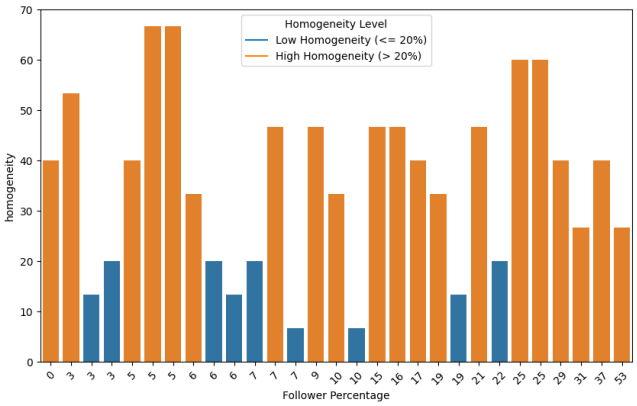

Figure 6: In orange, we can see all the users with a homogeneity level above 20%. While it is possible to find users with high homogeneity and a low following percentage, users with a higher following percentage (second half of the chart) tend to register fewer cases of low homogeneity.

Again, we can compute the ATE (Average Treatment Effect) of user choice on homogeneity:

$$P(\text{user = following, recommendation = homogeneous}) = P(F, H) = 0.276$$
$$P(\text{user = not following, recommendation = homogeneous}) = P(\neg F, H) = 0.552$$
$$\text{ATE} = P(H|\text{do}(F)) - P(H|\text{do}(\neg F)) = P(H|F) - P(H|\neg F) = \frac{P(H,F)}{P(F)} - \frac{P(H,\neg F)}{1-P(F)} = 0.238$$

This means that user choice plays a discrete role in the homogeneity of the recommendations, even though the role of the algorithm still predominant. By choosing to follow the recommendation, the users lose some of their individuality and become more similar to other users in terms of future recommendations.

7.3 Evaluation of the Algorithm's Role

Because there is no recommendation without algorithm, is hard to estimate its role looking at the recommendation list. However, we can consider users to be in a filter bubble if the sequence of movies they consume lacks diversity. This can result from the algorithm (filter bubble) or from the user's own choices (echo chamber). To distinguish between filter bubbles caused by the algorithm and user tastes, we can examine whether the user follows the recommendations.

We can evaluate the algorithm's effect by examining its impact on users who follow at least 20% of the recommendations in their watch history. Specifically, because different users have watch history of different length, we focus on the last 23 entries. This provides a way to assess the algorithm's influence on filter bubble formation. We use the tag genome metric to measure diversity, applying a 45% threshold to define a diverse sequence. Based on this, we compute the Average Treatment Effect (ATE):

$$P(\text{user = following, watch-history = diverse})$$
$$= P(F, D) = 0.138$$
$$P(\text{user = not following, watch-history = diverse})$$
$$= P(\neg F, D) = 0.103$$
$$\text{ATE} = P(D|\text{do}(F)) - P(D|\text{do}(\neg F)) = P(D|F) - P(D|\neg F)$$
$$= \frac{P(D,F)}{P(F)} - \frac{P(D,\neg F)}{1-P(F)} = 0.357$$

We observe that the algorithm has a significant impact on the diversity of the watch history. As in previous findings, the results indicate that a collaborative filtering algorithm helps increase, rather than reduce, the diversity of consumed content. Furthermore, we find that the control the algorithm has over the user's watch history is greater than the control the user has over recommendations.

Again, given the nature of the collaborative filtering algorithm, it is better to look at the homogeneity score of watch histories to gain insight into the algorithm's role in filter bubble formation.

$$P(\text{user = following, watch-history = diverse})$$
$$= P(F, H) = 0.069$$
$$P(\text{user = not following, watch-history = diverse})$$
$$= P(\neg F, H) = 0.034$$
$$\text{ATE} = P(H|\text{do}(F)) - P(H|\text{do}(\neg F)) = P(H|F) - P(H|\neg F)$$
$$= \frac{P(H,F)}{P(F)} - \frac{P(H,\neg F)}{1-P(F)} = 0.202$$

So, if followed, the algorithm has an effect on watch history homogeneity: if a user decides to follow the algorithm, he ends up with a watch history similar to that of other users.

7.4 Results

We can conclude that users play a limited but not negligible role in the diversity of future recommended contents of a collaborative filtering algorithm. By deciding to follow the recommendations, they increase the diversity of future suggestions. On the other hand, if followed, the algorithm plays a limited but not negligible role in shaping the user's watch history, helping to increase diversity. In any case, when considering a collaborative filtering algorithm, neither the user nor the algorithm is the cause of a filter bubble in terms of a loss of diversity in the recommendations.

Things change when looking at inter-user similarity. We can say that the algorithm is responsible at 80% of the increase in inter-user homogeneity of recommendation, while the user is responsible for only 20%. So, both the user and the algorithm play a role in this type of filter bubble. The influence of the algorithm is predominant, but the user's choice not to follow the recommendation can reduce the effect. The same trend is observed with the watch history, where the algorithm plays a 20% role in creating a watch history that is more homogeneous with that of other users.

We can summarize our results as follows:

- Collaborative filtering does not reduce but instead increases the diversity of watch history and recommendations.
- The algorithm is primarily responsible for "between-user filter bubble" in recommendation, and the user's choice plays only a limited role.
- The ability to choose other films makes the user primarily responsible for "following the recommendations" and falling into a "watch-history filter bubble", although the algorithm indirectly plays a role.

Finally, based on our results, we suggest the following approach to detect a filter bubble that the user doesn't want to be in: if there is a lot of homogeneity and the user chooses outside the recommendations, they are probably experiencing the annoying effects of the filter bubble. This combination of factors can be used to trigger a modification to the algorithm, such as switching to a content-based approach, or an 'injection of diversity,' which is one of the methods proposed by Jiang et al. [6] to reduce the filter bubble effect. In our context, given that collaborative filtering leads to inter-user similarity, diversity refers to movies that are less frequent among users.

8 DISCUSSION

We have provided two different causal approaches to analyze the role of users and algorithms in filter bubble formation. The first experiment shows a significant effect of the algorithm in shaping user choices, while the second experiment proves that the user has a limited but not negligible role in filter bubble formation.

We find that the filter bubble phenomenon is very complex to analyze. Indeed, the choice of threshold, the amount of data considered, the metric structure, and the amount of training can all impact the results. Furthermore, the inability to conduct a real-life simulation with real users makes it even more challenging to obtain more accurate results. The limitations of the first experiment lie in the relatively small number of agents designed to mimic human behavior. In large-scale systems, where the number of human users is often very high and interactions are more complex, these results may not fully generalize. By extending the experiment to include multiple recommender algorithms and a larger, more diverse set of users, we could achieve a more comprehensive understanding and better address the research question. For the second experiment, the main limitations are the size of the treatment group, constrained by computational limits, as well as the selected metrics and thresholds.

An interesting future direction could be to estimate the role of a content-based recommendation system in filter bubble formation. Indeed, this type of algorithm suggests items based on similarity to previously watched movies, rather than interaction with other users. Considering the results of our experiment, we hypothesize that this algorithm will have an effective role in reducing diversity, but not in increasing similarity with other users.

Other future steps may involve the development of new metrics and thresholds to better capture the filter bubble phenomenon, the use of more computational resources to increase the number of "treated users," and the application of graph acyclification techniques to handle the feedback loop in the SCM.

Finally, it would be valuable to further develop the method we sketched at the end of section 7.4 to detect the presence of annoying filter bubbles. If properly fine-tuned, this approach could enhance both user satisfaction and economic returns, while mitigating the negative societal impact of filter bubbles.

9 GITHUB

The code for the experiment is available in the following repository: https://github.com/saishridhar/CS520---Project.git

REFERENCES

[1] A.J.B. Chaney, B.M. Stewart, and B.E. Engelhardt. 2018. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proceedings of the 12th ACM Conference on Recommender Systems*. 224–232.

[2] S. Messing E. Bakshy and L. A. Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348, 6239 (2015), 1130–1132.

[3] Argueda et al. 2023. Echo Chambers, Filter Bubbles, and Polarisation: A Literature Review. *Journal Name* X, Y (2023), Z–ZZ. https://example.com

[4] C. Gao, Y. Zheng, W. Wang, F. Feng, X. He, and Y. Li. 2024. Causal inference in recommender systems: A survey and future directions. *ACM Transactions on Information Systems* 42, 4 (2024), 1–32.

[5] F Maxwell Harper and Joseph A Konstan. 2015. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)* 5, 4 (2015), 1–19.

[6] 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* (Honolulu, HI, USA) *(AIES '19)*. Association for Computing Machinery, New York, NY, USA, 383–390. https://doi.org/10.1145/3306618.3314288

[7] R. Jiang, S. Chiappa, T. Lattimore, A. György, and P. Kohli. 2019. Degenerate feedback loops in recommender systems. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 383–390.

[8] Greg Linden. 2011. Eli Pariser is Wrong. http://glinden.blogspot.com/2011/05/eli-pariser-is-wrong.html

[9] P. Liu, K. Shivaram, A. Culotta, M.A. Shapiro, and M. Bilgic. 2021. The interaction between political typology and filter bubbles in news recommendation algorithms. In *Proceedings of the Web Conference 2021*. 3791–3801.

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

[11] T.T. Nguyen, P.-M. Hui, F.M. Harper, L. Terveen, and J.A. Konstan. 2014. Exploring the filter bubble: The effect of using recommender systems on content diversity. In *Proceedings of the 23rd International Conference on World Wide Web*. 677–686.

[12] Eli Pariser. 2011. *The Filter Bubble: What the Internet Is Hiding from You.* Penguin UK.

[13] Eli Pariser. 2011. *The Filter Bubble: What the Internet Is Hiding from You.* Penguin UK. 10 pages.

[14] J. Vig, S. Sen, and J. Riedl. 2012. The tag genome: Encoding community knowledge to support novel interaction. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 2, 3 (2012), 1–44.

[15] W. Wang, F. Feng, L. Nie, and T.-S. Chua. 2022. User-controllable recommendation against filter bubbles. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1251–1261.

[16] D. Xu, C. Ruan, E. Korpeoglu, S. Kumar, and K. Achan. 2020. Adversarial counterfactual learning and evaluation for recommender system. In *Advances in Neural Information Processing Systems*, Vol. 33. 13515–13526.