Exploring the Filter Bubble: The Effect of Using Recommender Systems on Content Diversity

Tien T. Nguyen Pik-Mai Hui F. Maxwell Harper Loren Terveen Joseph A. Konstan

GroupLens Research
Computer Science and Engineering
University of Minnesota
Minneapolis, MN 55455
{tien,hui,harper,terveen,konstan}@cs.umn.edu

ABSTRACT

Eli Pariser coined the term *'filter bubble'* to describe the potential for online personalization to effectively isolate people from a diversity of viewpoints or content. Online recommender systems - built on algorithms that attempt to predict which items users will most enjoy consuming - are one family of technologies that potentially suffers from this effect. Because recommender systems have become so prevalent, it is important to investigate their impact on users in these terms. This paper examines the longitudinal impacts of a collaborative filtering-based recommender system on users. To the best of our knowledge, it is the first paper to measure the filter bubble effect in terms of *content diversity* at the *individual* level. We contribute a novel metric to measure content diversity based on information encoded in user-generated tags, and we present a new set of methods to examine the temporal effect of recommender systems on the user experience. We do find that recommender systems expose users to a slightly narrowing set of items over time. However, we also see evidence that users who actually consume the items recommended to them experience lessened narrowing effects and rate items more positively.

Categories and Subject Descriptors

H.1 [**Information Systems**]: Personalization, Recommender systems; H.1 [**Human-centered computing**]: Collaborative filtering

Keywords

filter bubble; recommender system; content diversity; user experience; tag-genome; longitudinal data

1. INTRODUCTION

In less than two decades, recommender systems have become ubiquitous on the Internet, providing users with personalized product and information offerings. They play a significant role in companies' profit margins. For example: Amazon once reported that 35% of its sales came from its recommendation systems [7]. Netflix in 2012 reported that 75% of what its users watched came from recommendations [1]. Recommender systems have greater influence on users' choices than peers and experts [14]. They lower users' decision effort, and improve users' decision quality [20].

But from the early days of recommender systems, researchers have wondered whether recommender systems might cause the *'global village'* to fracture into tribes [12], leading to *'balkanization'* [17]. Pariser [11] characterizes this worry in terms of a *'filter bubble'* – a self-reinforcing pattern of narrowing exposure that reduces user creativity, learning, and connection.

Investigating the filter bubble effect requires access to a longitudinal dataset that represents users' interaction with a recommender system and consumption of information items. We also must be able to distinguish users who act on the system's recommendations from those who do not.

In this paper, we meet these challenges by analyzing long-term users of the MovieLens recommender system. We look at whether recommendations *received* become more narrow over time, but more important we also look at whether content consumed by using recommendation systems becomes more narrow. And because the essence of the risk of filter bubbles - that is people enter them willingly because they provide appealing content - we also explore the question of whether recommenders indeed provide that positive experience – leading their users to consume content they enjoy better. We frame two specific research questions:

- **RQ1**: Do recommender systems expose users to narrower content over time?
- **RQ2**: How does the experience of users who take recommendations differ from that of users who do not regularly take recommendations?

To answer these questions, we develop two new research methods to isolate and measure the effect of accepting recommendations from recommender systems. First, we separate users into categories based on how often they actually consume recommended content. This separation lets us focus on users where a filter bubble is possible, and to compare these users against a control group who use the same system

WWW'14, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2744-2/14/04.
http://dx.doi.org/10.1145/2566486.2568012.

but do not regularly follow recommendations. Second, we introduce a method and metric for exploring changes in the diversity of consumed items over time. This method looks at the items consumed (in our case, rated) in a time window, and then uses the *tag genome* – a content coding derived from the community of users – to measure the diversity of those consumed items of a user. Hence, these analyses let us address the question of the filter bubble where it is most relevant – at the individual level.

In answering these research questions, we make several contributions. First, we introduce a novel set of methods to study the effect that recommender systems have on users. Second, we provide quantitative evidence suggesting that users who take recommendations receive a more positive experience than users who do not. Third, we find evidence that while top-recommended items become more similar, the reduction in diversity is relatively small. Finally, we find that recommendation-takers consume more *content diverse* movies than non-recommendation-takers, and that these users are actively seeking to watch more diverse movies.

2. RELATED WORK

While there is little debate about the efficacy of recommender systems in commerce, the debate about whether recommender systems are harmful to users has plenty of scholars on each side.

Pariser [11] argued that the root of human intelligence is the ability to adjust and adopt with new information, and that recommender systems trap a user into an unchanging environment. This unchanged environment, which he coined the *filter bubble*, reduces creativity and learning ability, and strengthens the belief of the user.

Tetlock [16], a political scientist, ran a study in which he asked different people with various background for opinions on political and economic issues. Surprisingly, he found that normal people gave more accurate predictions than the experts. A reason for the low prediction accuracy of the experts is that their views of the world are strengthened after years of study, leading to bias in making predictions.

Sunstein [15] went further and argued that by absorbing experiences that are personalized to them, users share fewer and fewer common experience with each other. He argued that '*Without shared experiences, a heterogeneous society will have a much more difficult time in addressing social problems. People may even find it hard to understand one another*' (p. 6, [15]).

On the other hand, Negroponte, co-founder of the MIT Media Lab, suggested that users can use recommender systems in such a way that it helps them to learn and explore new things. One such way is explored in 'The Daily US', in which users have personal intelligent agents that explore and summarize topics that are not the users' interests [9]. Negroponte called these intelligent agents *'the unequivocal future of computing'* [8].

Linden, one of the authors of Amazon's recommender system, suggested that narrowing user choices is not what personalization via recommender systems does. He argued that users can't search for items that they are not aware of, therefore, personalization increases serendipity [5]. With the idea of helping users achieve a good balance of awareness of new things, Kamba et al. [4] implemented a personalized newsagent called *Krakatoa Chronicle*. With *Krakatoa Chronicle*, users can choose how to balance between news that was personalized for them and news selected by editors as important for the whole community.

Fleder et al. [2], in their simulation study about the effect of recommender systems on sales' diversity, argued that at the user-user level, users are directed towards a common experience. This is because recommender systems cannot recommend items with less data (i.e. ratings), even if these items are favorable to the users. Therefore, recommended items can be new to an individual user, but they are overall the same (i.e. popular items). Hosanagar et al. [3], with a two-group designed study with users using iTunes as a recommender platform, also found that users tend to consume more common items. They argued that users consume the same items because recommender systems help users widen their interests, leading to higher chances of consuming same items.

Although these studies present very interesting results regarding the debate about the filter bubble, they have some limitations. In Fleder et al.'s study, although they model user purchasing behaviors and how a recommender system works, their simulation does not capture the complexity of the user behaviors and their decision making processes. Furthermore, with only two items in the simulation, their study cannot model the complexity of the eco-system of a recommender system, in which new items are added, and users' preferences drift over time. In Hosanagar et al.'s study, they build networks based on the user purchasing behaviors. They then compute the properties of these networks (e.g. median degrees & distances) as a measure if users tend to purchase the same songs.

Prior work leaves open the question of whether taking recommendations leads to narrowing of consumed content. To the best of our knowledge, our study is the first study looking at *recommended content diversity*, user behavior over the time, and the effect of taking recommendations on *consumed content diversity*.

3. DATA & METRICS

In this section we describe our datasets and discuss our methods for identifying recommendation takers and computing the content diversity of movies.

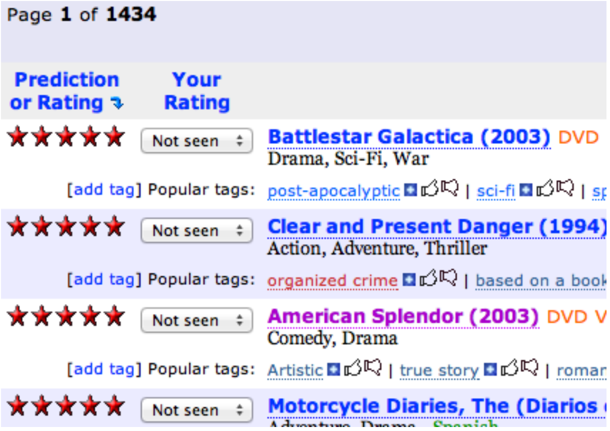
3.1 Dataset

To answer our research questions, we use data from MovieLens[1]. MovieLens is a movie recommender system that has been in continuous use since 1997. As of September 2013 , there are 217,267 unique users who have provided more than 20 million movie ratings for more than 20,000 movies. We use this data because it offers us three unique advantages: longitudinal data, a recommender system with a well-known recommender engine, and an expressive way to compute content diversity.

Longitudinal data: MovieLens provides us longitudinal data of user rating data. MovieLens logs capture timestamps and other information when users rate movies and when they view pages of recommended movies.

MovieLens provides a feature called *'Top Picks For You'* (shown in figure 1) that takes users to a page displaying movies the user has not seen, ordered from the highest predicted ratings to the lowest predicted ratings. By default, a *'Top Picks For You'* page displays 15 movies, though users

[1] http://www.movielens.org/



Page 1 of 1434

Prediction or Rating | Your Rating

★★★★★ Not seen Battlestar Galactica (2003) DVD
Drama, Sci-Fi, War
[add tag] Popular tags: post-apocalyptic | sci-fi | sp

★★★★★ Not seen Clear and Present Danger (1994)
Action, Adventure, Thriller
[add tag] Popular tags: organized crime | based on a book

★★★★★ Not seen American Splendor (2003) DVD V
Comedy, Drama
[add tag] Popular tags: Artistic | true story | romar

★★★★★ Not seen Motorcycle Diaries, The (Diarios

Figure 1: Top Picks For You

can change this default number[2]. Since May 2003, MovieLens started to log all user access to *'Top Picks For You'* pages and recommended movies with their respective positions in the recommendation lists at the time users accessing the page.

Knowing when and what movies users rated, and when and what was recommended to users helps us identify if users are taking our recommendations and how consistently they consume the recommendations.

A recommender system with a well-known recommender engine: MovieLens uses an item-item collaborative filtering (CF) algorithm [3], a well-known and broadly-used recommendation algorithm that is robust in performance and scalability with high dimension data [13]. Due to these advantages, Amazon - one of the early industrial recommender systems - used it in production [6]. We think that analyzing the longitudinal data generated from one of the well-known and broadly-used algorithms makes our case more generalizable.

An expressive way to compute content diversity: MovieLens provides tag genome data, an expressive way to characterize movie content. *'Tag-genome'* is an information space in which for any pair of a movie and a tag, a relevance score is computed to indicate how best the tag describes the movie. Since 2006, MovieLens has provided a feature that allows users to apply tags (words or short phrases) to movies. Vig et al. [19], based on this tagging feature and the tags that MovieLens users have applied, built tag genome to help users navigate and choose movies where all dimensions, but one, are the same as those of the compared movie. In section 3.3, we will describe the tag genome data in details and illustrate why we use this data to measure content diversity.

It is important to use data that is independent from the *'Top Picks For You'* computations. The tag genome data is only used to help users navigate through MovieLens' movie collections, and not be part of the *'Top Picks For You'* computations. At the time we took a snapshot of the tag genome data (April 2013), it consisted of 9,543 distinct movies described by 1,128 distinct tags (10,764,504 pairs). In our analyses, all of the movies are in this information space.

In this study, we analyze data in the period from February 2008 to August 2010 (21 months). We choose to analyze this period because of missing log data from February 2007 to December 2007 and from October 2010 to May 2012. Due to the potential missing data in January 2008 and September 2010, these periods are served as our buffer zones.

3.2 Identifying recommendation takers

To study the effect of 'taking' recommendations, we need to classify the users in our dataset into those that do 'take' recommendations and those that do not. In this section, we describe how we define these two groups.

Rating Block.

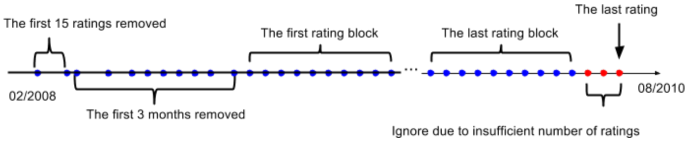

The first 15 ratings removed
The first rating block
The last rating block
The last rating
02/2008
The first 3 months removed
08/2010
Ignore due to insufficient number of ratings

Figure 2: Rating Block Ilustration

Our objective in this study is to examine the temporal effect of recommender systems on users throughout their lifecycles. To do so, we have to divide the rating history of a user into discrete intervals.

Before we define these smaller intervals, for each user we remove the first 15 ratings because these ratings are given based on the movies suggested by MovieLens in order to gain knowledge about the preference of the new user. Then, we remove all of the ratings from the first three months after the first 15 ratings. We do this for three reasons:

- some users rated an abnormally high numbers of movies in the first three months. This is potentially due to the fact that these users had watched many movies before joining MovieLens – they rated these movies to help MovieLens understand their preference better. However, in this study, we want to capture the consumed movies which were recommended;
- we want to give users sufficient time to learn how to use MovieLens;
- we want to give MovieLens enough time to understand users' preferences better, in order to improve the quality of recommendations.

After removing these initial movies, we formulate intervals for the remaining rating history of a user. There are several ways to define an interval. One is to define an interval as a login session. Another is to define an interval as a block of n consecutive months. However, both of these approaches possess some potential problems for our temporal analyses. First, the number of ratings per interval between users is different. These differences are due to the frequencies using recommender systems are different among users. Second, the numbers of ratings provided by users diminishes over the time. Hence, with an interval defined as n consecutive months (or even as a logging-in session), some intervals will have a lot of ratings, at others will not. These problems potentially make our analyses unreliable because the effect of recommender systems are different in different intervals.

[2] Our analyses suggest that only 3.2% of our users change this number.

[3] MovieLens switched to item-item CF algorithm in 2003.

To address both of the potential problems described above, we define an interval as a block consisting of 10 consecutive ratings [4]. With this definition, from now on we refer to an interval of 10 ratings as a *rating block*. With this constant number of ratings per block, we make sure that all users have the same levels of using recommender systems throughout a defined rating block. We choose 10 ratings per rating block because we want a rating block sufficiently long enough to capture the long-term effect, and because our analyses show that 10 is the median of the distribution of numbers of ratings per 3 months, a sufficiently long time interval. If there are not enough ratings to form the last rating block, we will drop these ratings because we want to make sure all rating blocks have the same number of ratings. Figure 2 summarizes our method of forming a rating block.

We only select users whose first ratings were in the *analyzed period* (i.e. in the period of February 2008 - August 2010 as defined in the section 3.1). We include only those users who have three or more ratings blocks in the analyzed period. To simplify our writing, in the remainder of this paper we refer to this selected group of experimental subjects simply as users.

Overall, we have 1,405 users in our analyses. These users made at least 3 rating blocks and at most 203 rating blocks (mean= 12, $\sigma = 15$). In our analyzed period, February 2008 to August 2010, the 1,405 users provided 173,010 ratings on 10,560 distinct movies. 100% of these movies are in tag genome database described above. They accessed their *'Top Picks For You'* 150,759 times.

Identifying consumed recommendations in a rating block.

To investigate the effect of recommender systems on users, we need to identify which movies in each rating block were explicitly recommended to the user in the interface. With these recommended movies identified, we can measure the level of recommendation intake of a user during his rating history. Furthermore, with recommended movies identified in a rating block, we can examine the user experience when taking and not taking recommendations at the same time (i.e. within a rating block). Based on individual levels of recommendation intake, we classify users into two groups - those who take recommendations and those who do not. We will discuss our classification method in more detail in the next section.

We define if a movie was recommended to user u by checking if the movie was in the *'Top Picks For You'* before. Specifically, for any user u, a movie in his i^{th} rating block is defined as *recommended to him* if and only if the movie was in *'Top Picks For You'* between 3 hours and 3 months before user u rated this movie. Figure 3 visualizes our definition.

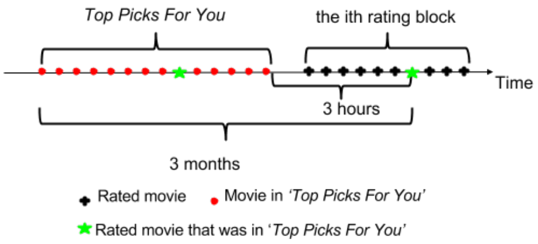

Top Picks For You
the ith rating block
Time
3 hours
3 months
Rated movie
Movie in 'Top Picks For You'
Rated movie that was in 'Top Picks For You'

Figure 3: Identifying if a rated movie was recommended before.

We require at least three hours to avoid the case where user u rated a movie upon seeing it in his *'Top Picks For You'* (an indication that the user rated it because they had seen it previously, not because they took the recommendation and watched it on the spot). We believe that three hours is sufficient time for a user to watch a movie, then rate it. We set a limit of three months to accomodate the fact that some users might need substantial time to rent and consume a movie; we capped the time limit to accomodate the reality that as time passes, the likelihood of a causal link between the recommendation and consumption diminishes.

In the next section, we discuss how we classify our users into two groups - a group that took recommendations (*Following Group*) and a group that did not (*Ignoring Group*).

Ignoring Group v.s. Following Group.

The purpose of our study is to investigate the long term effect of using recommender systems on content diversity. To this end, it is useful to draw comparisons between two groups of users - one that consumes recommendations consistently over the time, and one that does not.

Suppose that we classify user u solely based on the ratio of his rated movies that were recommended over the number of the rated movies in his rating history. Some users might always take recommendations towards the beginning of their rating histories, then do not take any recommendations towards the end. With potentially high ratios, these users could be classified as recommendation takers. However, the effects of the recommender systems on these users are only towards the beginning of their rating histories.

In order to estimate the consistent recommendation intake of a user over his rating history, we first look at whether the user took at least one recommendation in one of his rating blocks using the proposed method in the previous section. We argue that as long as within a rating block, user u took a recommendation, there was an effect of the recommendation system on that user in that rating block. We then compute the percentage of that user's rating blocks in which the user took at least one recommendation.

With these per-user percentages computed, we rank our users from the highest percentage to the lowest percentage. That said, the users who took recommendations in all of their rating blocks (i.e. percentage = 100%), are placed on top, those that did not take recommendations in any of their rating blocks (i.e. percentage = 0%) are placed bottom. Users who did not take any recommendations in any of their rating blocks are classified as non-recommender takers and placed in the *Ignoring Group*. Users who took recommendations in at least 50% of their rating blocks are classified as recommender takers and placed in the *Following Group*. Overall, the *Following Group* consists of 286 users, and the *Ignoring Group* consists of 430 users. Of these 430 users in the *Ignoring Group*, 52 never access to *'Top Picks For You'* and 378 accessed *'Top Picks For You'* but never consumed any recommendations. Figure 4 visualizes our classification method.

[4] We also analyzed with other block sizes (e.g. a block consisting of 15 (or 5) consecutive ratings), and we observed the similar results.

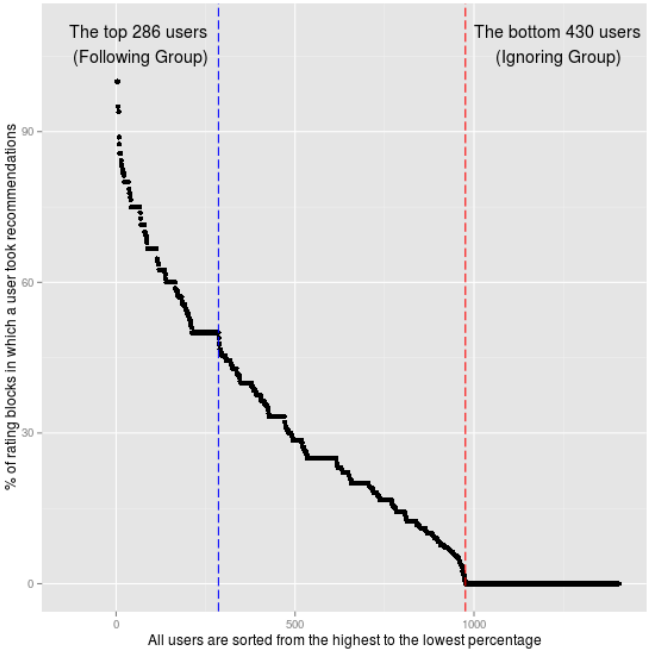


Figure 4: The visualization of our methodology to identify recommendation takers and non-recommendation takers. All users are sorted from the highest to the lowest percentages. The two cut-off points blue and red are at 50% and 0% respectively.

3.3 Measuring Content Diversity

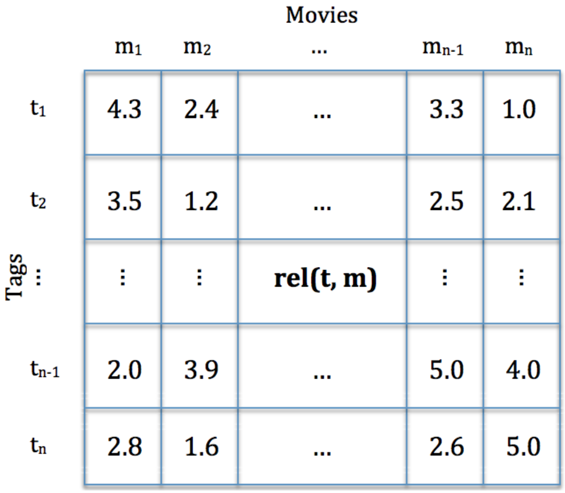
Our study examines the effect of recommender systems on the content diversity of recommended and consumed (rated) movies. In this section, we describe the tag genome data, our method to compute content diversity using tag genome and discuss why we use tag genome.

The *tag genome* [19] is an information space containing a set M of movies, and a set T of tags. The set T consists of tags that are highly descriptive about movies. How well tag t describes movie m is expressed via the relevance score $rel(t, m)$. The relevance score $rel(t, m)$ takes on values from 1 (does not describe the movie m at all) to 5 (strongly describes the movie m) [6]. Each movie m_i is represented as a vector of size $|T|$ where entry i, j is the relevance of tag j to movie i. Figure 5 visualizes the tag genome information space.

To measure the similarity of two movies, we compute the Euclidean distance between two movie vectors. That is:

$$d_{(m_i,m_j)} = \sqrt{\sum_{k=1}^{m}[rel(t_k,m_i) - rel(t_k,m_j)]^2}$$

Lower numbers indicate greater similarity. We use Euclidean distance instead of cosine distance because the movie × tag genome matrix is dense (i.e. $rel(t_k, m_i) > 0 \quad \forall i, k$). The minimum distance in MovieLens dataset is 5.1, representing the distance between two movies in the *'Halloween Series'*: *'Halloween 4: The Return of Michael Myers (1988)'*, and *'Halloween 5: The Revenge of Michael Myers (1989)'*. The maximum distance in MovieLens dataset is 44.24, representing the distance between two movies *'Paris was a woman'* and *'The Matrix'*. The average distance is 23.44 representing the distance between two movies *'Chronicle (2012)'* and *'End of Watch (2012)'*. The standard deviation of movie distances is 4.45.



Movies

Tags	m_1	m_2	...	m_{n-1}	m_n
t_1	4.3	2.4	...	3.3	1.0
t_2	3.5	1.2	...	2.5	2.1
⋮	⋮	⋮	rel(t, m)	⋮	⋮
t_{n-1}	2.0	3.9	...	5.0	4.0
t_n	2.8	1.6	...	2.6	5.0

Figure 5: The visualization of tag genome.[5]

We use tag genome because it provides an expressive way to describe the content of a movie. This expressive way is better than the traditional method of computing movie content diversity via user rating vectors. If two movies have similar user rating vectors, that means they are similarly liked, not that their content is similar. It is also better than computing movie content diversity based on meta-data such as genres, actors, or directors, etc., because two movies that share actors or directors (or even are *' comedies'*) may not actually be similar.

The strength of our tag genome-based method lies in how tag genome computes the relevances between the set of tags T and the set of movies M. These relevances are computed based on a community-supervised learning approach. In this approach, users provide the training dataset by evaluating how strongly a tag describe a movie. With the training dataset and other sources of tags such as IMDB, MovieLens predicts the relevances for other pairs based on different machine learning models. Furthermore, the relevance of any pair of tag t and movie t is constantly refined via feedback from users. Hence, these relevance scores are better at describing the content of movies than user rating vectors and properties such as genres and directors. Due to its unique advantages, researchers have shown that the *tag genome* can help users navigate through a collection of thousands of movies [18], and can assist users in remembering what movies are about [10].

To illustrate the difference of using user rating vectors and the *tag genome* for computing content diversity, we look at the following example. Based on the tag genome, the movie that is the most content similar to movie *'Halloween 4: The*

[5] This figure is reproduced based on the original figure in Vig et al.

[6] Vig et al. originally proposed the range to be from 0 to 1. However, after several revisions, as of 09/26/2013 MovieLens uses the range of 1 to 5.

Tag	*'Halloween 4 ...'*	*'Halloween 5 ...'*	*'The Front P...'*
creepy	2.551	2.328	1.087
revenge	3.185	3.920	1.311
franchise	4.992	4.993	1.162
suspense	3.864	3.890	1.261
nudity (topless)	2.712	2.848	1.071
supernatural	3.386	3.434	1.055
serial killer	4.940	4.943	1.065
splatter	2.089	3.439	1.435
teen movie	3.450	3.949	1.347

Table 1: The 9 tags describe the three movies.

Revenge of Michael Myers' is *'Halloween 5: The Revenge of Michael Myers'*. However, based on the user rating vectors of MovieLens data, the most similar movie to *'Halloween 4 ...'* is *'The Front Page'* (with the cosine similarity is 0.991). Clearly, taking the content similarity in consideration, it is obvious that the former is more accurate the later because *'Halloween 5 ...'* is the fifth movie in the 'Halloween film series' whereas *'The Front Page'* has a different story line. Table 1 shows the content of the three movies are described by 9 tags. We choose these 9 out of 1,128 tags because these tags tell what *'Halloween 4 ...'* is and is not about.

3.4 Measuring The Effect of Recommender Systems

In this study, we measure the effect of recommendation systems on content diversity as well as the user experience. In the next section, we describe the metrics to compute content diversity and user experience. Then, we discuss how we measure the effect of recommender systems.

3.4.1 The Metrics:

***Content Diversity*:** We compute the content diversity distribution of a group of users by computing the *movie distance* distribution of the group. Specifically, the content diversity of a list of recommended movies to user u is the average pair-wise distances of the movies in the list. We also do the same to compute the content diversity of consumed movies. Measuring the diversity of a list of items by averaging pairwise diversity scores was developed by Ziegler et al. [21]. To make our study more robust, we also use the maximum value of the pair-wise distances as the content diversity metric. In our results, we will report both the average as well as the maximum pair-wise distances of a list of movies.

For recommended movies, we compute the content diversity of the top 15 recommended movies per user. We choose only the top 15 because for most of our users who consulted *'Top Picks For You'*, MovieLens always captured at least the top 15 recommended movies for them. This is because 15 is the default number of recommended movies shown on the first page when a user clicked on *'Top Picks For You'*, Furthermore, only 0.05% of the MovieLens' users changed the default number to less than 15.

For the consumed movies, we measure the content diversity of rated movies. Since we divide a user history into smaller rating blocks, we compute the content diversity of all 10 rated movies in a rating block for all rating blocks.

***User Experience*:** For user experience, we measure how much users enjoy movies via their given ratings in MovieLens. Specifically, we compute per user rating average of movies in a given rating block.

3.4.2 Group Comparison

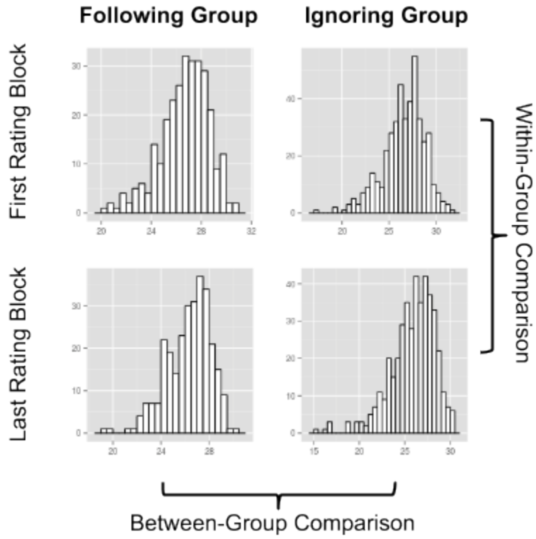


Figure 6: The visualization for our within and between group comparison for the content diversity of the consumed movies, where x is movie distance and y is a number of users. (For the recommended movies, we replace the first rating block and the last rating block as the beginning and the end of users' rating histories.)

Since all the content diversity and user experience distributions are approximately normal, we investigate the effect of recommender systems by measuring the shift in means of the two distributions of a group.

Specifically, to examine the effect of recommender systems on content diversity of a group, we measure the shift in means of the content diversity distributions at the beginning and at the end of the rating histories of all users in the group. We do the same for measuring the effect on user experience. Comparing the effect of recommender systems at the beginning and the end of a rating period is used by other researchers (for example Hosanagar et al. [3]). We call this within-group comparison.

To examine how the effect on content diversity or user experience is different between two groups, we measure the shift in means of the distributions of the two groups at the beginning rating histories of all users in the group. We do the same for the two distributions of the group at end of user rating histories. We call this between-group comparison. Figure 6 visualizes our comparison method.

Since all the content diversity (i.e. movie distance) distributions are approximately normal, we use t-tests to compare the means of the two distributions. Specifically, for the within-group comparison, we use a paired t-test. For the between-group comparison, due to the different sizes of the populations, (286 users in the *Following Group* vs. 430 users in the *Ignoring Group*, we use Welch's t-test. All the above t-tests can be performed using the R statistical package [7]. We do the same for the user experience distributions.

[7] http://www.r-project.org

4. RESULTS

We present our results as they relate to our two research questions.

RQ1: Do recommender systems expose users to narrower content over time?

To answer this question, we compare the content diversity of recommended movies at the beginning and at the end of a user's observed rating history.

Table 2 shows the content diversity of all users [8]. We observe that for all users the average pair-wise distance of the top-15 recommended movies becomes smaller over the time with a drop from 25.02 to 24.67. The p-value for the t-test is 2.43e-06, showing that the difference in the means between the two distributions are statistically significant. Therefore, although the drop in content diversity of the recommendations is small, it is statistically significant.

	At the beginning	At the end	Within-group p-value
All users	25.02	24.67	2.43e-06
Following Group	25.22	24.80	0.014
Ignoring Group	24.74	24.51	0.087
Between-group p-value	0.0037	0.0406	

Table 2: The *a*verage content diversity of the top 15 recommended movies

These drops in content diversity are also observed in *Following Group* as well as the Îgnoring Group with the within-group p-values are 0.014 and 0.08 respectively. That means for the movies recommended to the *Following Group* became more and more similar, and this trend is statistically significant at 95% confidence interval. Although the drop of the *Ignoring Group* is not statistically significant at 95% level but at 90% level, this drop carries a significant meaning. The *Ignoring Group* did not take recommendations from *'Top Picks For You'*, leading to minimal changes on in the recommendation lists. This change is due to the fact that MovieLens still learned about the preferences of these users via ratings. MovieLens then made adjustments in the recommendation lists, and recommended movies that were more similar to these users.

Interestingly, we also observe that the recommended movies to the *Following Group* seems to be more content diverse that those recommended to the *Ignoring Group* (the between-group p-value = 0.0037 at the beginning and p-value = 0.0406 at the end of user rating histories). However, the difference in the content diversity of the two groups becomes smaller over the time (0.48 at the beginning v.s. 0.29 at the end). Eventually the content diversity of the *Following Group* may become less than that of the *Ignoring Group*. However, this is an issue for future work.

Negroponte [9], Linden [5], Kamba et al. [4], and other researchers have proposed that users can use recommender systems as tools to explore new things that they are not aware of. Hence, potentially the content of consumed movies might be diverse. Thus, it is of our interest to investigate how taking recommendations affects the users' consumed content diversity and user experience. In the next section, we set out to answer our second research question:

RQ2: How does the experience of users who take recommendations differ from that of users who do not regularly take recommendations?

To answer our second research question, we set out to answer the following questions:

a) Does taking recommendations lower the consumed content diversity?

Our results, as shown in table 3, suggest that at the beginning, there is no difference in the content diversity of the consumed (rated) movies by the two groups (26.67 vs. 26.59 with p value of the t-test = 0.6162). This suggests that, after using recommender systems for the first three months [9], the effect of recommender systems on the consumed movies of both groups is not significantly different.

Rating Block	The First	The Last	Within-group p-value
All users	26.60	26.01	1.542e-12
Following Group	26.67	26.30	0.01007
Ignoring Group	26.59	25.86	8.236e-07
Between-group p-value	0.6162	0.006468	

Table 3: The average content diversity of the consumed movies of the two groups

However, our results also suggest that after using MovieLens sufficiently long enough, we can see the effect on content consumed by users. At the end of our observed periods, the content diversity of both groups is reduced. With p-values of approximately zero showing that the reductions are significant. Interestingly, we also observe that compared to *Following Group*, the *Ignoring Group* had higher drop.

We observe similar results when we define the content diversity as the maximum distance of a pair movies in the movie list (table 4). Using this metric, we find no differences between the two groups during the first three months (p-value = 0.237), and we find that users consume less diverse movies over time (p-value = 8.903e-07). Again, the following group consumed more diverse content than the ignoring group.

Rating Block	The First	The Last	Within-group p-value
All users	34.56	34.00	8.903e-07
Following Group	34.73	34.36	0.127
Ignoring Group	34.45	33.73	0.000
Between-group p-value	0.237	0.008	

Table 4: The maximum content diversity of the consumed movies of the two groups

Given the finding that the *Following Group* watched more diverse movies than the *Ignoring Group*, we ask:

b) Did the *Following Group* have better experience?

By their nature, movies recommender systems help users find movies that they may enjoy. Enjoyment is expressed via ratings: the higher the rating, the more enjoyable the movie. However, we observe that for all users (N = 1405), the rating averages at the first rating block and at the last rating block are 3.69 and 3.57 respectively, suggesting the drop of

[8] Of 1405, 4 changed the default number to less than 15; 52 users never accessed to the *'Top Picks For You'*. Thus the number of users analyzed for this analysis is 1349.

[9] We recall that the first three months of usage history are removed before forming the first rating block (see section 3.2).

	Rating Block	0.5 - 1 stars	1.5 - 2 stars	2.5 - 3 stars	3.5 - 4 stars	4.5 - 5 stars
All Users	The First	2.7%	5.3%	17.8%	46.5%	27.7%
	The Last	2.8%	6.3%	22%	46.4%	22.5%
Following Group	The First	2.2%	6.0%	17.8%	46.2%	27.8%
	The Last	1.8%	5.1%	19.0%	49.2%	24.9%
Ignoring Group	The First	2.4%	4.6%	18.0%	45.3%	29.7 %
	The Last	3.6%	6.9%	21.5%	45.1%	22.9%

Table 5: The percentage of rated movies in the respective rating ranges.

0.12. This drop surprised us since we expected that recommender systems should have helped users identify movies better suited to their tastes.

To analyze the user experience further, we look at the percentage of movies all users rated at the rating scale from 0.5 to 5 stars as shown in table 5. The percentages of watched movies that were rated higher or equal to 3.5 stars drop (from 74.2% to 68.9%), whereas for the other rating stars, the percentages increase. We observe that overall, our users watched less enjoyable movies.

Interestingly, we observe that the *Following Group* consumed more enjoyable movies. The percentage of the movies rated from 3.5 - 4 stars for this group increases from 46.2% to 49.2%, and that of the movies rated from 0.5 - 1 star decrease from 2.2% to 1.8%. Furthermore, the percentages of the movies rated from 4.5 - 5 stars of *All Users* and *Ignoring Group* receives higher drop than that of the *Following Group* (5.2%, 6.8% and 2.9% respectively).

To verify that the trend that the users in the *Following Group* watched more enjoyable movies than the users in the *Ignoring Group* is statistically significant, for each group, we compute the distributions of per user rating mean in the first rating block. We also do the same for ratings in the last rating block.

Since these distributions are normal, and have the same number of users with approximately the same variance, we perform t-test on these distributions. Like the methodology visualized in figure 6, we compare the between-group distributions, and within-group distributions. However, this time our distributions are the distributions of rating mean. Our results, as shown in table 7, suggest that in the first rating block, the users in the *Ignoring Group* had better experience than those in the *Following Group.* However, the enjoyment difference between the two groups (measured by the difference in rating mean) is not statistically significant at the 95% level of confidence interval (p-value = 0.2129). However, in the last rating block, the *Ignoring Group* watched less enjoyable movies than the *Following Group.* The enjoyment difference between the two groups is statistically significant (p-value = 0.001719). Furthermore, although the *Following Group* watched less enjoyment movies over the time, this drop is not statistically significant (p-value = 0.7). For the *Ignoring Group*, the drop in enjoyment is statistically significant (p-value = 3.128e-11).

Rating Block	The First	The last	Within-group p-value
All users	3.69	3.57	2.2e-16
Following Group	3.69	3.68	0.7
Ignoring Group	3.74	3.55	3.128e-11
Between-group p-value	0.2129	0.001719	

Table 7: Rating Mean of the two groups

We look further into the experience users receive when they consumed movies that were and were not recommended. Specifically, we look at the experience of the *Following Group* since this group consumed a significant amount of recommended movies. As shown in table 6, we observe the users in the *Following Group* consumed more enjoyable movies. The group gave at least 3.5 stars for 85% and 84.7% of consumed recommended movies in the first and the last rating block respectively. On the other hand, the group only gave 72% and 72.8% of consumed non-recommended movies at least 3.5 stars in the first and the last rating block respectively. These numbers mean that the group received worse experience when watching movies that were not recommended for them.

These results suggest that the users who followed recommendations received a better experience than those who did not follow the recommendations. However, as we mentioned above, some users in the *Following Group* did not always take recommendations. To verify whether taking recommendations indeed improves the experience of a user, we seek to answer the following question:

c) What does the change of rating average mean? To clearly understand what it means when the rating changes 0.12, we define the positive experience index as the percentile of per use rating average. That means the change in percentile of a user (or a group) indicates how the positive experience of that user (or a group) changes is comparing to the population. If the average rating of a user is at 90th percentile, that means he receives more positive experience than the rest of 90% of the population.

	Rating mean change	Percentile change
All users	-0.12	-11.97
Following Group	-0.01	-1.20
Ignoring Group	-0.19	-18.86

Table 8: The change in percentile corresponding the change in rating mean.

With this analogy, we build the percentile table based on the ratings of all of the users (N = 1405) in the analyzed period (i.e. from February 2008 to August 2010). We observe that overall at the first rating block with the average rating of 3.69, the *Following Group* is at 58.93^{th} percentile. That means the group had a better experience than more than half of the population. Whereas, the *Ignoring Group* in the first rating block with the average rating of 3.74 (at $63.63^{th}percentile$), had even better experience than the *Following Group.* However, in the last rating block, the percentile of the *Ignoring Group* drops to 44.77^{th} percentile, a 18.86 drop whereas the drop of the *Following Group* is 1.21. That implies as time went by, the *Ignoring Group*

Group	Rating Block	In Predictions	0.5 - 1 stars	1.5 - 2 stars	2.5 - 3 stars	3.5 - 4 stars	4.5 - 5 stars
Following Group	The First	Yes	1.6%	3.8%	9.6%	50.2%	34.8%
		No	2.5%	6.1%	19.4%	45.4%	26.6%
	The Last	Yes	1.6%	2.9%	10.8%	52.0%	32.7%
		No	2.2%	5.2%	19.8%	48.8%	24.0%

Table 6: The percentage of rated movies in the respective rating ranges.

received worst experience, or watched significantly less enjoyable movies, than the *Following Group* did. Table 8 summarizes our results.

5. DISCUSSION

We set out to better understand the broadening or narrowing influence of an online recommender system on its users: did it tend toward a filter bubble? We found evidence for two forms of narrowing when analyzing all users - the items recommended by the system and the items rated by users both became slightly narrower (less diverse) over time. However, the results for *all* users obscure the most interesting part of the story. The narrowing effect actually was mitigated for users who appeared to "follow" the recommender (operationalized as having rated movies that appear in their top-n recommendation lists); in other words, *taking recommendations lessened the risk of a filter bubble*.

First, recommendation-following users received more diverse top-n recommendation lists than non-following users. Because recommenders are personalized, user actions affect their output. In the case of the relatively standard item-item algorithm evaluated in this research, rating recommended movies (rather than movies chosen via other means) appears to encourage the algorithm to broaden its future recommendations. Second, recommendation-following users narrowed the content diversity of their rated movies more slowly – these users were still narrowing (significantly, but slightly), but the effect was smaller than for those users who never rated the movies the recommender showed to them.

This begs the question - is there a "natural" narrowing effect over time, at least in the domain of movies? After all, we form habits based on what we've watched recently, and as we watch more, we solidify our preferences. In the movie domain, we face the additional possibility that the best movies are relatively diverse in content, but limited in number; once we get through those, we turn to newer movies closer to our comfort zone. If this is true – if there is a natural tendency to narrow our consumption of movies (or other media) over time – then collaborative filtering-based recommenders appear to help mitigate the tendency, and thus may play a broadening role.

What can the designers of recommender systems do to discourage the narrowing tendency? First, they can use collaborative filtering algorithms like those in MovieLens, which slows the narrowing effect over time. It is an open question if content-based recommenders have the same effect as collaborative recommenders; we suspect the content-based alorithms will more strongly push users towards narrow consumption. Second, recommender systems can inform users about the diversity of their consumption. Be it movies or news, a site can display diversity metrics or summary statistics that help users better understand if they have in fact gone too far into a particular interest of theirs. Finally, if recommenders aren't enough to reduce the narrowing effect, we should explore further steps to intentionally increase diversification of recommendation lists. This is consistent with Ziegler's finding [21] that diversification can improve user satisfaction.

Our work has several limitations. We cannot be sure if people are really following MovieLens recommendations, since we are using log data analysis methods. Additionally, users may be influenced by recommendations from other information sources or from their friends. To verify recommendation-following behavior would require contacting users to develop baseline measures for recommendation awareness. Furthermore, due to the design of the MovieLens logging infrastructure, we are restricted to analyzing the "top picks for you" interface. A superior set of log data would facilitate analysis across all recommendation interfaces in the system. Perhaps most importantly, we are attempting to find generalizable learning from a particular system (MovieLens + item-item CF) with a particular kind of item (movies). There is plenty of room for studying the differential rates of narrowing (or broadening) across media, and across algorithms. We hope our methods and results can be applied to inform the study of those domains.

6. ACKNOWLEDGEMENT

This work is supported by the NSF grants IIS IIS 08-08692 and IIS 10-17697. We also thank the late Professor John T. Riedl for his advice and support in the first half of this project. We thank Shou Chang, Daniel Kluver, and Aaron Halfaker for feedback.

7. REFERENCES

[1] X. Amatriain and J. Basilico. The netflix tech blog: Netflix recommendations: Beyond the 5 stars (part 1). http://techblog.netflix.com/2012/04/netflix-recommendations-beyond-5-stars.html, visited on 2013-09-06.

[2] D. Fleder and K. Hosanagar. Blockbuster culture's next rise or fall: The impact of recommender systems on sales diversity. *Management Science*, 55(5):697–712, May 2009.

[3] K. Hosanagar, D. M. Fleder, D. Lee, and A. Buja. Will the global village fracture into tribes: Recommender systems and their effects on consumers. SSRN Scholarly Paper ID 1321962, Social Science Research Network, Rochester, NY, Oct. 2012.

[4] T. Kamba, K. A. Bharat, and M. C. Albers. The krakatoa chronicle-an interactive, personalized newspaper on the web. 1995.

[5] G. Linden. Eli pariser is wrong. http://glinden.blogspot.com/2011/05/eli-pariser-is-wrong.html, visited on 2013-09-13.

[6] G. Linden, B. Smith, and J. York. Amazon.com recommendations: item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76–80, 2003.

[7] M. Marshall. Aggregate knowledge raises $5m from kleiner, on a roll | VentureBeat. http://venturebeat.com/2006/12/10/aggregate-knowledge-raises-5m-from-kleiner-on-a-roll/, visited on 2013-09-06.

[8] N. Negroponte. 000 000 111 - double agents. http://www.wired.com/wired/archive/3.03/negroponte_pr.html, visited on 2013-09-13.

[9] N. Negroponte. *Being Digital*. Random House LLC, Jan. 1996.

[10] T. T. Nguyen, D. Kluver, T.-Y. Wang, P.-M. Hui, M. D. Ekstrand, M. C. Willemsen, and J. Riedl. Rating support interfaces to improve user experience and recommender accuracy. *To appear in the seventh ACM Recommender System Conference, RecSys 2013*, Oct. 2013.

[11] E. Pariser. *The Filter Bubble: What the Internet is Hiding from You*. Penguin, Mar. 2012.

[12] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. GroupLens: an open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work*, CSCW '94, pages 175–186, New York, NY, USA, 1994. ACM.

[13] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, WWW '01, pages 285–295, New York, NY, USA, 2001. ACM.

[14] S. Senecal and J. Nantel. The influence of online product recommendations on consumers" online choices. *Journal of Retailing*, 80(2):159–169, 2004.

[15] C. R. Sunstein. *Republic.com: XA-GB. ...* Princeton University Press, 2002.

[16] P. E. Tetlock. *Expert political judgment: How good is it? How can we know?* Princeton University Press, 2005.

[17] M. Van Alstyne and E. Brynjolfsson. Global village or cyber-balkans? modeling and measuring the integration of electronic communities. *Management Science*, 51(6):851–868, 2005.

[18] J. Vig, S. Sen, and J. Riedl. Navigating the tag genome. In *Proceedings of the 16th international conference on Intelligent user interfaces*, pages 93–102. ACM, 2011.

[19] J. Vig, S. Sen, and J. Riedl. The tag genome: Encoding community knowledge to support novel interaction. *ACM Trans. Interact. Intell. Syst.*, 2(3):13:1–13:44, Sept. 2012.

[20] B. Xiao and I. Benbasat. E-commerce product recommendation agents: use, characteristics, and impact. *MIS Q.*, 31(1):137?209, Mar. 2007.

[21] C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web*, pages 22–32. ACM, 2005.