Hypothetical Recommendation

James Schaffer, Tobias Höllerer, John O'Donovan
Department of Computer Science
University of California, Santa Barbara
{james_schaffer, holl, jod}@cs.ucsb.edu

ABSTRACT

Explanation and feedback to a user during the recommendation process can influence the user experience. In many real-world recommender systems, however, creating a profile modification based on this feedback involves buying a product, listening to a piece of music or some other costly action. This paper studies the effects of what we call hypothetical recommendations: low-cost, exploratory profile manipulations ("what-if" scenarios) based on informative dynamic feedback from the recommender system. An experiment (N=263) designed to track a user's profile manipulation behavior in the presence and absence of recommender feedback is described. Results from this study suggest that (i) feedback during profile manipulations has a biasing effect on user opinion, (ii) iterated profile manipulations after receiving even very few recommendations tends to increase centrality and reduce diversity, (iii) users are more likely to perform profile manipulations while using a visual recommender, and (iv) recommendation accuracy improves linearly with profile manipulations, however, each individual manipulation has more effect on the recommendation accuracy when feedback is not present.

Categories and Subject Descriptors

Information Storage and Retrieval [**H.3.3**]: Information Search and Retrieval - Relevance Feedback

Keywords

User Interfaces, Visual Knowledge Representation

1. INTRODUCTION

Recommender systems have evolved to help users get to the right information at the right time [18, 19]. In recent years, a number of researchers and practitioners have argued that the user experience with recommendation systems is equally, if not more, important than accuracy of predictions made by the system [7]. Research has shown that providing feedback to users during the recommendation process can have a positive impact on the overall user experience, in terms of user satisfaction and trust in the recommendations, in addition to accuracy of predictions [1, 14, 9, 8]. In many real-world recommender systems, however, providing an update to a profile based on dynamic feedback can be costly. For example, purchasing a product on Amazon or listening to a tune on Pandora or Spotify have financial and time costs respectively. While researchers have explored the benefits, limitations and effects of conversational recommender systems [11], these studies focus on granular refining of requirement specifications for individual product search. In this paper, we focus on evaluating how the experience of the recommendation consumer is affected by low-cost, exploratory profile manipulations, which we call *hypothetical* recommendations. These are scenarios that allow a user to ask questions of the form "what if I purchased product x?", "what if I listen to these 10 songs?". Specifically, this paper describes a study involving 263 participants, designed to answer the following three research questions:

1. When a user is provided with an enabling framework to make low-cost manipulations to their preference profile, what do they tend to do?
2. After producing a set of "low-cost" profile adjustments (additions, deletions, re-rates), what is the effect of the new profile on recommendation accuracy, diversity, user satisfaction and trust in the system?
3. What is the impact of dynamic feedback (through a visualization) on both of the above

Previous work on profile elicitation for collaborative filtering systems has focused on passive [17] and active [2] approaches. The experiments discussed in this paper can also be classed as a form of active profiling, since we encourage the user to create "hypothetical" additions, deletions and re-rates from a previously established profile. This research also considers the role and impact of an interactive user interface for eliciting and encouraging profile manipulations from the user, both in the presence of interactive feedback and without it. Before we proceed with our discussion of the experiment itself, the following sections frame the experiment in the context of previous research on explanation and interaction aspects of recommender systems.

2. ENGAGING USERS

The majority of research in recommender systems is focused on improving recommendation algorithms (e.g. [10, 18, 13]), without specific focus on user experience. This research builds on a number of related research efforts that deal with

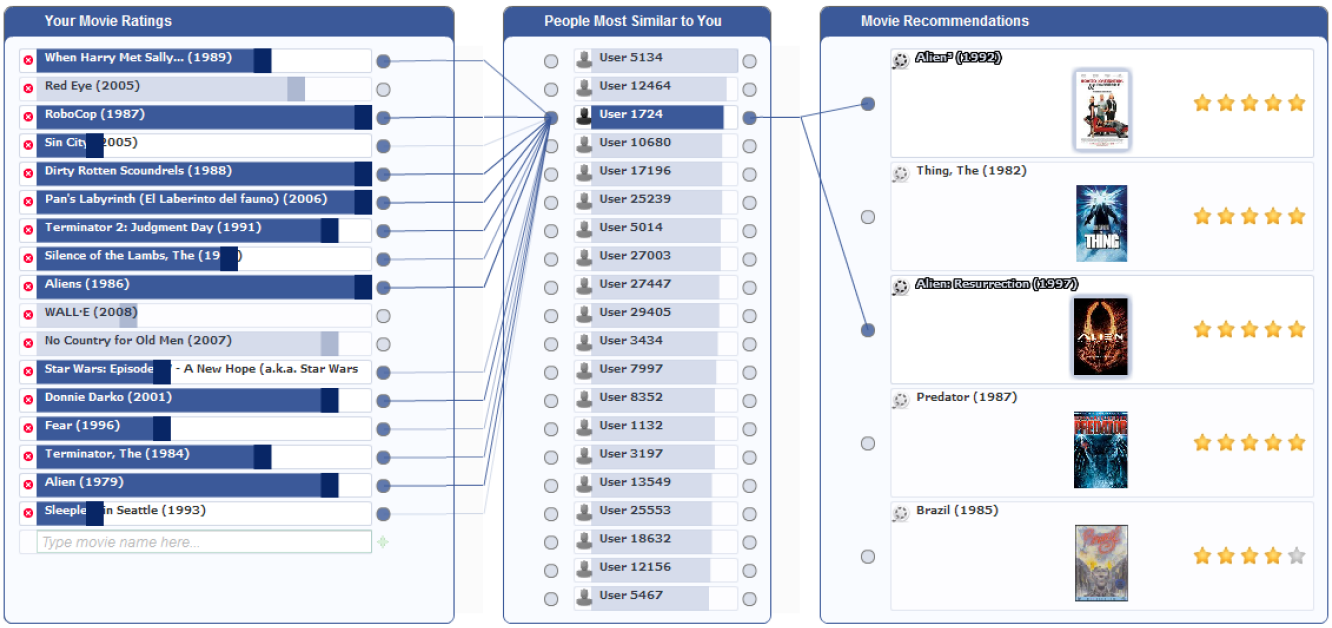


Figure 1: The Fluo user interface, configured for a visualization of collaborative filtering. From left to right- the columns display: a user's profile items; most similar users from the MovieLens dataset; current top movie recommendations. User similarities and movie recommendations are recalculated and re-ranked as a user adds, deletes, or re-rates items. The dark blue lines appear when a user clicks a node and show provenance among movies and users. Clicking a movie recommendation on the right side of the page opens the movie information page on Rottentomatoes.com.

visualization, interaction and control of recommender systems. Earlier work by Herlocker [6] demonstrated that explanation interfaces for recommender systems can improve the user experience, increasing the trust that users place in the system and its predictions. Cosley et al. [4] build on the explanation study to explore how explanations can change the opinions of a recommendation consumer, particularly in terms of rating behavior. They focus on consistency of re-rating behavior, impact of the rating scale and of dynamic feedback. Our experiment differs from Cosley's study [4] in that feedback is not explicitly controlled to be high or low quality, placing the focus on the true impact of hypothetical profile manipulations on the overall user experience. Work by Swearingen and Singha [21] finds that users tend to have higher trust in recommender systems that predict items that they already know and like. They posit two important considerations for interaction design: what user needs are satisfied by interacting and what specific features of the system lead to satisfaction of those needs? In the context of our experiment, we believe that the user "need" is a desire to explore and probe the information space, and that a low-cost "hypothetical recommendation" feature provided by an interactive visualization tool can fulfil this user requirement. Other researchers have focused on iterative profile refinement in so-called conversational recommender systems [11], but at the level of specifing thresholds for various item-specific metadata, for example, a minimum requirement on the battery life for a digital camera. This approach has proven to produce more accurate predictions and increased satisfaction over time, but is a task-specific interaction, targeted at finding one particular item. Our experiment differs in that the post-manipulation profile is used to generate many recommendations covering multiple items. Work by Tintarev and Masthoff in [23] provides guidelines for designing user-centered recommendations, focusing on the explanation process, and a survey and discussion of the advantages and pitfalls in explaining recommendations is provided in [22], highlighting the importance of transparency, scrutability, persuasiveness and trust.

2.1 Interactive Recommendation Systems

Recent work in this area focuses on visual interactive explanation and control mechanisms for recommender system algorithms. O'Donovan et al. [14] describe an interactive visualization tool that supports genre-based manipulations of the k-nearest neighbors used in a collaborative filtering algorithm. They argue that "over-tweaking" can reduce the quality of recommendations if the interactive manipulations are not well balanced with the pre-existing user profile information. Gretarsson et al. [5] describe an interactive visualization of a social recommender using the Facebook API. They discover that visual representations are especially important in social recommendations because identification of a known neighbor can trigger an infusion of pre-existing knowledge from the participant. Bostandjiev et al. [1] describe a visual interface to a hybrid recommender system that supports user guided transitions between social and semantic recommendation sources, and this system is leveraged by [9] in an experiment to study the effect of inspectability and control in social recommender systems. In particular, Knijnenburg et. al finds that both inspectability and control have a positive impact on user satisfaction and trust in the recommender system, but caution that many users may not want to see low level details of the recommendation process. Verbert et al. [24] further analyze the impact of information visualization techniques on user involvement in the recommendation process. Their evaluation of the Conference Navigator system [24] shows that the effectiveness of recommendations and the probability of item selection increases when users are able to explore and interrelate entities.

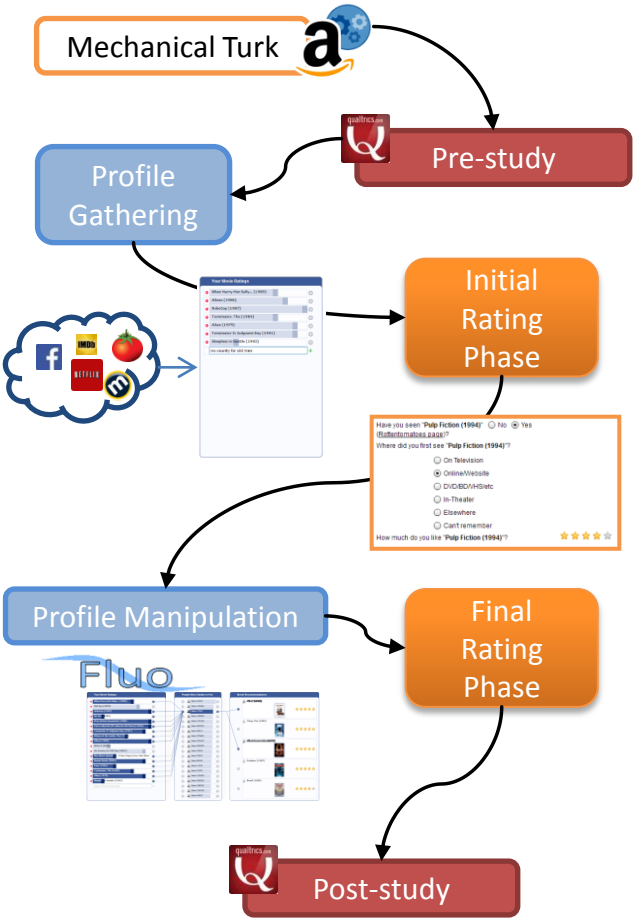


Figure 2: Overview of the crowdsourced experiment to evaluate hypothetical recommendation based on interactive modifications to pre-existing profile data.

3. EXPERIMENT OVERVIEW

The goal of our experiment is to a) evaluate how low-cost profile manipulations to aging profile data influences both the user profile, and the resulting "hypothetical" recommendations, and b) evaluate how dynamic feedback from the recommender affects profile manipulation and the resulting predictions. Figure 2 provides a high-level overview of the experiment. Participants in the experiment interacted with a collaborative-filtering recommender system operating on the MovieLens 10M dataset through two different configurations of the interactive user interface, shown in Figure **??**, one with dynamic feedback and one without. Pre-existing profile information was retrieved by participants through a web service of their choice (Netflix, IMDb, etc.) and we asked users to rate recommendations from the system based on this initial profile as a benchmark. Next, users updated their profiles using the interactive recommender and received iterative feedback from the recommender based on a treatment (feedback or no feedback), and were then asked to rate a second set of movies.

4. SETUP

The participant task consisted of two phases: a profile gathering phase and a profile manipulation phase (Figure 2). Participants were asked to rate a list of recommendations at the end of each phase. Additionally, in the second phase, participants were also put into one of two between-subjects treatments: feedback or no-feedback. Participants in the feedback condition received recommendations generated by the system on the fly as they manipulated their profile. By comparing ratings from the first phase against the second phase, and between treatments, we were able to examine how manipulation of profiles affected satisfaction, trust, and recommendation quality in the presence and absence of interactive feedback from a recommender system. To use our earlier analogy, profile manipulations can be used to establish "what-if" scenarios at low cost to the user. Our aim is to assess how users go about this, and what the resulting outcome is for their final recommendations, and subsequent user opinions.

The recommender system was deployed on Amazon Mechanical Turk (AMT) and data was collected from AMT workers. The AMT web service is attractive for researchers who require large participant pools and low cost overhead for their experiments. However, there is valid concern that data collected online may be of low quality and require robust methods of validation. Numerous experiments have been conducted, notably [3] and [16], that have attempted to show the validity of using the service for the collection of data intended for academic and applied research. These studies have generally found that the quality of data collected from AMT is comparable to what would be collected from supervised laboratory experiments, if studies are carefully set up, explained, and controlled. Previous studies of recommender systems have also sucessfully leveraged AMT as a subject pool [1, 9]. We carefully follow recommended best practices in our AMT experimental design and procedures.

4.1 Recommendation Strategy

Since the focus of this paper is on examining the profile manipulation behavior of users, and not specifically on the underlying recommendation algorithm, we chose a popular collaborative filtering recommendation algorithm and dataset that was easy to implement, robust, and comparable to other work in the field [12].

4.2 Dataset

The Movielens 10M dataset[1] was used in the experiment. The dataset spans over 70000 users and just over 10000 movies, with a fairly complete library from before 2008. To speed up interaction in our web-based visual recommender, the dataset was reduced from to 30000 users. This facilitated on-the-fly recalculations of similarity scores with every interaction a user made in the system. The number of movies was not decreased. With this reduction, updates to recommendations calculated by our Java-based server backend ranged from 1-3 seconds on average in our testing, which is noticeable to users but, with appropriate progress feedback, reported as "tolerable".

4.3 Algorithm

A collaborative filtering algorithm was chosen for this experiment because of its popularity, and since it lends itself well to visualization and users have reported being able to easily understand visual representations of the algorithm [6, 1].

[1]http://grouplens.org/datasets/movielens/

Specifically, a predicted rating $\hat{r}_{u,m}$ for some movie $m \in M$ was calculated by the simple aggregation function shown in Equation 1. u is the taget user receiving a recommendation and $u' \in U$ is the set of all users in the Movielens dataset. $r_{x,m}$ is the rating user x gives to movie m. $P(x, y)$ is the similarity function used and k is a normalizing constant (Equation 4).

$$\hat{r}_{u,m} = k \sum_{u' \in U} P(u, u') r_{u',m} \tag{1}$$

Equation 2 describes our formulation of Pearson correlation between users, M_x represents all movies that user x has rated, and $\overline{r_x}$ is the mean rating for user x.

$$P(u, u') = \frac{hdamp(u, u') \sum_{m \in M_u \bigcap M_{u'}} (r_{u,m} - \overline{r_u})(r_{u',m} - \overline{r_{u'}})}{\sqrt{\sum_{m \in M_u \bigcap M_{u'}} (r_{u,m} - \overline{r_u})^2 \sum_{m \in M_u \bigcap M_{u'}} (r_{u',m} - \overline{r_{u'}})^2}} \tag{2}$$

Pearson correlation is calculated on the intersection to improve performance, so we apply Herlocker damping (Equation 3) as in [20] to increase recommendation quality.

$$hdamp(u, u') = \frac{\min(|M_u \bigcap M_{u'}|, 50)}{50} \tag{3}$$

$$k = \arg\max_{m \in M} r_{u,m} \tag{4}$$

4.3.1 User Interface

Figure 1 is a visualization of a collaborative filtering algorithm in a novel system called Fluo. Built on research from [1], Fluo shows a column-based visualization that is suitable for visualizing the top-N results from ranking algorithms. Fluo can ingest any data that can be modeled as a node-edge graph and specializes in visualizing similarity scores between semantic groups of nodes (e.g. the effect of an item's rating on the correlation of a user in a recommendation database). Nodes are organized into columns (referred to as tubes), which can be placed serially (visually creating an upstream/downstream relationship) or in parallel (when multiple sources are weighted together). Each node in the visualization has a corresponding "score" which is shown as a light blue slider and can be mapped to any value corresponding to the underlying algorithm. Users can provide feedback (or change the behavior of) the underlying algorithms by specifying a score using the horizontal slider (if available) on each node. User interaction with nodes in left-hand columns prompts a re-computation of scores and re-sorting of downstream items. In this way, an implicit relationship between categories and individual nodes is communicated to the user. Explicit relationships in the form of edges are only drawn between columns and only on-demand, eliminating the visual clutter associated with many other node-edge visualizations, e.g., as discussed in [5].

Our experiment uses a three column representation of collaborative filtering, similar to [1]. From left to right, a user sees his or her movie profile, then similar users in the collaborative filtering database, and finally a list of top movie recommendations. The underlying algorithm represents results from collaborative filtering as a directed graph, connecting the user's profile items to database users with at least one overlapping item and specifying edge strength as a similarity score. This score is shown as a light-blue gauge on the node for simplicity. Thus, if a user clicks on a movie he has rated, they can see which other similar users have rated it, and which recommendations are a result of those ratings. The recommendation column uses a star notation rather than a bar, provides visuals from the movie in the form of a teaser poster, and, when clicked, takes the user to RottenTomatoes.com to get more information about the movie.

Treatment	*First Phase*	*Second Phase*
1	Gathering	Manipulation (no feedback)
2	Gathering	Manipulation (w/ feedback)

Table 1: Breakdown of Participant Task and Treatments

4.4 Procedure

Figure 2 shown earlier gives a high level overview of the experimental procedure. The following sections details each individual phase.

4.4.1 Study Access and Pre-study

After accessing the experimental system through AMT, participants were presented with a pre-study questionnaire using the Qualtrics survey tool[2], collecting basic demographic and movie expertise information. Next, they were directed a to page that asked them about their online movie profiles. Quick links were provided to a collection of the most popular online or social web services for movie ratings. Participants were asked to select the service they used most frequently or to provide the name of a service not in the displayed list. Participants that were not active users of online webservice were not allowed to do the main experiment. The resulting distribution was: Netflix 163; IMDB 55; Facebook/Other 33; RottenTomatoes 10; Metacritic 2.

4.4.2 Gathering Phase

Participants were presented with a single-column view (profile only) of Fluo and tasked with copying their old profile items into the user interface. They were instructed to log into the respective webservice, navigate to their saved movie ratings, and use the interface to copy the ratings exactly as they appear. For the (n=230) participants that used Netflix, RottenTomatoes, IMDb, or Metacritic, direct links were provided that eased the process.

Once the user's profile data was collected, recommendations were computed the top 5 were presented to users before further profile manipulation occurred. These initial recommendations provided a baseline for both treatments and allowed for a before-and-after analysis of recommendation satisfaction, accuracy, and trust. For each movie recommended, we asked the participant if they had seen the movie and via what medium, then asked for a rating on a (1-5) Likert scale. If the participant had not seen the movie, we provided a direct link to the title's RottenTomatoes page where they could view a trailer, plot synopsis, user reviews, and critic reviews. In these cases, the participant was also asked how much they thought they would like the movie.

[2]www.qualtrics.com

Metric Name	*Explanation*
Recommendation Accuracy	Mean of ratings collected from participants for movie recommendations given in the initial and final rating phases, using a 1-5 Likert scale.
Subjective Satisfaction	The participant's reported satisfaction with the recommendations, given on a scale of 1-100.
Subjective Trust	The participant's reported trust in the recommender, giveon on a scale of 1-100.
Subjective Accuracy	The participant's qualitative assessment of the accuracy of the recommender, given on a scale of 1-100.
Interaction	The total number of manipulations performed by the user during the profile manipulation phase. Does not include manipulations required to gather the profile. Each manipulation corresponds to an add, delete, or re-rate.
Centrality	A participant's degree in the recommendation neighborhood; a two users are connected if they have at least one item in common.

Table 2: Key dependent variables in the study.

4.4.3 Manipulation Phase

After the first set of recommended movies were rated by participants, they entered the manipulation phase of the experiment. Participants were instructed to update the profile entered earlier any way they saw fit, either from memory or by looking at a web resource of their choice. They were instructed to take their time and try to create a profile that they thought would generate the best recommendations for them. Participants were not forced to make manipulations, since some may have been perfectly satisfied with their original profile (for instance, because they meticulously kept it up to date), so participants could advance the study at any time without performing further manipulations. After advancing, they were required to rate a second wave of recommendations, which included the top five recommendations generated by their final profile, but also the top recommendation from each manipulation iteration. On average each user rated 6 items, but some rated as many as 15-20.

4.4.4 Post-Study

After completing the main study, participants were redirected back to Qualtrics to complete a short post-study. Upon completing this, they were provided with a unique completion code to enter into AMT to receive payment. A summary of selected post-study questions and average responses are shown in Figure 5 and discussed in the following section.

5. EXPERIMENTAL RESULTS

More than 300 users started the study, but a small percentage did not finish. We assume that these trials were abandoned for other arbitrary reasons, such as external distractions. To check the integrity of data used in our analysis, manipulation records were run back through the Java implementation of our collaborative filtering recommender and compared with the recommendation records. Though few participants had erroneous logs with errors that could not be determined (and others dropped out of the study partway through), quality data was collected from 263 users. The average rating over all recommendations in the study was 3.741, and 73 % of recommendations were reported as known by the user.

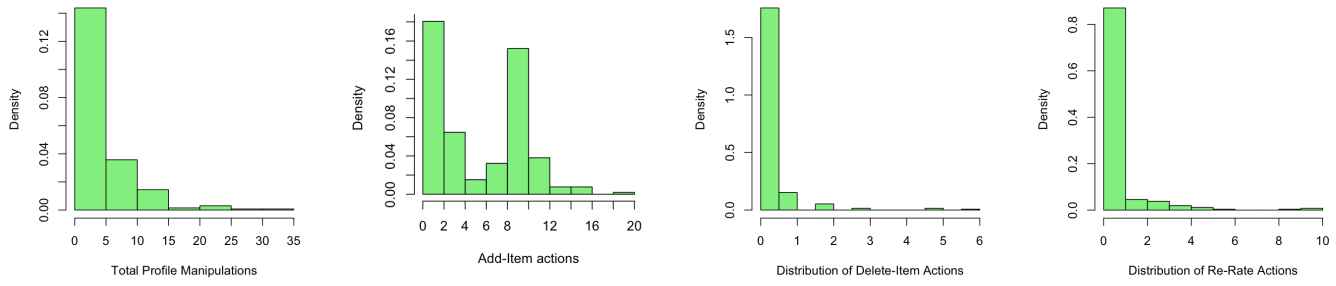
5.1 Demographics

Participant age ranged from 18 to 65, with an average of 31 and a median of 29. 53% of participants were male while 47% were female. Participants were asked questions on a five-point Likert Scale to gauge their familiarity with related technology such as social networks and recommender systems applications. On average, participants had 293 Facebook connections. When asked how often they update their profiles on social media sites, they reported an average of 2.77 with a variance of 1.37. Most users reported that they were familiar with recommender systems (115/283). The mean reported value frequency of movie watching was 3.56/5; self-reported movie expertise was 3.32; self-reported score (1:completely disagree, 5:completely agree) for "rarely watch movies" was 2.1. When asked if they usually pay close attention to detail, 57% of users strongly agreed. When asked if they frequently changed their mind (compared to their peers), most users reported that they were average on the scale. 24% of users reported that they updated their profiles on Facebook very frequently; 43% had intermittent updates and 26% updated rarely.

5.2 Profile Manipulation Behavior

Analysis of manipulations was done after the profile gathering step. The manipulations of participants as shown in Figure 3 occur strictly after the gathering phase. That is, the initial additions while transferring items from their pre-existing profiles were not counted in the analysis. Most participants re-rated and deleted just a few items (usually <5). More profile additions were created. The peak in Figure 3(b) around 10 manipulations occurrs becuase that was the default number of slots shown on the user interface. Participants were free to add more by clicking on a button, but only a relatively small number added more than 10 items at this step.

In the experiment participants were simply instructed to update their profile until they were satisfied it would produce the best recommendations possible for them. Unfortunately, many participants immediately reported satisfaction with their profile and many might have originally copied an up-to-date profile or were simply satisficing [15] to claim the study reward. Participants were split into bins based on their interaction behavior to better summarize the results. As many bins as possible were created to create a good picture of the data without sacrificing balancing, and the resulting breakdown is shown in Table 3. For participants with a non-zero interaction level, average interaction was 6.14 manipulations, with a median of 5 interactions (both of which fall into the medium interaction bin). Note that the group of participants in the "No Interaction" bin is quite large, as a large number of participants reported they were satisfied with their profile after the initial step and did not interact further. Additionally, note that the feedback group $\mu = 4.66$



(a) (b) (c) (d)

Figure 3: Distribution analysis of profile manipulations.

Interaction Groups				
Group	*Manipulations*	*NF*	*F*	*Total*
No Interaction	0	76	64	140
Low Interaction	1-3	16	17	33
Medium Interaction	4-7	38	15	53
High Interaction	8+	14	24	48

Table 3: Binned participant behaviors "Group": the group used for analysis, "M": number of total manipulations after first rating phase, "NF": number of participants in this bin for the no-feedback treatment, "F": number of participants in this bin for the feedback treatment, "Total": total number of participants in this bin regardless of treatment.

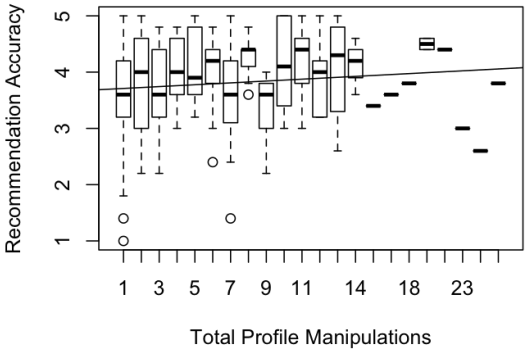


Figure 4: Combined Effect of Profile Manipulations on Recommendation Accuracy. Accuracy was recorded as as live user ratings of predictions on a 5-point Likert scale.

95% CI [3.57, 5.74] interacted, on average, 25% more than the no feedback group $\mu = 3.73$ 95% CI [3.72,3.74].

5.3 Perceived Satisfaction, Trust and Accuracy

Since user experience is a critical aspect of recommender system evaluation, the post-test study analyzed perceptions of trust, satisfaction with, and accuracy of the provided recommendations. The three groups in Figure 5 (caveat: the nonzero start on the y-axis) highlight the changes in these perceptions before and after the profile manipulation phase. These results are shown by the blue plots on the left of each group. There is a relative improvement in perceived trust of 11.3% for users in the dynamic feedback condition (i.e: those who saw the full graph in Figure 1 as they performed profile

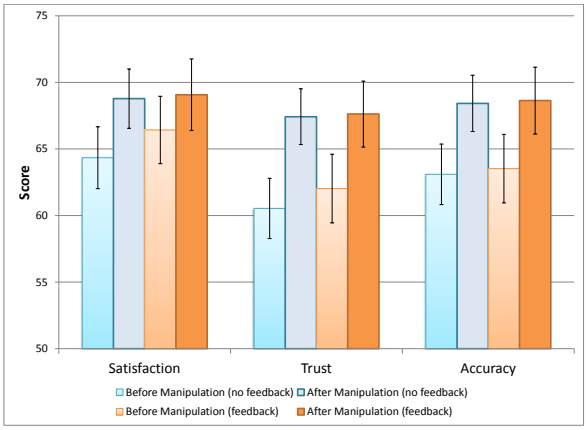


Figure 5: Result of post-study analysis showing perceived satisfaction, perceived trust in recommendations and perceived recommendation accuracy before and after profile manipulations were performed in both conditions. All metrics show increaing trends after profile manipulation has occurred.

manipulations. Though no statistical significance was found, perceived accuracy showed a relative increase of 8% for the feedback condition, and perceived satisfaction showed a 7% increase. In the condition with no feedback, increases in all three metrics were also present, but they were less pronounced, as shown in Figure 5.

5.4 Empirical Accuracy

This section discusses how different levels of manipulation affected the accuracy of recommendations. Figure 4 shows a boxplot of total manipulations against accuracy, and an increasing trend is visible. However, Figure 6 compares the mean average rating of recommended movies between the first phase (completed by all participants) and the second phase (in the presence of feedback or no-feedback). Though one might have expected it to be the case, the collected data indicates no significant difference in accuracy of ratings before and after profile updates were completed by participants: Initial Recommendations (n=263) μ=3.73 95% CI [3.633,3.827], Final Recommendations (no-feedback) (n=144) μ=3.72 95% CI [3.6,3.86], Final Recommendations (feedback) (n=119) μ=3.81 95% CI [3.68, 3.94]. As the treatment variable (feedback, no-feedback) by itself is not suffi-

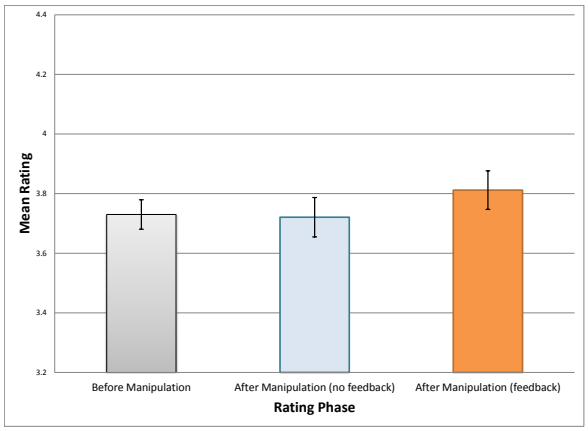


Figure 6: Old and New Profiles: Recommendation Accuracy - recommendations were collected at two points in the study for both treatments; error bars correspond to standard error. User ratings based on original profile data is shown on the far left. The next two bars show a change in ratings after users reported they were satisfied with their profile, broken up by treatment.

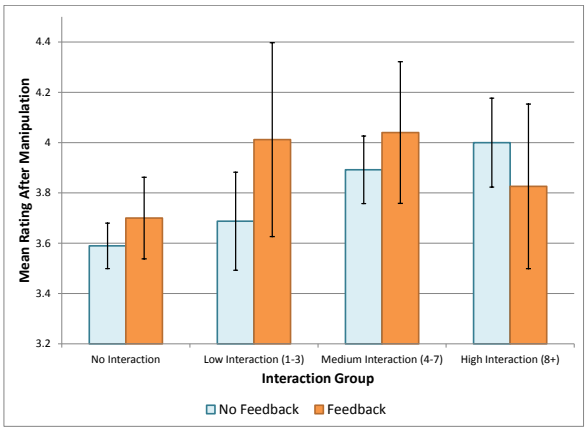


Figure 7: Effects of Manipulation on Recommendation Accuracy - participants binned into groups based on manipulation behavior and split on feedback treatment, error bars correspond to standard error. A trend of increasing rating accuracy can be seen only for the no-feedback group.

cient to examine the effects of manipulation based on the study setup, we performed a more in-depth analysis of recommendation accuracy by examining an interaction effect of the feedback condition and manipulation behavior.

Figure 7 shows a breakdown of how participant manipulation behavior in each treatment affected prediction accuracy. A trend seems to appear for participants in both treatments, with the no-feedback trend appearing more consistently. To verify this result, we ran a joint regression analysis on recommendation accuracy after manipulation against the treatment and observed manipulations. The resulting model showed that each additional manipulation participants made (regardless of treatment) corresponded to a 0.0414 ($p = 0.0155$) increase in rating accuracy (Likert-scale units), while simply seeing the visual recommender corresponded to a one-time 0.24 ($p = 0.0329$) increase in rating accuracy. As expected, interaction propensity and feedback were not separable in our analysis across all participants, as participants using the visual recommender showed much more interaction overall. This result motivated an examination of effect of interaction in each treatment separately. For the no-feedback group in 7, a linear model showed that each additional manipulation a participant made corresponded to a 0.037 increase in rating accuracy ($p = 0.0403$), however, this effect is marginal for the feedback users (0.0165, $p = 0.0825$). In summary, participants in the no-feedback group showed a significant increase in rating accuracy only in proportion to their efforts, while participants in the visual feedback condition appeared to rate items more highly due to a combination of manipulation and satisfaction with the visual feedback.

5.5 Centrality

To better understand any biasing effect the visual recommender had on the feedback group, we re-simulated the experiment, and compared each participant at each time-step with all users in the MovieLens dataset. For each manipulation that was performed by a participant, we re-computed Pearson correlation for the participant's current movie profile and generated a vector of how that correlation changed over time. Figure 8 shows a breakdown of the degree of connectedness when the user reported they were satisfied with their profile. The degree corresponds to the total number of users in the MovieLens database that had at least one item in common with the user after all manipulations had ceased. For both treatments, there is a clear trend towards increasing connectivity. Given the well known issues with cold-start and sparsity in CF systems [18], this is a potentially useful property of the visual recommender tool, but needs closer study to understand the resulting impact on diversity of recommendations.

There are a few possible explanations for the increase in connectivity over time regardless of feedback. Recall that, during the post-phase participants were instructed to manipulate their profiles to generate the best recommendations possible. Users in both groups may have rated popular items in an effort to define their tastes and thus increased their connectivity over time. Another possible explanation is that the system generated movie recommendations during the first rating phase that the users had previously seen and which were immediately added by the participant into their profile. The increasing similarity for users in the feedback treatment reinforces the result in [4] regarding the ability of an interactive recommender to bias a participant's ratings over time.

6. FUTURE WORK

The results from this study suggest that creating recommenders that aid users but minimize biasing effects may be beneficial. According to the model that fits our data, after a certain number of manipulations participants in the no-feedback treatment would have overtaken participants in the feedback group in terms of recommendation accuracy. In a followup study, we will alter the user interface to more closely examine the effects among manipulation quantity, manipulation quality, and the presence of recommendation feedback. We plan a second follow-up study to assess what aspects of the visual recommender capture the attention of and influence the user during interaction. This will be a

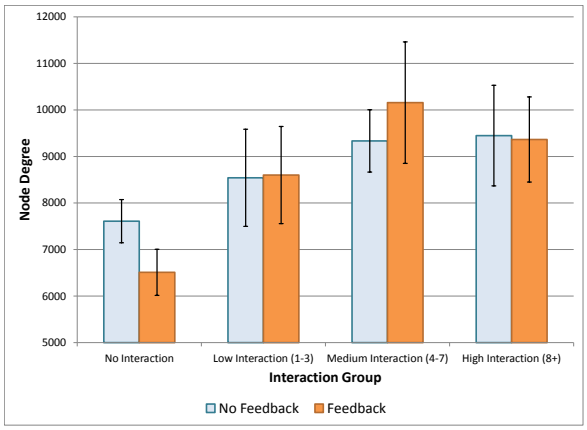


Figure 8: Effects of Manipulation on Connectedness - participants binned into groups based on manipulation behavior and split on feedback treatment; the vertical axis shows the mean number of neighbors that had at least one item in common with the active user. Bars correspond to standard error.

lab-based study performed using eye-tracking equipment to capture gaze patterns on the Fluo interface.

7. CONCLUSION

This paper analyzed the results an experiment (N=263) to evaluate the impact of low-cost, exploratory manipulations on a preference profile for collaborative filtering recommender systems. The experiment tested one condition in which visual feedback on profile manipulations was provided, and one with no feedback. These low-cost manipulations were used to generate "hypothetical recommendations". The experiment recorded four main metrics: degree/type of manipulations, impact on recommendation accuracy, impact on diversity and centrality, and perceived metrics of trust, accuracy and user satisfaction. Our data supports the following claims: (i) dynamic feedback during profile manipulations has a biasing effect on user opinion (this is in agreement with findings by Herlocker in [6]), (ii) profile manipulations tend to increase centrality and reduce diversity of the user profile in both conditions, (iii) participants are more likely to perform profile manipulations in the presence of dynamic feedback from a recommender system, and (iv) recommendation accuracy improves linearly with profile manipulations, however, the effect of each manipulation is stronger in the absence of recommendation feedback. This may be a result of exploratory behavior induced by the interactive visualization.

8. REFERENCES

[1] S. Bostandjiev, J. O'Donovan, and T. Höllerer. Tasteweights: a visual interactive hybrid recommender system. In P. Cunningham, N. J. Hurley, I. Guy, and S. S. Anand, editors, *RecSys*, pages 35–42. ACM, 2012.

[2] C. Boutilier, R. S. Zemel, and B. Marlin. Active collaborative filtering. In *In Proceedings of the Nineteenth Annual Conference on Uncertainty in Artificial Intelligence*, pages 98–106, 2003.

[3] M. Buhrmester, T. Kwang, and S. D. Gosling. Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6(1):3–5, 2011.

[4] D. Cosley, S. K. Lam, I. Albert, J. A. Konstan, and J. Riedl. Is seeing believing?: How recommender system interfaces affect users' opinions. In *Proceedings of the Conference on Human Factors in Computing Systems*, pages 585–592. ACM Press, 2003.

[5] B. Gretarsson, J. O'Donovan, S. Bostandjiev, C. Hall, and T. HÃűllerer. Smallworlds: Visualizing social recommendations. *Comput. Graph. Forum*, 29(3):833–842, 2010.

[6] J. L. Herlocker, J. A. Konstan, and J. Riedl. Explaining collaborative filtering recommendations. In *Proceedings of ACM CSCW'00 Conference on Computer-Supported Cooperative Work*, pages 241–250, 2000.

[7] J. L. Herlocker, J. A. Konstan, L. G. Terveen, John, and T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, 22:5–53, 2004.

[8] Y. Kammerer, R. Nairn, P. Pirolli, and E. H. hsin Chi. Signpost from the masses: learning effects in an exploratory social tag search browser. In *CHI*, pages 625–634, 2009.

[9] B. P. Knijnenburg, S. Bostandjiev, J. O'Donovan, and A. Kobsa. Inspectability and control in social recommenders. In P. Cunningham, N. J. Hurley, I. Guy, and S. S. Anand, editors, *RecSys*, pages 43–50. ACM, 2012.

[10] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, Aug. 2009.

[11] K. McCarthy, J. Reilly, L. McGinty, and B. Smyth. Experiments in dynamic critiquing. In *IUI '05: Proceedings of the 10th international conference on Intelligent user interfaces*, pages 175–182, New York, NY, USA, 2005. ACM Press.

[12] B. Mobasher, R. Burke, R. Bhaumik, and C. Williams. Toward trustworthy recommender systems: An analysis of attack models and algorithm robustness. *ACM Trans. Inter. Tech.*, 7(4):23, 2007.

[13] J. O'Donovan and B. Smyth. Trust in recommender systems. In *Proceedings of the 10th International Conference on Intelligent User Interfaces*, pages 167–174. ACM Press, 2005.

[14] J. O'Donovan, B. Smyth, B. Gretarsson, S. Bostandjiev, and T. Höllerer. Peerchooser: visual interactive recommendation. In *CHI '08: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pages 1085–1088, New York, NY, USA, 2008. ACM.

[15] D. M. Oppenheimer, T. Meyvis, and N. Davidenko. Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, 45(4):867–872, 2009.

[16] G. Paolacci, J. Chandler, and P. G. Ipeirotis. Running experiments on amazon mechanical turk. *Judgment and Decision Making*, 5:411–419, 2010.

[17] R. Rafter, K. Bradley, and B. Smyth. Passive profiling and collaborative recommendation. In *Proceedings of the 10th Irish Conference on Artificial Intelligence and Cognitive Science, Cork, Ireland*. Artificial Intelligence Association of Ireland (AAAI Press), 1999.

[18] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. Grouplens: An open architecture for collaborative filtering of netnews. In *Proceedings of ACM CSCW'94 Conference on Computer-Supported Cooperative Work*, pages 175–186, 1994.

[19] B. M. Sarwar, J. A. Konstan, A. Borchers, J. Herlocker, B. Miller, and J. Riedl. Using filtering agents to improve prediction quality in the grouplens research collaborative filtering system. In *Proceedings of ACM CSCW'98 Conference on Computer-Supported Cooperative Work*, Social Filtering, Social Influences, pages 345–354. ACM Press, 1998.

[20] J. B. Schafer, D. Frankowski, J. Herlocker, and S. Sen. In *The Adaptive Web: Methods and Strategies of Web Personalization*, chapter 9.

[21] R. Sinha and K. Swearingen. The role of transparency in recommender systems. In *CHI '02 extended abstracts on Human factors in computing systems*, pages 830–831. ACM Press, 2002.

[22] N. Tintarev and J. Masthoff. Effective explanations of recommendations: user-centered design. In *Proceedings of the 2007 ACM conference on Recommender systems*, pages 153–156. ACM, 2007.

[23] N. Tintarev and J. Masthoff. A survey of explanations in recommender systems. In *Data Engineering Workshop, 2007 IEEE 23rd International Conference on*, pages 801–810. IEEE, 2007.

[24] K. Verbert, D. Parra, P. Brusilovsky, and E. Duval. Visualizing recommendations to support exploration, transparency and controllability. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces*, IUI '13, pages 351–362, New York, NY, USA, 2013. ACM.