The Value of Recommender Systems: Decomposing the Informational and Discovery Gains

Guy Aridor* Duarte Gonçalves†

Daniel Kluver‡ Ruoyan Kong‡ Joseph Konstan‡

May 11, 2021

Project Outline

1. Overview

Recommendation systems are nearly ubiquitous in the digital economy, from online grocery shopping to the consumption of cultural goods on streaming platforms such as Spotify and Netflix to news platforms. While recommendation systems undeniably affect consumption choices, the mechanisms that drive their impact are not well understood empirically.

In this project we run a longitudinal field experiment in order to decompose the mechanisms that drive the influence recommendation systems have on consumption choices. Our study follows a within-subjects design that randomizes the set of movies that users are exposed to and recommended. This allows us to understand the role that the following mechanisms play in influencing consumption choices.

The first mechanism that could be at play is that recommendation provides consumers with information that changes their beliefs about the match value of an item and may induce them to consume it. Furthermore – as argued in Aridor et al. (2020) –, the consumption of items may induce changes in their beliefs about similar items. However, characterizing user beliefs about their product valuation and causally identifying how such beliefs

*Department of Economics, Columbia University; g.aridor@columbia.edu.

†Department of Economics, University College London; duarte.goncalves.ds@gmail.com.

‡GroupLens Research and Department of Computer Science and Engineering, University of Minnesota - Twin Cities; *Kluver*: kluve018@umn.edu; *Kong*: kong0135@umn.edu; *Konstan*: konstan@umn.edu.

change as a result of recommendation and consumption has not been, to our knowledge, studied empirically.[1]

The second mechanism that could be at play is that recommendation enables consumers to more efficiently search through large product spaces. By expanding users' consideration sets, such systems reduce product discovery costs and enable consumers to discover products that they would hardly find if they were instead randomly searching. Thus, with this interpretation, recommendation provides no informational value, as users can directly infer how much they value the product through evaluation of its features (e.g. looking at genre, actors, tags, reviews), but may have difficulties in doing this across all products and so recommendation allows them to restrict their attention to products likely to be a good fit for them.

In reality, both mechanisms likely play some role. In this project, we aim to identify and decompose the value of recommendation in terms of its informational value and its product discovery value.

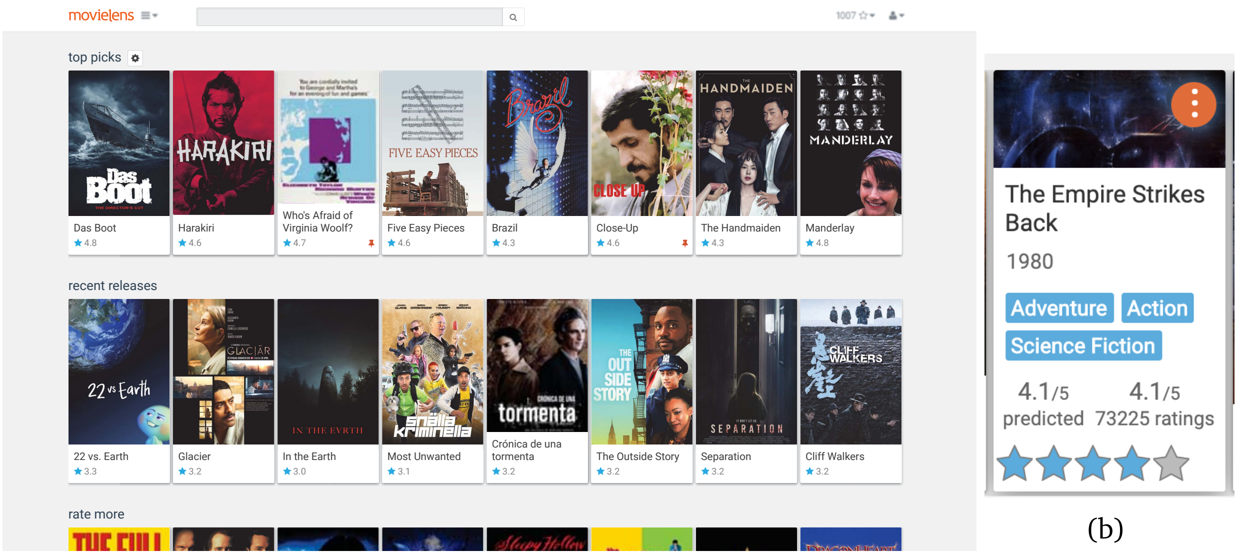
2. Setting

This study relies on an experimental intervention on a movie recommendation online platform, MovieLens, which is utilized by the recommender system community as a benchmark dataset for evaluation of new recommender system algorithms (Harper and Konstan 2015). Users sign up to the platform and initially rate a set of movies and then are presented with an ordered list of lists interface as seen in Figure 1. The first row shows "top picks" which are the top recommended movies for this user, and the rest of the rows are recent releases, unrated movies, and other categories of potential interest. When a user hovers over a movie title, they see the genres of the movies, the predicted rating according to the recommendation algorithm, and the number and average of community ratings for the movie.

The platform is mainly used as a movie discovery tool. Thus, the life-cycle of a user in this context is that they periodically use the platform to find movies to watch and then rate

[1]While there is a voluminous literature that studies psychological aspects of recommender systems, to our knowledge there there has been surprisingly little empirical work that views the problem from the lens of economic theory.

them after watching. If a user clicks into the movie page they see Figure 2, which provides information on tags associated with the movie, derived from the tag genome (Vig et al. 2012), as well as similar movies. As the platform is free to use and noncommercial, users have no reason to not truthfully report their ratings as it is in their benefit to provide the platform with truthful information in order to get the best possible recommendations.



(a) (b)

Figure 1. **MovieLens Interface**

Notes: Panel (a) exhibits the MovieLens home page, where the "top picks" or recommended movies are always at the top. Panel (b) shows the interface experienced when a user hovers over a movie.

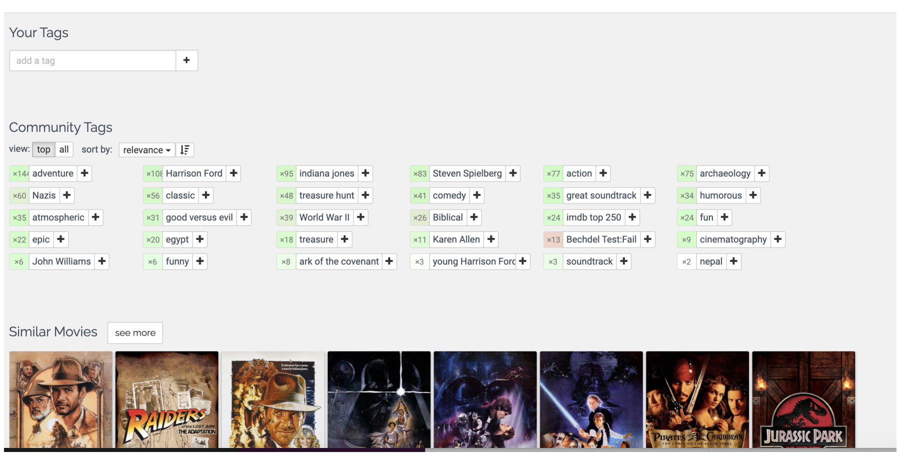


Figure 2. **Product Page - Tags and Similar Movies**

In the context of the platform – and for the purposes of this study – displaying a movie as one of the 8 "top picks" on the very top of the platform's home page is interpreted as a recommendation. This is not only true from a search perspective, as showcasing the movie forces it into the user's consideration set, but also by the fact that it also provides the user with valuable information in the form of the movie's expected rating (as per the platform's prediction algorithm).

3. The Impact of Recommendations on User Consumption and their Beliefs

Our proposed intervention is simple: We elicit users' beliefs about how much they value a set of movies – their expected rating and a measure of uncertainty – and examine how beliefs evolve over time with observed recommendations, individual search, and consumption patterns. Recommended movies correspond to a set of 8 movies that are singled out on the home page. Since it is not feasible to elicit beliefs over the entire set of movies available on MovieLens, we select a subset of movies elicit beliefs as follows.

At the beginning of the experiment, we collect the set of the $3N$ movies with the highest predicted rating for each user i as generated by the platform and according to the recommendation algorithm selected by the user. We then randomly assign these movies to one of three different sets, $X_{i,0}$, $X_{i,1}$, and $X_{i,2}$. The movies in $X_{i,0}$ are set aside and are never shown neither as recommended movies nor will we elicit the user's beliefs about them; this set effectively corresponds to our control group. The set $X_{i,1}$ corresponds to movies about which we elicit the user's beliefs but that are never recommended. Finally, the movies in $X_{i,2}$ are used for both belief elicitation and recommendation. Concretely, this means that only movies from $X_{i,2}$ are shown in the top picks row of the MovieLens homepage. The sets are user-specific, have the same number of elements, and are constructed in a manner that controls for the user's idiosyncratic taste in a manner specified below.

In order to formalize our hypotheses, we now introduce some notation. Let $C_{i,t}$ denote the set of movies watched by user i by to time t and $C^I_{i,t}$ the set of movies watched by the user during the intervention up to time t. We omit the time index to refer to the sets at the end of the intervention. We will call $r^x_{i,t}$ user i's expected rating for movie x at time t as predicted by the platform's algorithm and $b^x_{i,t}$ the same expected rating but as predicted by the user. We also collect user i's self-reported degree of uncertainty regarding $b^x_{i,t}$, which is given by $u^x_{i,t}$. Lower scores of $u^x_{i,t}$ are associated with lower reported uncertainty.

Our first two hypotheses are as follows

Hypothesis 1. Exposure to a movie increases the likelihood that the movie is watched relative to no exposure.

Hypothesis 2. Recommendation of a movie increases the likelihood that the movie is watched relative to mere exposure.

If Hypothesis 1 captures the effect of mere exposure on consumption – reflecting an expansion of the user's consideration set –, Hypothesis 2 focuses on whether recommendations have an impact on consumption beyond mere exposure. We leverage two traits of our design in order to test these hypotheses. First, that by eliciting beliefs about a movie while providing minimal information we obtain as a by-product user exposure to such a movie. In particular, during belief elicitation no information on the movies' platform-predicted rating is provided. Second, we rely on the fact that, by construction, movies are randomly sorted into $X_{i,0}$, $X_{i,1}$, and $X_{i,2}$. Then, testing Hypotheses 1 and 2 corresponds to testing $\mathbb{P}\left(x \in C_i^I \mid x \in X_{i,1}\right) \geq \mathbb{P}\left(x \in C_i^I \mid x \in X_{i,0}\right)$ and $\mathbb{P}\left(x \in C_i^I \mid x \in X_{i,2}\right) \geq \mathbb{P}\left(x \in C_i^I \mid x \in X_{i,1}\right)$, respectively.

If indeed recommendations affect consumption beyond forcing the user to consider a movie, then a natural mechanism underlying this effect is that recommendations affect users' beliefs. By providing the consumer with information on the predicted rating for a given movie, it is reasonable to conjecture that recommendations will not only affect the user's expected rating but especially their degree of uncertainty about how much they would enjoy the movie. This constitutes the core of our third hypothesis:

Hypothesis 3. A recommendation (i) shifts the user's predicted rating towards the platform's predicted rating and (ii) decreases the user's degree of uncertainty about their predicted rating.

By providing information on the (user-specific) platform-predicted movies, we expect users' expected rating to shift toward the platform's prediction. Such behavior has been documented with user rating of already seen movies in (Cosley et al. 2003) who observed that the rating a user gives a movie can be artificially inflated or deflated based on the predicted rating provided by the platform at the time of rating. While the first conjecture being verified could be due to a similar effect, our conjecture about the effect of recommendations on the degree of uncertainty provides support to interpreting a shift in expected ratings following a recommendation as a consequence of the recommendation's informational effect. In order to test these hypotheses, our experimental design explicitly elicits beliefs before and after recommendation for a randomly chosen subset of recommended movies. Then, if the recommendation for movie x occurs at time t, we expect (i) $\mathbb{E}\left[\left|r^x_{i,t} - b^x_{i,t}\right|\right] \leq \mathbb{E}\left[\left|r^x_{i,t-1} - b^x_{i,t-1}\right|\right]$, and (ii) $\mathbb{E}\left[u^x_{i,t}\right] \leq \mathbb{E}\left[u^x_{i,t-1}\right]$. A second way we will test these hypotheses is by comparing the change in the expected rating and in the degree of uncertainty before and after recommendation of movie $x \in X_{i,2}$ with the same change for movie $x' \in X_{i,1}$ for which there was no recommendation. Then, we will assess whether (i) $\mathbb{E}\left[\left|r^x_{i,t} - b^x_{i,t}\right| - \left|r^x_{i,t-1} - b^x_{i,t-1}\right|\right] \leq \mathbb{E}\left[\left|r^{x'}_{i,t} - b^{x'}_{i,t}\right| - \left|r^{x'}_{i,t-1} - b^{x'}_{i,t-1}\right|\right]$, and (ii) $\mathbb{E}\left[u^x_{i,t} - u^x_{i,t-1}\right] \leq \mathbb{E}\left[u^{x'}_{i,t} - u^{x'}_{i,t-1}\right]$, as well as whether the right-hand side of both these inequalities are statistically different from 0.

We then look at whether reported beliefs are of any consequence to understand consumption.

Hypothesis 4.1. Movie watching activity depends on the user's beliefs.

Formally, Hypothesis 4.1 can be assessed by testing for independence between (b^x_i, u^x_i) and $\mathbf{1}_{x \in C^I_i}$ for $x \in X_{i,2}$, e.g. via testing distance correlation (Székely et al. 2007).

We posit a specific relationship between a user's beliefs and their movie-watching activities:

Hypothesis 4.2. The likelihood a user watches a movie is (i) increasing in their expected rating and (ii) decreasing in reported uncertainty.

Hypothesis 4.2(i) reflects an intuitive monotonicity principle: holding uncertainty fixed, the user is more likely to watch movies that the user expects to enjoy more. On the other hand, holding the expected rating fixed, uncertainty aversion underlies the intuition for Hypothesis 4.2(ii). And, in fact, insofar as users are sufficiently myopic in their movie-watching decisions – i.e. the undertaking exploration of the space of movies is not too appealing –, one would expect Hypothesis 4.2 to hold. We test it by assessing whether $\mathbb{P}\left(x \in C^I_i | b^x_i, u^x_i\right)$ is in fact increasing in b^x_i and decreasing in u^x_i.

Our last set of hypotheses pertains user's beliefs across different movies. As we argued elsewhere (Aridor et al. 2020), users' beliefs about valuations of different movies are likely to be correlated. We hypothesize users are likely to expect similar enjoyment from watching similar movies. We will rely on a particular notion of similarity used in previous studies conducted on MovieLens data (Nguyen et al. 2014). This notion of similarity is based on the existence of a finite set of movie tags T and a tag-score function $\tau : X \to \mathbb{R}^{|T|}$ that associates with each movie x a vector $\tau(x)$ describing how well each tag describes the movie. Then, taking any metric d on $\mathbb{R}^{|T|}$, one can obtain a well-defined notion of similarity via an induced pseudometric $s : X \times X \to \mathbb{R}_+$ such that $s = d \circ \tau$. We consider a tag space and a tag-score function τ as developed by the platform, MovieLens, making use of machine-learning techniques applied to content produced by users, including reviews,[2] and, in line with Nguyen et al. (2014), we take d to be the Euclidean metric. Making use of this similarity notion, we formulate the following hypothesis:

Hypothesis 5.1. Beliefs about movies are positively correlated with movie similarity.

In order to test Hypothesis 5.1, we will assess whether $Cov(b_i^x, b_i^{x'})$ and $Cov(u_i^x, u_i^{x'})$ are decreasing in $s(x, x')$.

If beliefs about movies are correlated and the correlation is stronger the more similar the movies are, then, obtaining information about a given movie – be it directly through consumption or through recommendations – will affect users' beliefs about similar movies.

Hypothesis 5.2. Uncertainty about the expected rating

(i) is lower for movies that are most similar to those already watched by the user;

(ii) decreases after watching a similar movie;

(iii) decreases after a similar movie is recommended.

We capture the notion of the watched movies most similar to movie x by considering the set of the n most similar watched movies

$$X_i^n(x) := \arg \min_{X' \subseteq C_i : |X'| = n} \sum_{x' \in X'} s(x, x').$$

[2] See Vig et al. (2012) for details.

With this set, we simply average similarity between movies in $X_i^n(x)$ and x, i.e.

$$\sigma^n(x) := \sum_{x' \in X_i^n(x)} \frac{s(x,x')}{n}.$$

We let n as a free parameter so as to be able to go beyond how similar is the most similar watched movie ($n = 1$) and assess how much exposure the user had to movies in the neighborhood of x by considering larger values for n. We can then test Hypothesis 5.2(i) by considering whether u_i^x is increasing in $\sigma^n(x)$. For Hypotheses 5.2(ii)-(iii), we will test whether uncertainty decreases with movie-watching and recommendation by considering how u_i^x changes when movies similar to x are watched or recommended.

4. Experimental Design

There are two main components to the design: enrollment and intervention. The enrollment phase includes the following:

- Eligibility and recruitment;
- Generation of user-specific sets of movies, $X_{i,0}$, $X_{i,1}$, and $X_{i,2}$.

The crux of our intervention is combined with a data collection exercise; specifically, it comprises:

- Movie-watching survey;
- Belief elicitation survey.

This data collection takes place every time that a user logs into the platform, up to a maximum of once a day. After the conclusion of these surveys, the user accesses the platform and is exposed to movie recommendations.

Both of these components involve a number of procedures that underlie the identification assumptions enabling us to test the hypotheses listed in the previous section. In this section, we describe in detail all these procedures.

4.1. Eligibility and Recruitment

In our intervention, we target a random sample from a subset of the platform's users that we call the set of eligible users. Any individual over the age of 18 is free to sign up to the platform.[3] The users are able to decline to participate in the study and can opt-out at any moment.[4]

We consider a user as **eligible** if the user satisfies all the following conditions below:

(1) the user rated more than 100 movies in total;

(2) the user rated fewer than 3,000 movies in total; and

(3) over the previous m months, the user rated a minimum of $\lceil 1.5m \rceil$ movies, for some $m \in \{1,2,3,4\}$.

Condition (1) is a minimum data requirement so that the recommender system algorithm utilized by the platform is able to provide valuable recommendations. Condition (2) is excluding high-powered users as these would constitute outliers. Condition (3) seeks to guarantee that the targeted user is minimally active on the platform over the recent past. The purpose of this restrictions is to mitigate the heterogeneity of treatment effects across users arising from differences in the quality of the recommendations. This is especially important given that, throughout the duration of the intervention, the recommendations for movies not seen are not updated with the new information. These criteria were chosen in consultation with the platform experts in order to ensure that the data is representative of the overall platform population.

The roll-out of the study is phased in order to control for implementation issues. On March 29th, 100 eligible users are to be randomly selected to participate in the study. On April 5th, the study is expanded to an additional 500 randomly selected eligible users. On April 15th, 4,000 additional eligible users are randomly selected to take part in the experiment. We aim for 1100 participants to opt into the study with an average of 20 survey responses per participant.

[3] The use of or access to the platform is prohibited to individuals under the age of 18, as per the platform's terms of service.

[4] See Figures 3 and 5 in Appendix A for screenshots of the interface.

4.2. Treatment Assignment

We utilize block randomization in order to assign movies to $X_{i,0}, X_{i,1}, X_{i,2}$. In particular, we want to ensure that the quality of the movies in each treatment arm is roughly the same for each user. As a result, we utilize a special case of block randomization by following a matched pairs design where we pair together movies according to a proxy for user-specific product quality, the predicted rating given by the platform's recommendation system. This is implemented as described below.

We sort the movies from 1 to $3N$, with lower values denoting movies with a higher predicted rating. Then, we randomly assign one movie for each subset $\{3n+1, 3n+2, 3n+3\}$ to one of the sets $X_{i,0}, X_{i,1}, X_{i,2}$, for $n = 0, ..., N-1$. When the user declares they have seen a movie in either of this sets, this movie is removed from the set. We set $N = 250$. These sets are held constant throughout the whole intervention.[5]

4.3. Tracking Movies Watched

Given that our hypotheses require us to keep track of when users watch movies, and as the platform does not have this feature built-in, we included it as part of our survey.

Upon signing in, the user is asked whether they have watched any movie since they were last on the platform. The user can search for the movie as they do usually on the platform and a number of options appear. If the user declares they have seen a movie, they are require to rate the movie and provide an approximate date of when they watched the movie. Figure 4 in Appendix A includes screenshots of the interface for this part of the survey.

Belief Elicitation Protocol

We provide additional details on the procedure for belief elicitation.

For each of the 10 selected movies – details below –, we ask the user whether they watched the movie before. If they answer affirmatively, we elicit their rating and for an approximate date of when they watched it, as in the previous section. If they declare not to

[5]In case one of these sets shrinks to less than 50 elements, all the three sets are regenerated; however, we do not expect this to occur. The timing of all these procedures is recorded.

have watched it, we elicit their expected rating based on the movie poster, corresponding to b_i^x. We also ask how certain the user is of their reported expected rating on a 5-point Likert scale, which we take as our measure of uncertainty u_i^x.

The selection of the movies used for belief elicitation is done according to Algorithm 1 described below.

Algorithm 1: Belief Elicitation Sampling Protocol

User logs in;
Survey on movies watched recently as per Section 4.3 ;
Add noise to predicted rating on $X_{i,1,t} \cup X_{i,2,t}$ and compute new sorting;
if $t == 0$ **then**
  $S_{i,\ell,t}$ = Subsample 2 from first 8 from $X_{i,\ell,t}$, $\ell = 1,2$;
  $S_{i,3,t}$ = Subsample 6 uniformly at random from $X_{i,1,t} \setminus S_{i,1,t}$;
  Elicit beliefs on $\bigcup_{\ell=1,2,3} S_{i,\ell,t}$;
end
if $t \geq 1$ **then**
  $X_{i,\ell,t} = X_{i,\ell,t-1} \setminus C_{i,t}$ for $\ell = 0,1,2$, removing movies watched ;
  $S_{i,\ell,t}$ = Subsample 2 from first 8 from $X_{i,\ell,t}$, $\ell = 1,2$;
  $S'_{i,1,t}$ = Subsample 2 uniformly at random from $X_{i,1,t} \setminus S_{i,1,t}$;
  Elicit beliefs on $\bigcup_{\substack{\ell=1,2,3 \\ \tau=t-1,t}} S_{i,\ell,\tau}$.
end

After the belief elicitation page, the user is taken to the platform's home page, where the 8 recommended "top picks" correspond to the top 8 movies in $X_{i,2,t}$, sorted according to the platform-predicted rating r_i^x with additive Gaussian independent noise $\epsilon_{i,t}^x \sim N(0,V)$.[6] The randomness induces variability in the presented set of movies as well as in the set of recommended movies. Recall that the sets of movies used for belief elicitation only, $X_{i,1}$, and for both belief elicitation and recommendation, $X_{i,2}$, were generated in a way that attempts to keep user preferences similar. Thus, the sets $S_{i,1,t}$ and $S_{i,2,t}$ comprise movies that are, in expectation, the same in terms of user preferences.

Another relevant feature of Algorithm 1 is that it features a subset of movies that persists from one period to the next – both movies that are recommended as well as movies that

[6] Trading off the goal of generating enough variation in the set of presented movies and the necessity of keeping recommendations meaningful for the user, we settled on $V = .2$. We converged on this calibration in consultation with the platform's experts and simulations of how such noise would impact the average quality of the recommended movies as well as the number of periods in which there would be overlap of recommendations / elicitations for the set of users targeted for the study.

are not recommended. Hence, we can identify the effect of recommendations on the user's beliefs as we elicit these before and after a movie is recommended.

5. References

Aridor, Guy, Duarte Goncalves, and Shan Sikdar. 2020. "Deconstructing the Filter Bubble: User Decision-Making and Recommender Systems." In *Fourteenth ACM Conference on Recommender Systems*, 82–91.

Cosley, Dan, Shyong K Lam, Istvan Albert, Joseph A Konstan, and John Riedl. 2003. "Is seeing believing? How recommender system interfaces affect users' opinions." In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 585–592.

Harper, F. Maxwell, and Joseph A. Konstan. 2015. "The MovieLens datasets: History and context." *ACM Transactions on Interactive Intelligent Systems (TIIS)* 5 (4): 1–19.

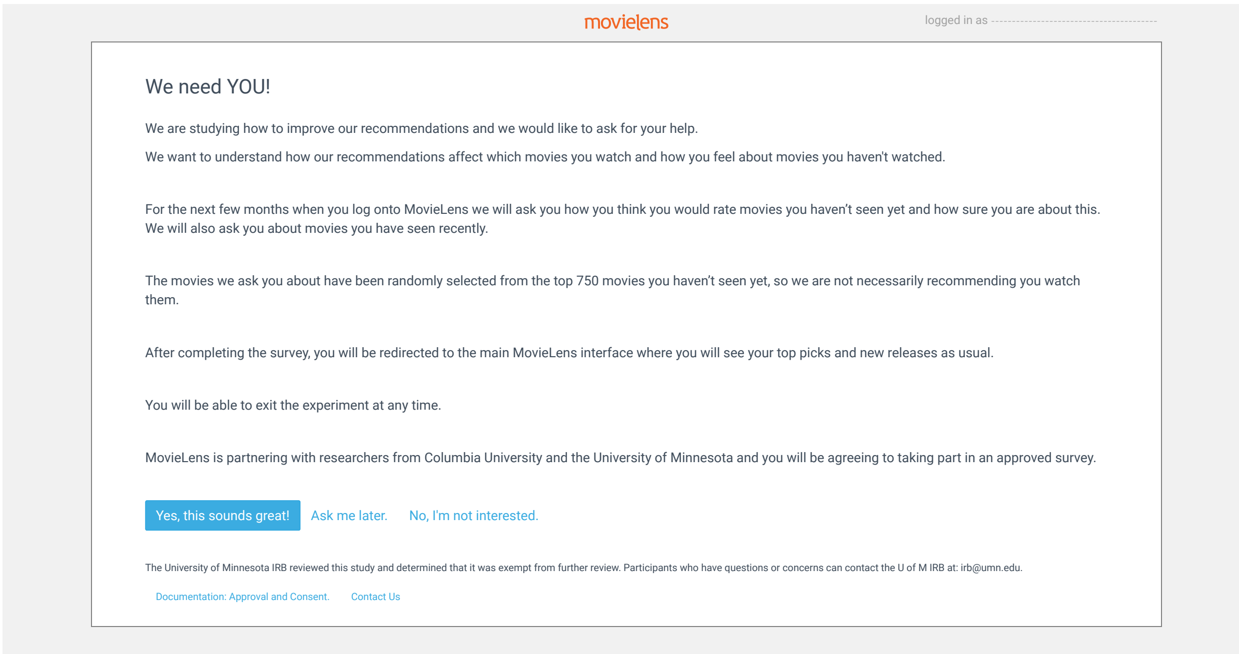
Nguyen, Tien T, Pik-Mai Hui, F Maxwell Harper, Loren Terveen, and Joseph A Konstan. 2014. "Exploring the filter bubble: the effect of using recommender systems on content diversity." In *Proceedings of the 23rd international conference on World wide web*, 677–686, ACM.

Székely, Gábor J., Maria L. Rizzo, and Nail K. Bakirov. 2007. "Measuring and testing dependence by correlation of distances." *The Annals of Statistics* 35 (6): 2769–2794. 10.1214/009053607000000505.

Vig, Jesse, Shilad Sen, and John Riedl. 2012. "The tag genome: Encoding community knowledge to support novel interaction." *ACM Transactions on Interactive Intelligent Systems (TiiS)* 2 (3): 1–44.

Appendices

A. Interface and Instructions



movielens

logged in as ---------------------------------------

We need YOU!

We are studying how to improve our recommendations and we would like to ask for your help.

We want to understand how our recommendations affect which movies you watch and how you feel about movies you haven't watched.

For the next few months when you log onto MovieLens we will ask you how you think you would rate movies you haven't seen yet and how sure you are about this. We will also ask you about movies you have seen recently.

The movies we ask you about have been randomly selected from the top 750 movies you haven't seen yet, so we are not necessarily recommending you watch them.

After completing the survey, you will be redirected to the main MovieLens interface where you will see your top picks and new releases as usual.

You will be able to exit the experiment at any time.

MovieLens is partnering with researchers from Columbia University and the University of Minnesota and you will be agreeing to taking part in an approved survey.

Yes, this sounds great! Ask me later. No, I'm not interested.

The University of Minnesota IRB reviewed this study and determined that it was exempt from further review. Participants who have questions or concerns can contact the U of M IRB at: irb@umn.edu.

Documentation: Approval and Consent. Contact Us

Figure 3. **Consent form**

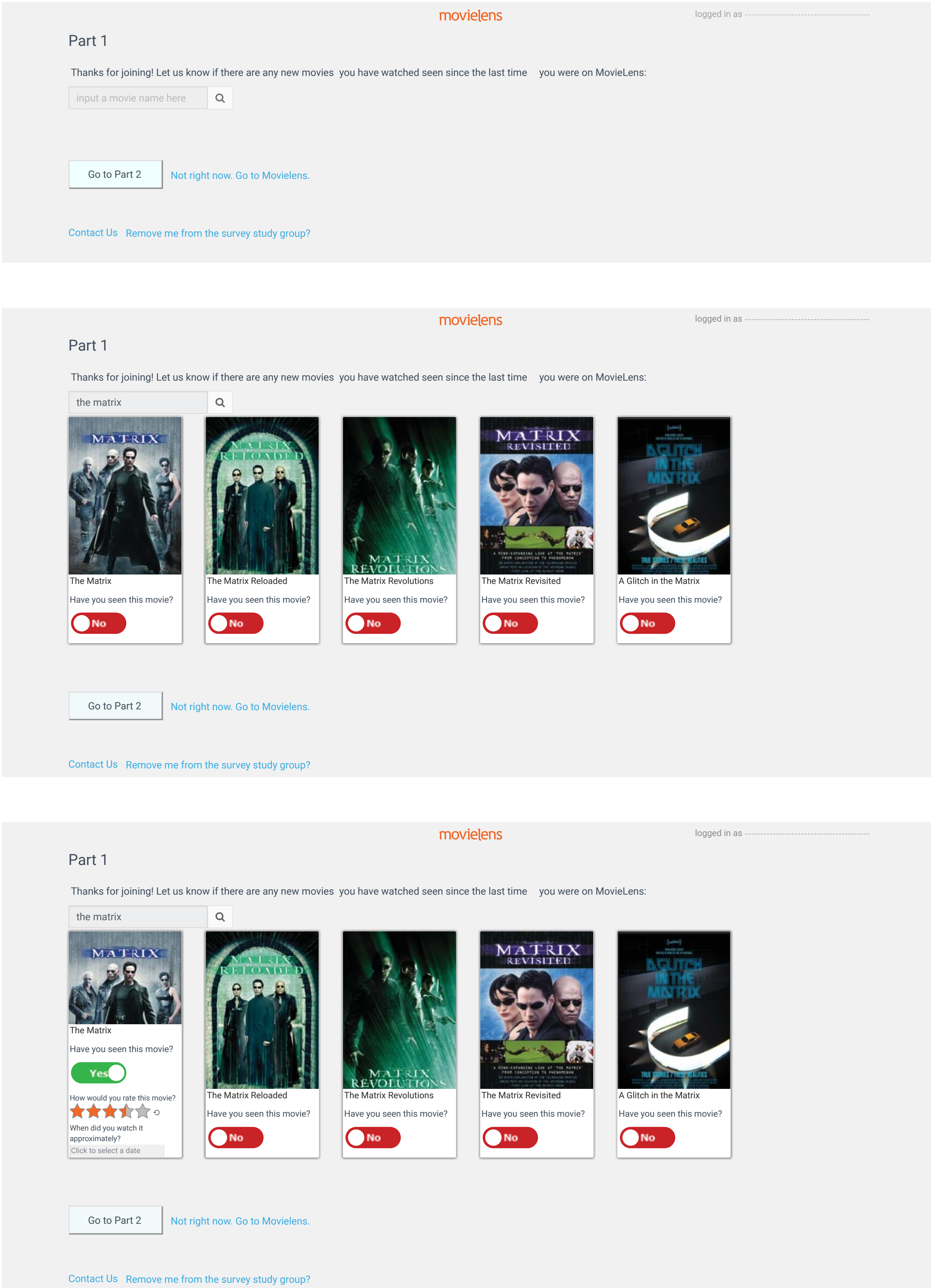


Figure 4. **Tracking Movies Watched**

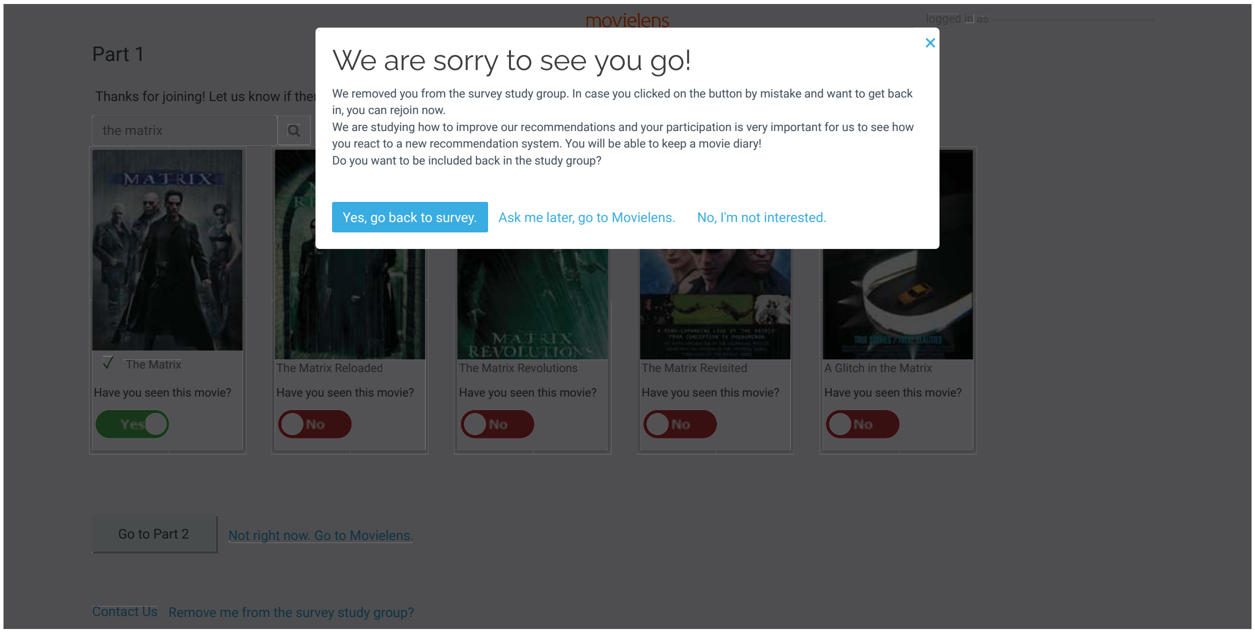


Figure 5. **Opting Out**

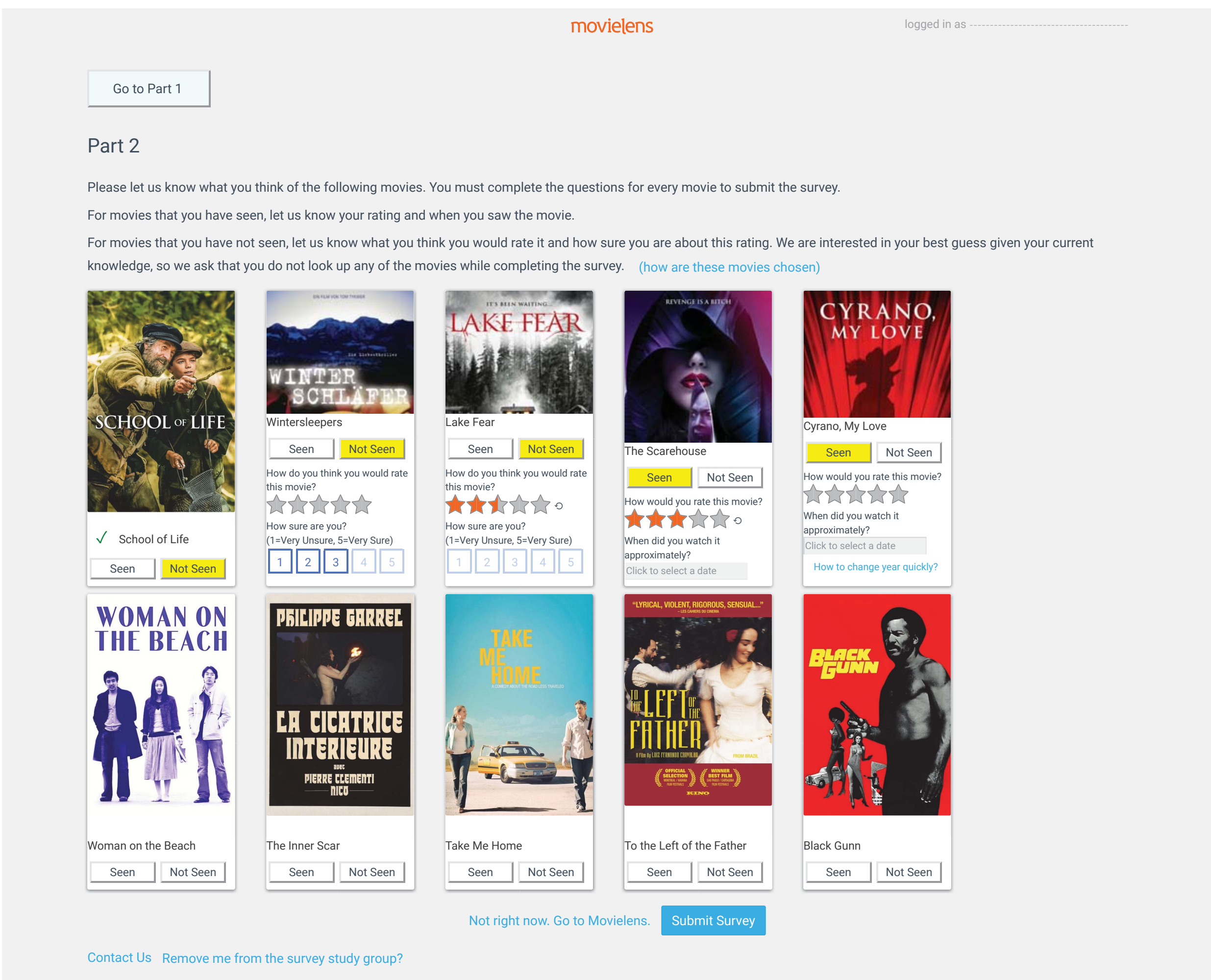


Figure 6. **Belief Elicitation**

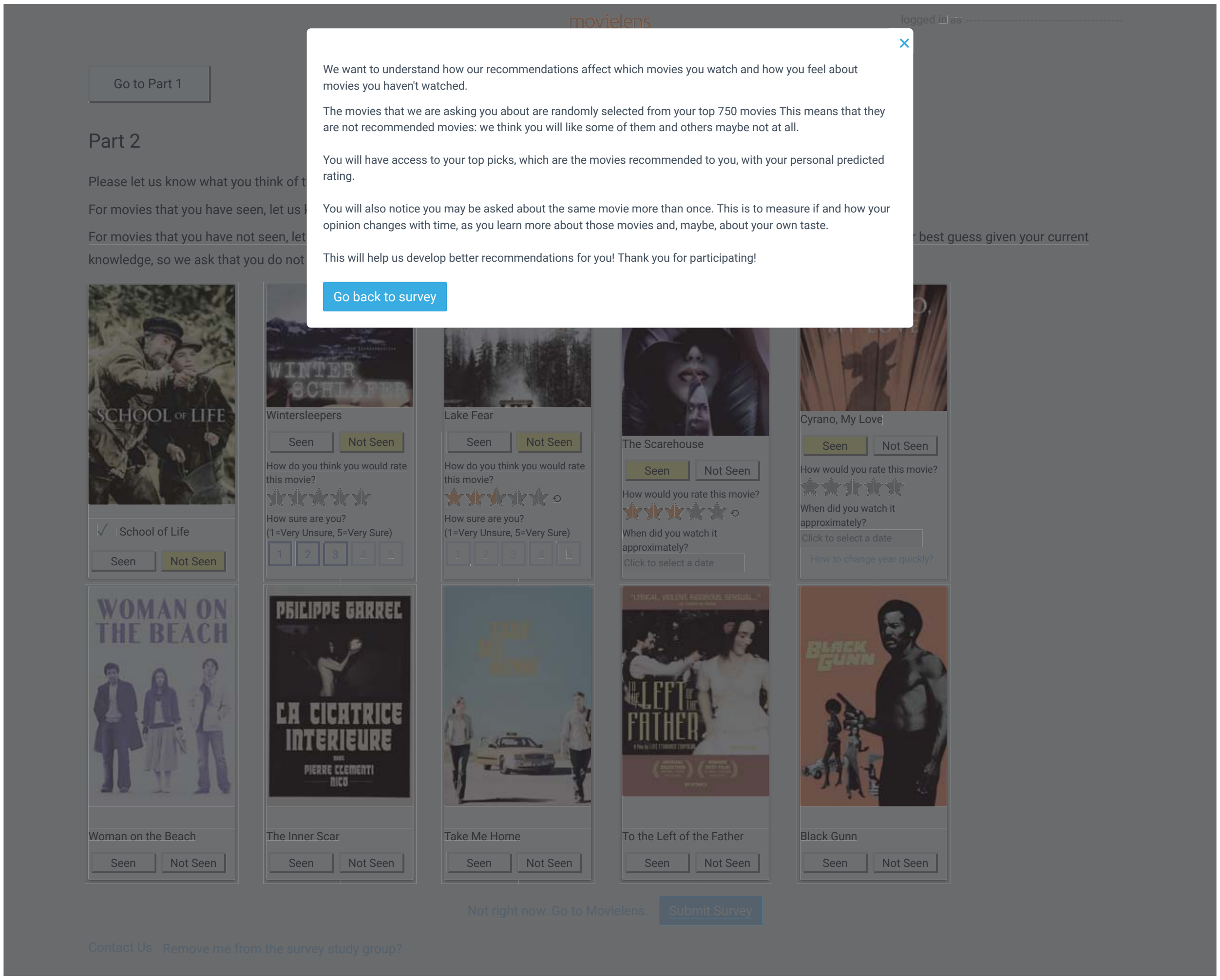


Figure 7. **Information on the sets of movies**