Film Friend Networks: Evidence of Spillover Effects in Film Watching Behavior from Letterboxd.com

Johnny Ma

June 14, 2017

1 Introduction

Numerous papers have identified the importance of friends' behavior and characteristics on one's own behavior. Outside of the typical peer effects in classrooms [1] and other microeconomic environments, there has been substantial work on the effect of a network of interactions on one's behavior. Udry and Conley [2] use a known network of social interactions for farmers in Ghana who are experimenting with a new crop, and find that farmers, who have farmer friends in their network succeed with different amounts of fertilizer, adjust their own fertilizer use to mirror those who are success, and avoid allocations for those who fail. This suggests a type of behavioral learning based on the revealed successes of friends, showing that one farmer's experiments have positive spillover effects on others, who learn in lieu of self experimentation. This holds conditional on the characteristics of both the farmers (users) and the land (items).

This type of learning can be extended to other fields where the outcome of experiments can be observed by the users, again conditional on the characteristics of both the users and items. A good arena to test this would be in fields with goods with subjective quality, such as music, film, food, and other "taste-based" goods. Many economists have provided microeconomic foundations for the process of taste formation [3], and have used pricing data to investigate behavior such as smoking [4]. However, many of the related studies lack the effect of peers on taste formation, a hole that is filled by the social interactions literature that examines the same smoking behavior [5].

Some papers have used models of social learning to examine consumption of taste based goods. Moretti has created a model of film going behavior based on peer effects. The idea of the paper is that when friends enjoy a film more than expected, conditional on the film's characteristics and the friends' taste profiles, they will increase the probability of other friends to watch the film. The paper is mostly concerned with aggregate movie sales, which leaves a lot of the empirical findings to be called into question.

This paper microfounds Moretti's paper by using data from a social network of film watchers, *letterboxd.com*. With data on each individual's film watching behavior, film rating behavior, and their individual friend network, we can create a complete microeconomic dataset. This adds to the literature by again showing that peer effects exist in networks (in this case, endogenous) and influence people's film watching behavior by a significant margin. Though more work needs to be done to identify the coefficients on peer effects in order to disentangle endogenous network formation and the reflection problem from actual peer effects, this paper shows that there is at least a correlation between friend formation and friend behavior on own behavior.

2 Model

Borrowing from both Udry and Moretti's idea of surprise being the important influencer of behavior (or at least an easily formed and salient metric of good versus bad experimentation), the regression we estimate is in the following form:

$$Pr(Watch_{i,j}) = \alpha_i + \beta_1 * s(good)_{k \in i,j} + \beta_2 * taste_i + \beta_3 * film_j + \epsilon_i$$

Where $Watch_{i,j}$ is defined as the probabilty of user i watching film j if they have already not seen film j (in which case, it is removed from the panel, making it unbalanced). α_i are shocks that affect users, such as a new film coming out and everyone going to see it simply because they watch new films. $taste_i$ is the idea of a user's taste profile, and $film_j$ is the film's characteristics. Though ideally one would use detailed data to construct these metrics, such as complete rating behavior or favorite movies for users and popularity, main actor, language, average rating, and director for films, we can also estimate this with a multiple fixed effects regression using users and films as categorical variables.

The coefficient of interest is β_i, which interacts with $s(good)_{k \in i,j}$. This refers to the share of friends k in the friend network of individual i who have seen film j and have rated it above their ($j's$) average rating for movies, suggesting that the film j was surprisingly good. Ideally this would also be conditional on all the characteristics of the films user j has seen, but this is a good estimation. Note that since this is a logit model, we have to pick a time frame in which users are making decisions about watching films. In our paper, we use from January 1st 2017 to June 1st 2017, but any time frame can be chosen.

Thus, β_i would capture the effects of having friends who seen a film and enjoy it more than they usually do. Of course, this specification is subject to both the reflection problem [7], which is the difficulty of identifying the effect of peer's characteristics versus their behavior, and the endogenous network effect, which is related in that users likely form friends based on having similar tastes. These effects make it difficult to identify the specific sign and magnitude of peer effects, likely biasing them upwards. A future work would use an identification strategy developed by Bramouille [8], using a friend-of-a-friend-but-not-my-friend as a but we can still say something about the structure of the friend networks.

3 Data

We decided to use Letterboxd.com, a small but vibrant film rating site. movie rating system has additional data that can ameliorate these issues. Letterboxd is a small but growing community of film fanatics, where users can rate movies, post reviews, and follow/like other users and their reviews. Since this is mainly a diary-style rating platform for movie enthusiasts, one can expect less sparsity and more truthful review/rating behavior, and not call into question the incentives of individuals to bias their ratings in any systematic way. In addition, a user can identify themselves by inputting their four "favorite movies." These favorites give a lot of information about the user supplied by the user himself that might be useful to construct the user taste profile to control for. Do they love older films, horror films, Japanese animation, etc. In addition, there is a watchlist that many users have that indicates future desire to watch movies. There is also more item (movie) information for each entry, such as genre tags, runtime, year and country of release, a delineated histogram of ratings (from 0.5 to 5 stars, in 0.5 star steps), and all crew and cast names, which is useful for creating the film characteristic controls.

In addition, users can log their film watching behavior using a "diary." This diary page contains all the users' watched films, the day they watched the film, if it was a rewatch, and their rating. Most users accurately record all their watching behavior, as this is the main purpose of the site. This provides an insanely detailed account of all users's film watching behavior, allowing us to run micro-level analysis.

To complement this data, users can also friend other users to "follow" them. Some features on the site involves the social networks. The homepage shows recently watched movies and reviews from friends, each film shows a users' top friend's ratings' for said film, and one can following friends to get notified of their watching behavior. Though it is difficult to understand the friending behavior, one would imagine that one would friend those with either similar taste, one's desired taste, or a funny reviewer. In any case, this is similar to people's following of film critiques that agree or disagree in a concise, eloquent, and interesting way. For the selection of users we chose, the median number of friends is 37, with a maximum of 500 and a minimum of 0 (these were dropped).

A negative of using this very detailed dataset involves having to use Data Scrappers to extract this information, as Letterboxd has no workable API. However, all users are in a large directory sorted by activity, and the site is wonderfully built with easy html tags for all possible data of interest. Overall, a great data source for investigations in rating systems. A sample profile is here: {letterboxd.com/johnnyma}. More figures of the site can be found in the appendix.

In the end, we web scrape the website to gather data. Functions for getting a user's watched films, diary, list of ratings (by which we use to calculate average rating and surprise of each film), and friend networks, as well as a film's characteristics (actors, director, etc.). We use 5000 users and 700 films, which are taken based on popularity. This number can be easily extended, and there might be more interest in looking at older films or less popular users, which might better estimate the effect we are looking for (rather than capturing only large personalities or recently popular films that everyone sees).

The desired data format would be the following: a panel of the 700s for each user, with each 700 containing the share of friends who found each film surprisingly good, and dummies for film and user. Of the 700, we would drop films that the user had already seen, as rewatch behavior is different as they already have their own expectations (though this might be a very fruitful and interesting extension) One would then run a multiple group fixed effect regression using R and the LFE package to estimate the effect of the shares on probability of watching.

4 Results and Discussion

As mentioned above, we ran the regression in our model section, using dummies for films and users. The following table are our results:

Table: Multiple Group Fixed Effects Model

	Dependent variable:			
	Watched in 2017 If Not Seen Already			
	(1)	(2)	(3)	(4)
shares	0.062^{***}	0.052^{***}	0.055^{***}	0.047^{***}
	(0.0004)	(0.0004)	(0.0004)	(0.0004)
Constant	0.024^{***}			
	(0.0003)			
Fixed effcts?	No	User	Film	Both
Observations	903,700	903,700	903,700	903,700
R^2	0.024	0.035	0.031	0.040
Adjusted R^2	0.024	0.035	0.031	0.040
Residual Std. Error	0.210 (df = 903698)	0.208 (df = 903696)	0.209 (df = 903696)	0.208 (df = 903694)
Note:				*p<0.1; **p<0.05; ***p<0.01

The shares coefficient is the β_i that was previously modeled. The regression adds fixed effects until the final column (4), which is the full specification. For the first regression there is a constant, which contains the film and user data. Note that observations is the value of 700*5000. The R-squared value is increasing by adding more fixed effects, suggesting that the fixed effects do have explanatory power. In addition, the coefficients are decreasing, another trend would would expect to find when including more controls.

Since the coefficient is 0.047 and highly significant, we can conclude that there is a correlation between friend's surprise of a film and a users' own probability of having seen that film (had they not already seen it) within the last six months by 4.7%. This is not the identified effect of peers, but rather captures both the idea of endogenous network formation and the reflection problem, meaning this tells us only about how people form friends.

There are multiple possible extensions. For one, a more detailed look into how individuals form friends (do they friend those with similar tastes or watching behavior?) can help alleviate the endogenous network formation confounder. In addition, using a more advanced specification, as aforementioned, can successfully isolate and identify peer effects in a social network. The criteria for this is having a friend-of-a-friend-who-is-not-your-friend, which is likely easily found in this dataset. In addition, we can exploit the variation in reviews for each user. Since for each film, reviews are populated first by three of a users' friends reviews (ranked by popularity) and then by popular reviews (overall site wide). If one has a friend who would be in the popular reviews section, another popular review would be populated. It is likely that this variation could be exploited to identify the differences between the peer effect of a popular reviewer and a review of a friend. In addition, we can try to estimate some idea of a social multiplier if we have an accurate magnitude of peer effects.

Other extensions would involve creating better controls for film and user characteristics. We can take advantage of the detailed data set to interact a users' favorite films with the characteristics of the films they are considering watching. For example, if a user notes that their favorite films are in French and contain Lea Seydoux, they would be more likely to watch French films or films that contain Lea Seydoux. We can thus build indicator dummy variables to control for these films that are in the users "taste" that they would likely watch in a world devoid of peer effects.

Another idea would be to use different slices of data. For now we are investigating popular users watching popular films, which are also likely to be recent films. If we were interested in looking at a more even environment, we would use moderately popular users who are considering watching old movies based on their friends' reviews. Both are valid and interesting, with the first possibly having a stronger peer effect (ala Moretti) but a weaker identification strategy, and vice versa with older films.

5 Conclusion

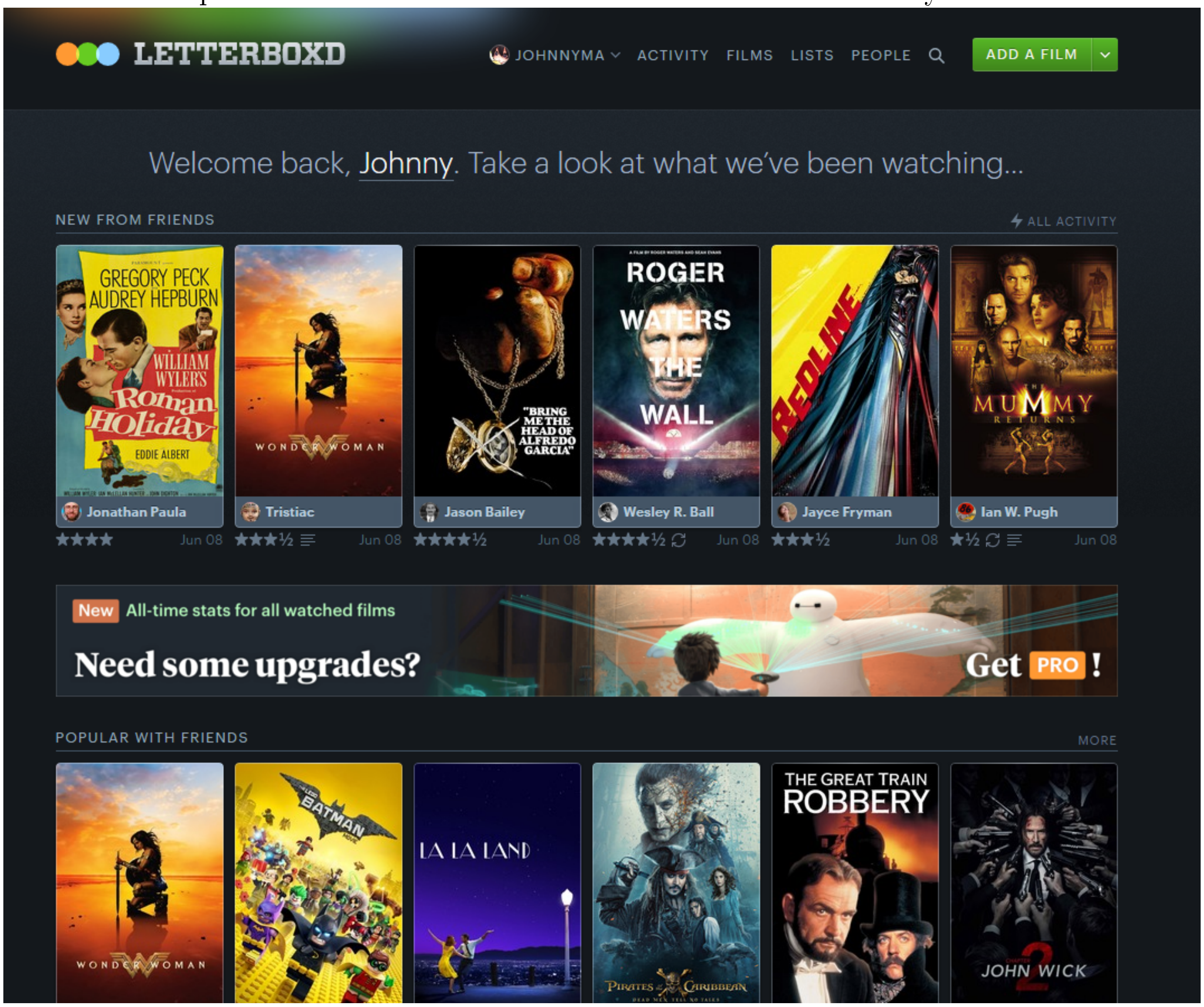
Using a microfounded dataset of film watching behavior in a social network context, we can use Moretti's model of social interactions to estimate peer effects or network effects of friends' surprise on a user's watching probabilty. Though these estimates are subject to identification problems, we find a positive correlation between friends' surprise and one's own behavior. Future extensions can add much more detail and nuance to this analysis, but the core idea of using a detailed dataset to look at social interactions remains within this paper.

6 Acknowledgements

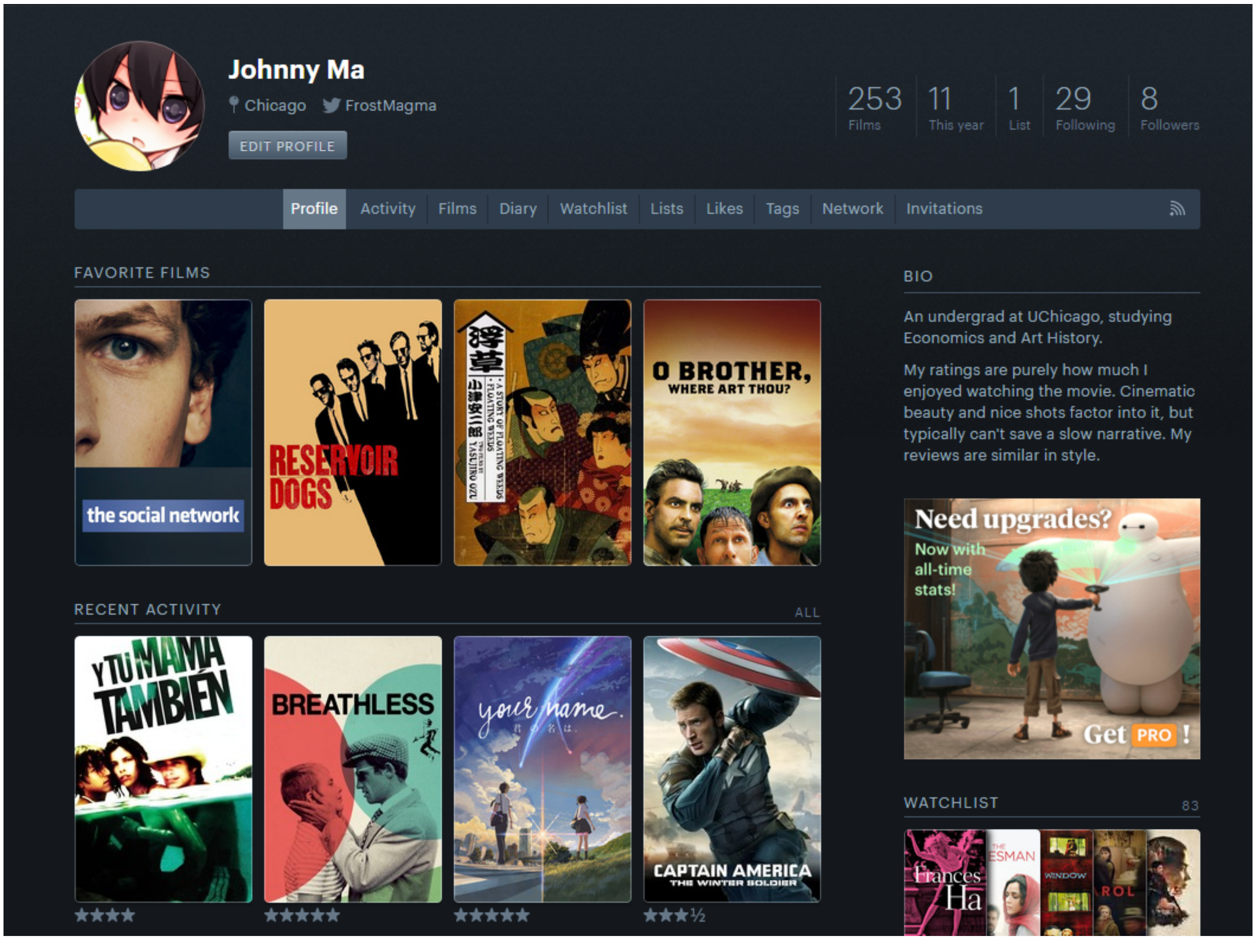
Thanks to Sylvia Klosin for help in creating the model and providing help with the regression. Thanks to Paul Beckman for general intuition about film watching. Thanks to William Jones for general support. Thanks to Professor Manresa for her great support in this project. Thanks for reading and thanks for the great quarter! This author learned a lot about identification of spillover effects and the difficulty of producing research in a short amount of time.

7 Appendix

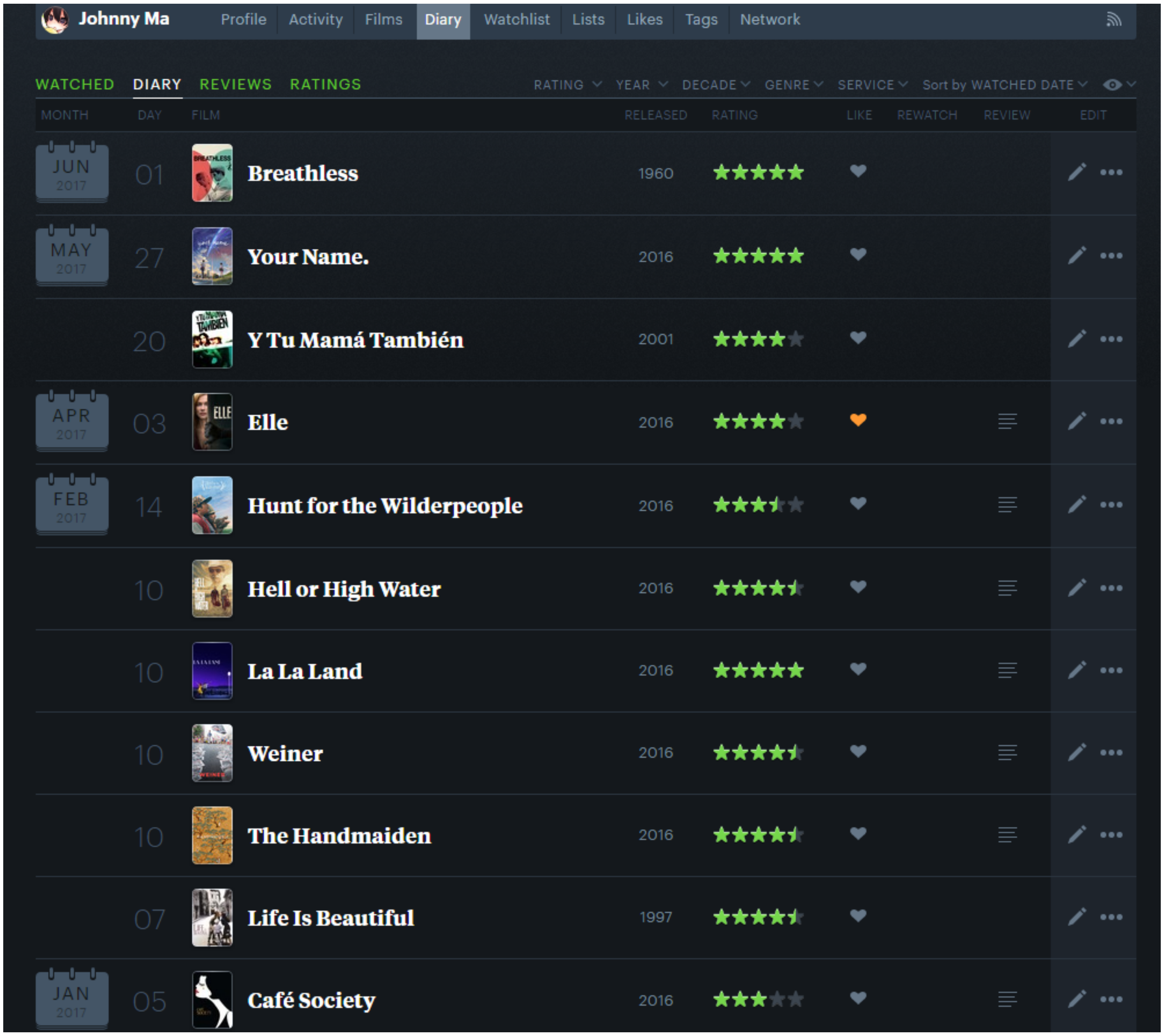
Here are some pictures of the format of the website and some summary statistics.



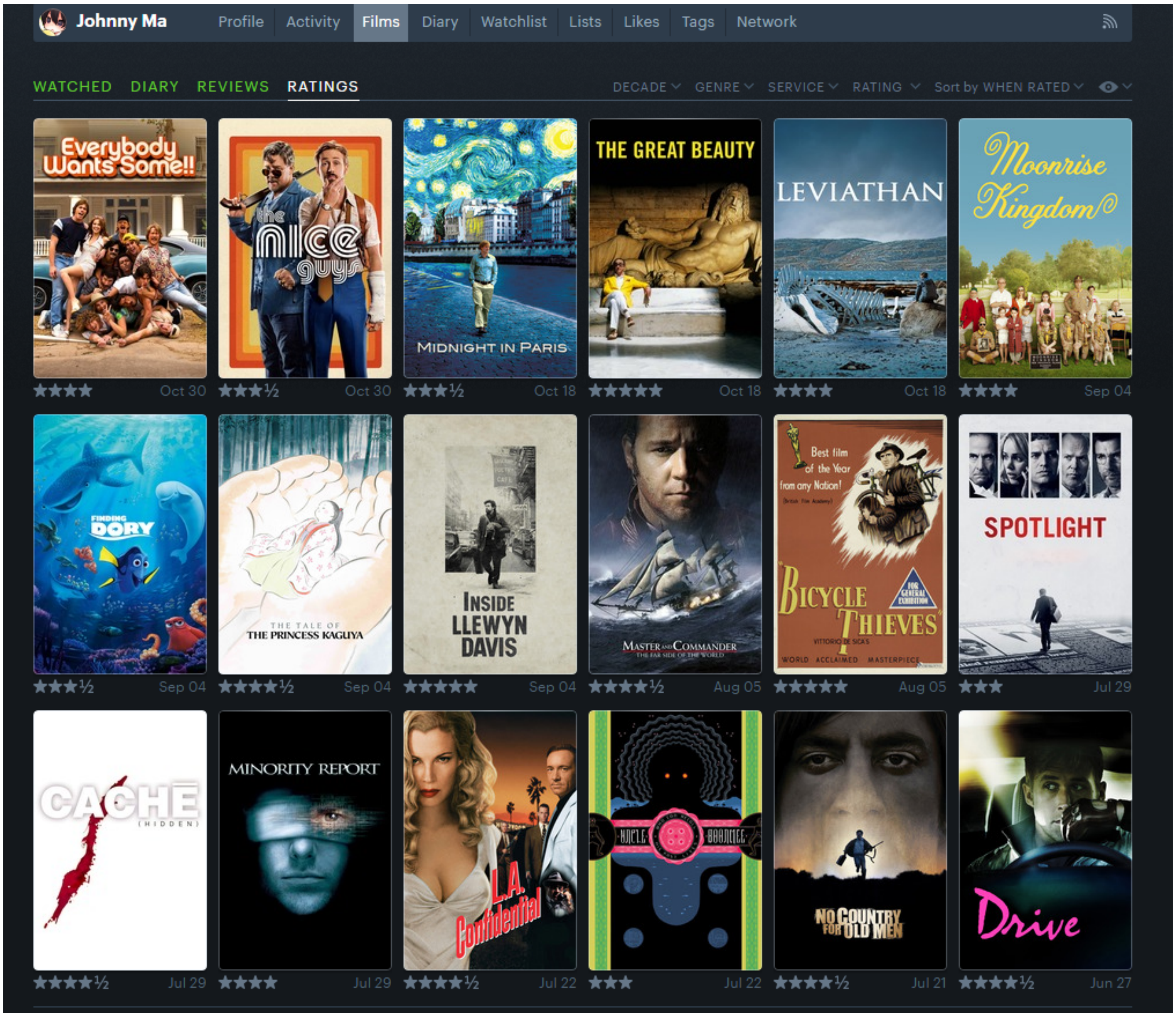
The Homepage of Letterboxd.com



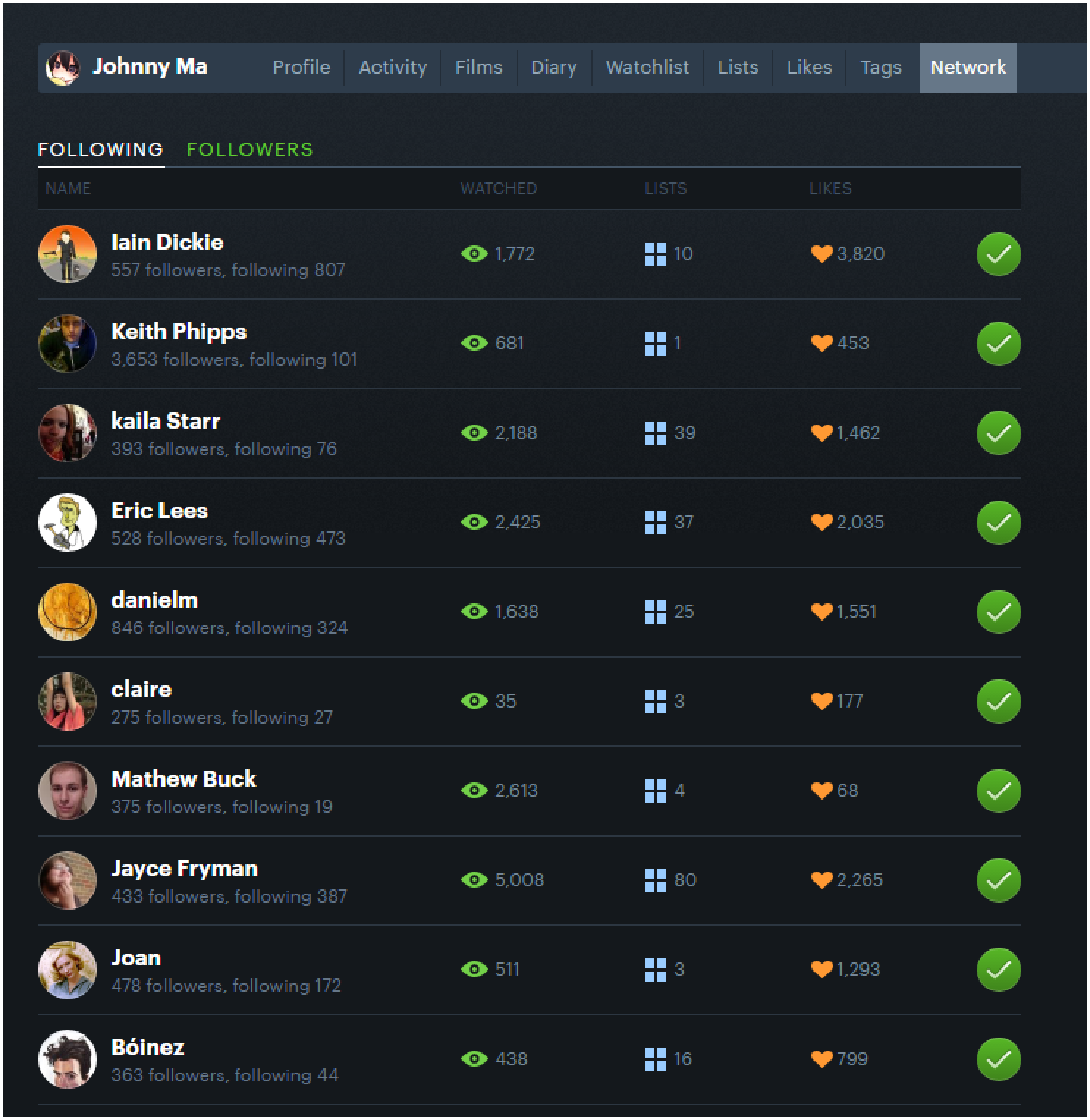
A User's Page. Note the Favorite Films.



A user's diary page.

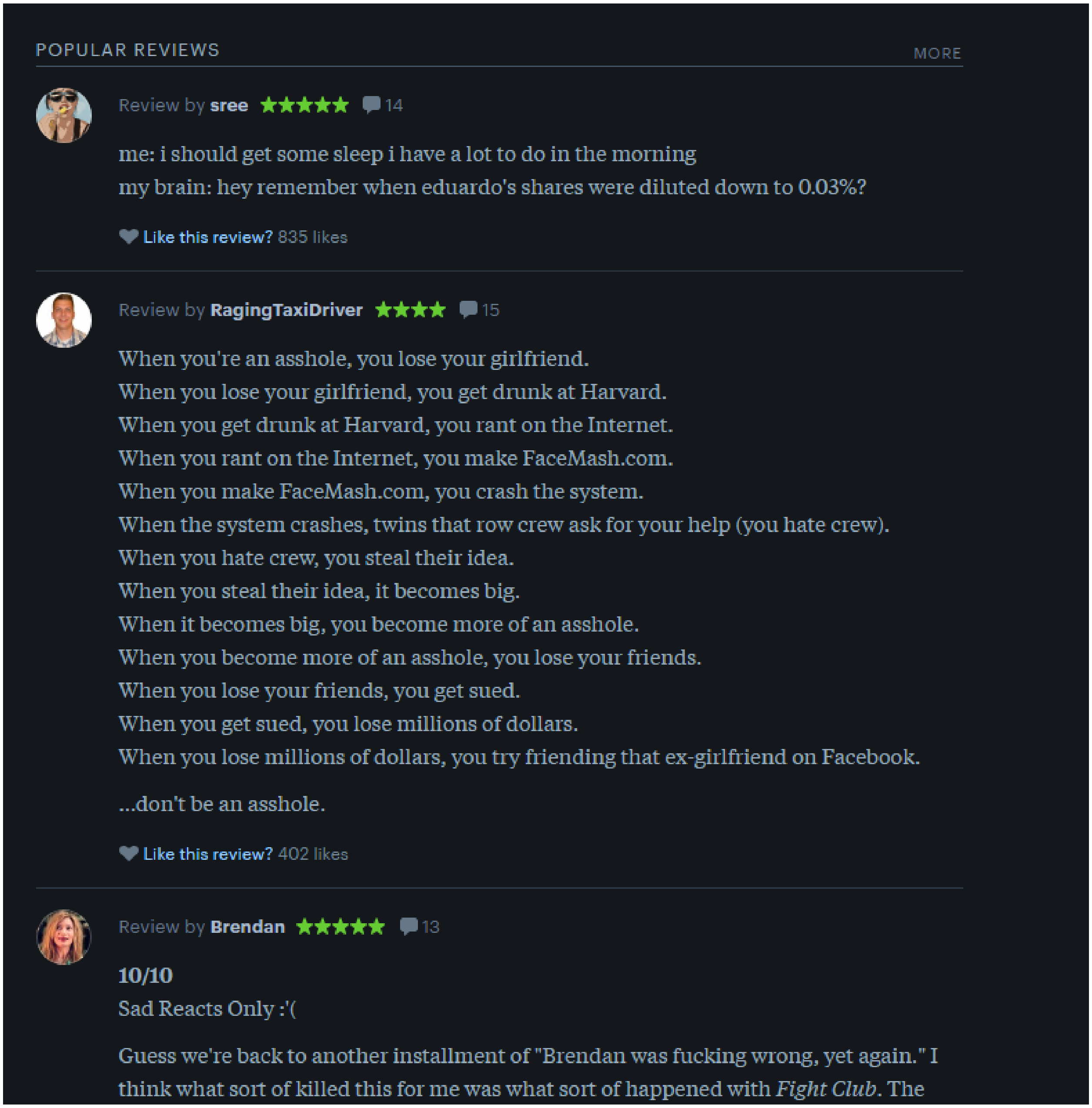


Pages of films and ratings for an individual.



A list of a user's friends.

A sample film page with great detail.


POPULAR REVIEWS MORE

Review by **sree** ★★★★★ 14

me: i should get some sleep i have a lot to do in the morning
my brain: hey remember when eduardo's shares were diluted down to 0.03%?

Like this review? 835 likes

Review by **RagingTaxiDriver** ★★★★ 15

When you're an asshole, you lose your girlfriend.
When you lose your girlfriend, you get drunk at Harvard.
When you get drunk at Harvard, you rant on the Internet.
When you rant on the Internet, you make FaceMash.com.
When you make FaceMash.com, you crash the system.
When the system crashes, twins that row crew ask for your help (you hate crew).
When you hate crew, you steal their idea.
When you steal their idea, it becomes big.
When it becomes big, you become more of an asshole.
When you become more of an asshole, you lose your friends.
When you lose your friends, you get sued.
When you get sued, you lose millions of dollars.
When you lose millions of dollars, you try friending that ex-girlfriend on Facebook.

...don't be an asshole.

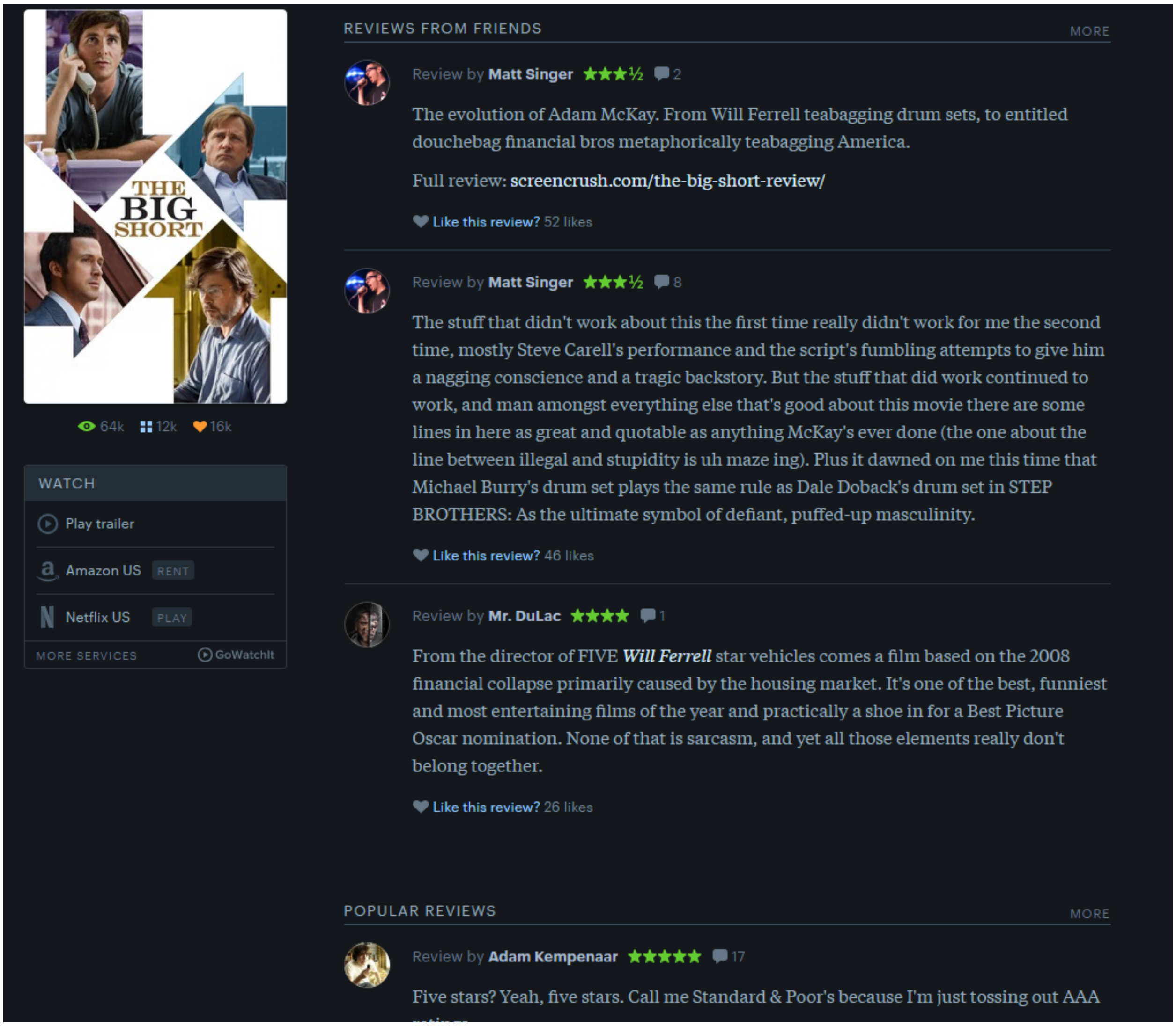
Like this review? 402 likes

Review by **Brendan** ★★★★★ 13

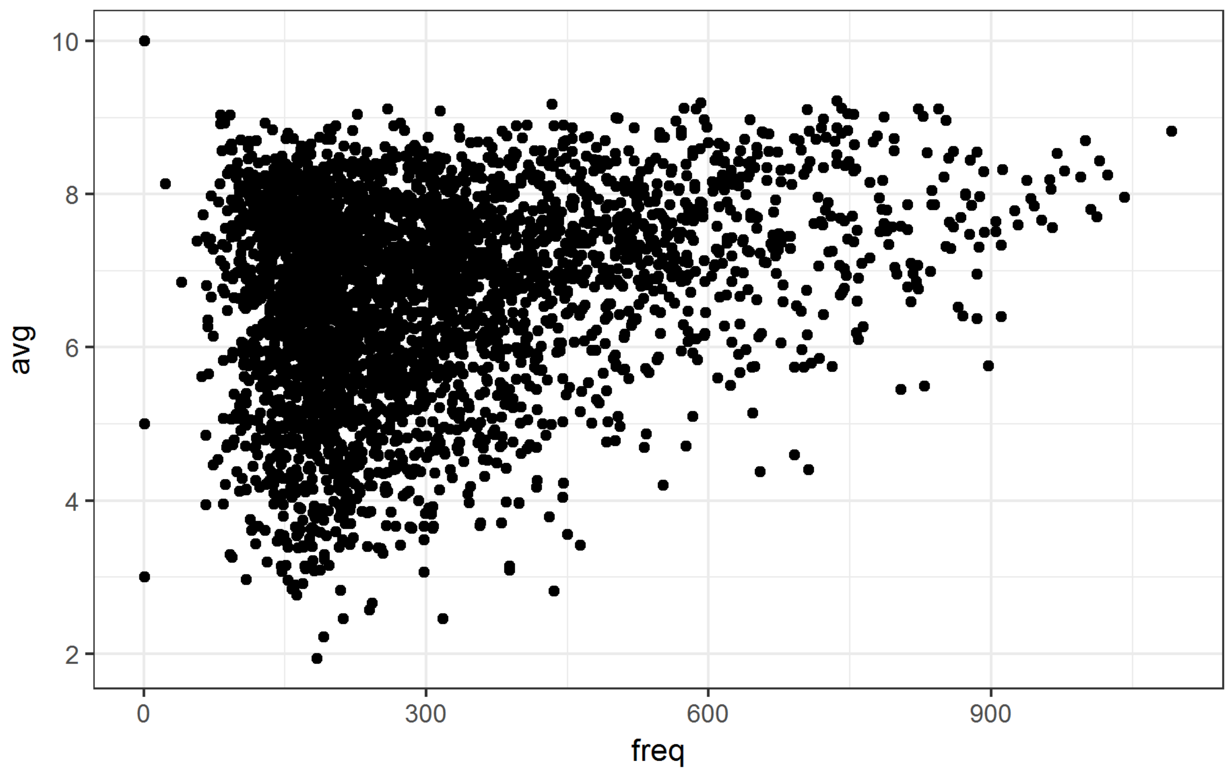
10/10
Sad Reacts Only :'(

Guess we're back to another installment of "Brendan was fucking wrong, yet again." I think what sort of killed this for me was what sort of happened with *Fight Club*. The

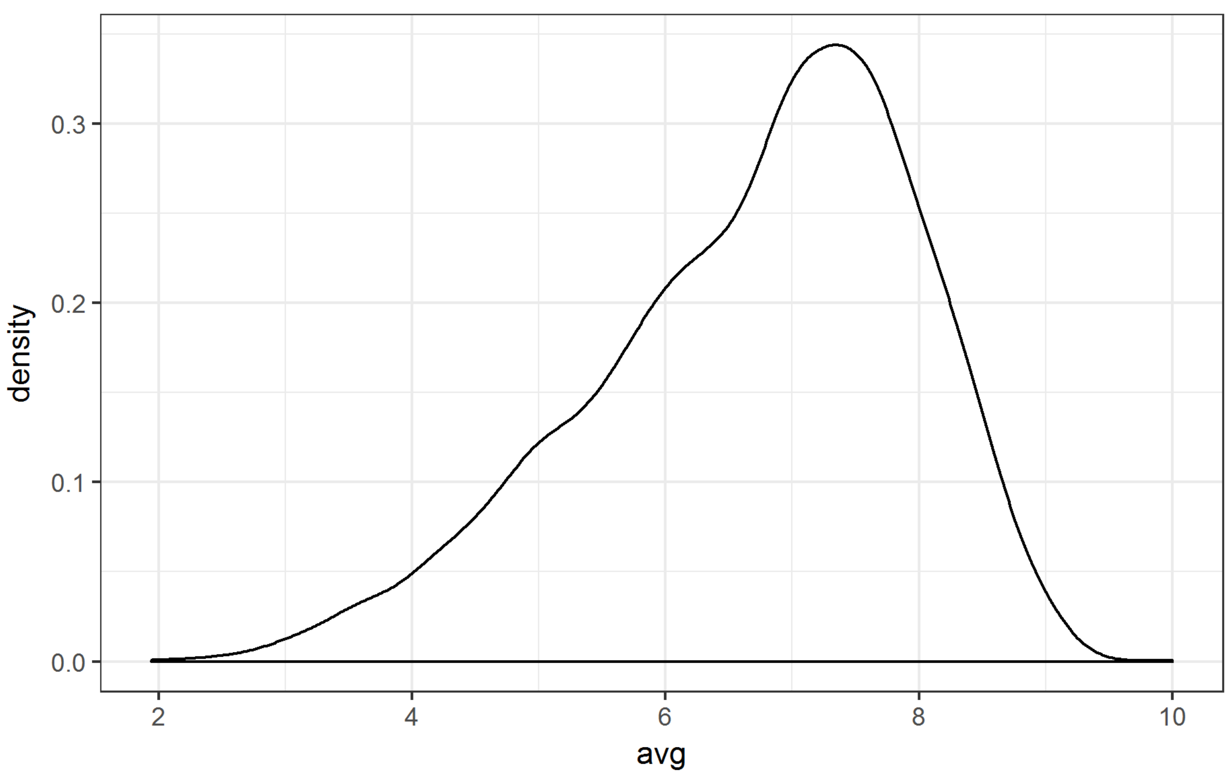
Sample reviews based on popularity.



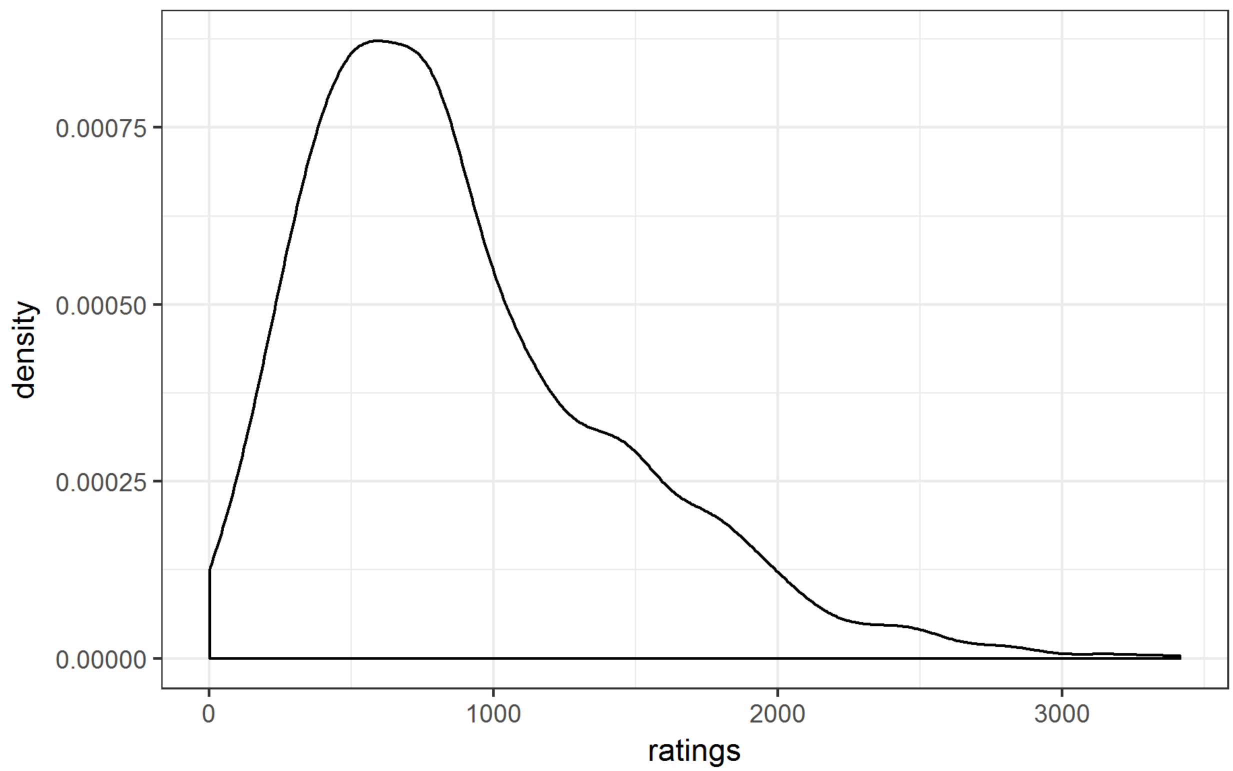
Sample review based on friend list and popularity.



a frequency map of num_ratings and ratings. upwards trending.



A density of ratings. Typically right skewed.



A density of number of films watched. Average is around 800, which is over a month of films.

References

[1] Sacerdote, Bruce. "Peer effects with random assignment: Results for Dartmouth roommates." The Quarterly journal of economics 116, no. 2 (2001): 681-704.

[2] Conley, Timothy G., and Christopher R. Udry. "Learning about a new technology: Pineapple in Ghana." The American Economic Review 100, no. 1 (2010): 35-69.

[3] Becker, Gary S., and Kevin M. Murphy. "A theory of rational addiction." Journal of political Economy 96, no. 4 (1988): 675-700.

[4] Becker, Gary S., Michael Grossman, and Kevin M. Murphy. "Rational addiction and the effect of price on consumption." The American Economic Review 81, no. 2 (1991): 237-241.

[5] Cutler, David M., and Edward L. Glaeser. Social interactions and smoking. No. w13477. National Bureau of Economic Research, 2007.

[6] Moretti, Enrico. "Social learning and peer effects in consumption: Evidence from movie sales." The Review of Economic Studies 78, no. 1 (2011): 356-393.

[7] Manski, Charles F. "Identification of endogenous social effects: The reflection problem." The review of economic studies 60, no. 3 (1993): 531-542.

[8] Bramoullé, Yann, Habiba Djebbari, and Bernard Fortin. "Identification of peer effects through social networks." Journal of econometrics 150, no. 1 (2009): 41-55.