Netflix case study: Genre vs decade

Alan Kreher, Chantal van de Luijtgaarden, Dion Verschuren, Dylan van Wonderen, Suzanne Ekhart, & Tawab Ghorbandi

[1] Jheronimus Academy of Data Science

[2] Eindhoven University of Technology

[3] Tilburg University

Netflix case study: Genre vs decade

Contents

Citations in papaja, detele appropriately later

Add the bibtex entry in the .bib file. You can find the entries in Google scholar, but double check since it is not always correct.

Call the citations in the text:

Citation within parentheses (Aust & Barth, 2020)

Multiple citations (Aust & Barth, 2020; R Core Team, 2021)

In-text citations Aust and Barth (2020)

Year only (2021)

Only if your citation appears in the text it will also show up in the Reference list. Don't manually modify the Reference list.

Executive Summary

(150 words) – 0.3 POINTS Summarize the report. Write this as the very last thing.

What is the main topic you are addressing?

what are your research questions and hypotheses?

what are your results and the main conclusion?

Introduction

Main topic. In this paper, we are going to study the presence of social networks within a movie streaming platform. We're focusing on the structure of links among a group of social players, which consist of users watching and rating movies on Netflix.

Users of Netflix's movie recommendation algorithms are frequently given specific questions about their interests for certain items (which they provide by liking or disliking them, for example). These choices are then immediately integrated into the underlying learning system for future suggestions. If a recommender system starts promoting unwanted products after incorporating new preferences, the user may try to steer the system in the future by correcting it or supplying alternate preference information.

Importance. It is important to study the presence of these socials networks because this could potentially improve the recommender engine that this currently in place. For example, if you know that a user is likely to like a movie that other users with the same "liking profile" also like, you can recommend that movie to the user. When these connections are studied thoroughly, you could have a high probability that the recommendation is successful. This could have a large impact on a movie streaming platform.

Existing studies. In this paper, we will be looking into the Netflix Price Dataset. In 2006 Netflix decided to start a competition with a grand prize of 1 million US dollars. The goal of the competition was to create a collaborative filtering algorithm to predict user ratings for films, based on previous ratings without any other information about the users or films. In order to win you had to at least improve on Netflix's own algorithm by 10%. During the competition a lot of literature has emerged about the dataset and the competition. (Bell & Koren, 2007; Narayanan & Shmatikov, 2006; Takács, Pilászy, Németh, & Tikk, 2008)

However, no papers or any other literature can be found on network analysis on this dataset. We like to fill this gap in the literature by analysing the network and network structure that arises from the subset of the data that we will use. (Guillory & Bilmes, 2011) developed an active movie recommendation system for Netflix. They found that a recommender system should not constantly ask questions to a user, because those reduces the user's mental image of how the recommendation system learns, prompting some participants to "lose track of what they were teaching." According to Amershi, Cakmak, Knox, and Kulesza (2014), this was because users are not always eager to act as simple oracles (repeatedly telling the recommendation system whether they like something or not). This is interesting to take into account for our research, because this would mean that a social network within a movie recommendation system can never be fully exposed.

Research Questions and hypotheses. In this dataset, we can easily see connections between users and movies, but not between just the users or just the movies. At least not, when we do not include one or the other. That is why we have defined the following research questions:

- RQ1: What are the effects of the genre on the likelihood of having both movies watched by the same user?
- RQ2: How much does liking the same movies influence disliking the same movies, and vice versa? (In a network of frequent reviewers)

In order to be able to answer these research questions, we have made set up the following hypothesis accordingly:

- Research Question 1:
  - Comparable movies, in terms of either genre and/or year, are more often watched by the same user than uncomparable movies.
  - Movies of genre horror are generally rated higher than other movie genres.
- Research Question 2:
  - Users that like the same movies to a certain degree, are likely to dislike the same movies to that same degree
  - Users that dislike the same movies to a certain degree, are likely to like the same movies to that same degree

Methodology

Dataset. The primary dataset that is being used during this study is provided by Netflix (REF: KAGGLE) . Netflix shared this data for the Netflix Prize competition where people could use the data to improve their recomender system for movies. The data

includes ratings for 17770 movies from 480189 users. In total, it contains around 24 million ratings on a scale from one to five and the date of the rating. For the movies, the dataset only contains the title of the movie and the year it was released. This was too little information about the movies for our project. Therefore, we decided to look for an additional data source to enrich the dataset. We used Amazon Prime movie genres to add genres to the dataset for further analysis.

The large size of the dataset made it difficult to work with and made running analysis models infeasable. Therefore, we decided to only use a selection of the ratings. The dataset has ratings from October 1998 to December 2005. For this analysis, only the data from December 2001, Januari 2002, and februari 2002 are considered. We chose this winter because this is the first time where users started to rate a good amount of movies while the number of ratings still did not completely explode yet.

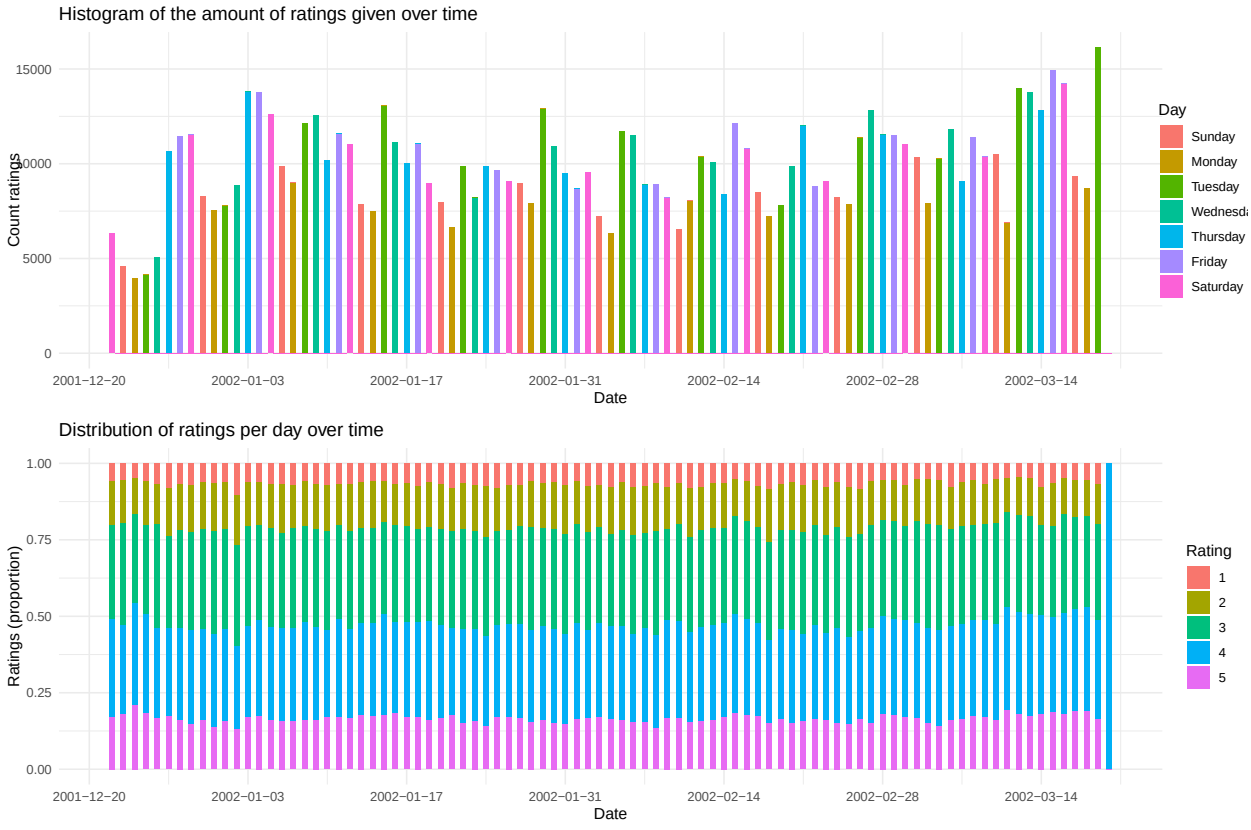
To answer our two research questions we had to transform our data further in two different ways. For the first research question, the network is defined as movies for the nodes and the movies have an edge between them if there is an user that likes both movies. Liking a movie is defined as a user giving a 4 or 5 rating for a movie. Of course, modelling the network this way means that movies that are highly rated are always connect. Therefore, the network only included movies that have between 20 and 50 ratings. That means that our analysis only concerns niche movies with a small number of ratings. Moreover, there is a lower threshold too to make sure that we only include movies that have at least a few connections. All in all, the network is just under 500 nodes and the attributes for the nodes are the year of release of the movie and the main movie genre that comes from Amazon Prime.

Descriptives.

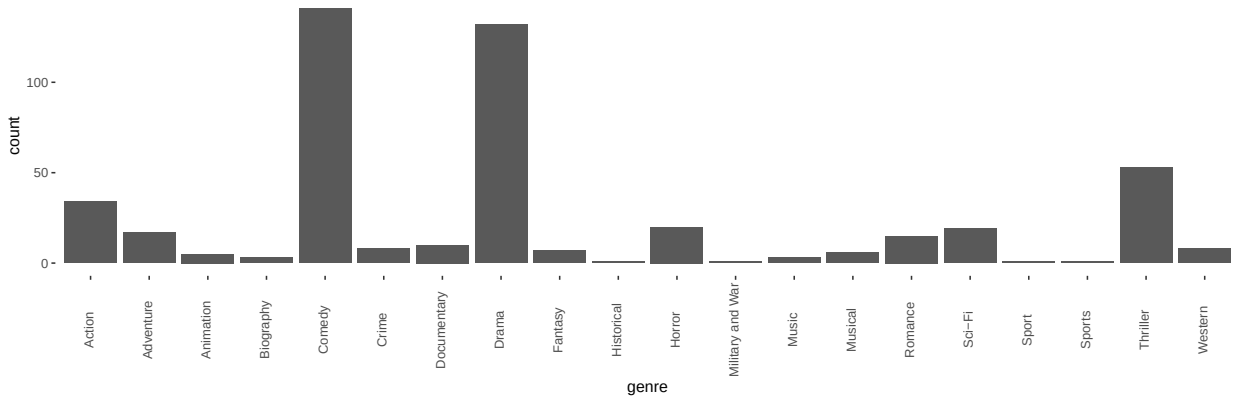
Looking at the average Rating of the users in the data, the average rating they gave was about a $3,36$.

	mean	sd	min	max	n
Rating	3.362346	1.114827	1	5	864610

Looking at the histogram of the amound of ratings given on a certain date we can see that a peak in ratings exists always in the middle of the week (around wednesday/thursday). Looking at the distribution of the ratings over time we don't see great differences. Over time, people give consistent ratings.



The genres in the data were determined using Amazone Prime. The first genre that Amazone Prime displayed is taken as the genre of the movie. As the genres that were displayed in Amazone Prime were not on alfabethic order, we assume that the first genre given is the main genre. 20 genres were found in the Netflix dataset. Most of the movies had *comedy* or *drama* as genre.



Data analysis (Research Rationale).

ERGM.

Many metrics like density and centrality are used to describe the structural characteristics of an observed network. These metrics, on the other hand, describe the observed network, which is just one of many possible alternative networks. The structural characteristics of this group of alternative networks may or may not be similar. A statistical model must consider the set of all possible alternative networks weighted in their similarity to an observed network to credit a statistical inference about the processes that influence the formation of network structures.

Network data, on the other hand, violates the independence and identical distribution assumptions of statistical models like linear regression because they are inherently relational. Alternative statistical models must reflect the uncertainty associated with a given observation and allow inference about the relative frequency of theoretically important network substructures. Exponential Random Graph Models (ERGM) help with this.

Using ERGMs is a powerful technique for conducting statistical inference on network data. ERGMs are a good choice as a network modelling framework for cases where the outcome of interest is in the presence or absence of edges.

QAP.

In general, QAP is very interesting, when you have two or more networks and wants to compare those. In addition, we also choose QAP regression, since we thought it would be interesting to use this model, based on its permutation of actually observed graphs and therefore staying closer to the nature of our networks.

There are three different QAP models from which one had to be chosen for our data and research question. The first is the simple QAP test where you measure the association between two networks. Furthermore, there is a QAP linear model, which can be used on a valued dependent network and one or more explanatory networks. Finally, there is the QAP logistic model, which differs from the linear model, in terms of a binary-valued dependent network.

Our weights in our network are not binary so the logistic model is not an option for us. Even though, we only have two networks we decided to go for the linear model since we would like to see if we could "explain" the first network with the second and vice versa.

As QAP is the only method (to our knowledge) that is able to compare two networks to each other, there aren't any other methods available to address the second research question.

Results

QAP Model.

Constructing data and the model.

During preprocessing of the data, an interesting artifact was encountered; When users were removed from the dataset that either liked or disliked a movie, it was clearly shown that it occurred much more often that user gave a 4/5 star rating to a movie than a 1/2 star rating. This is interesting to keep in mind, as this might pose a potential bias in the dataset.

To give a weight to the edges in terms of importance, we calculate the weight according to a formula:

$$weight = max((\frac{no.of mutual liked movies}{no.of liked movies}) * 100)$$

Results.

We have generated results for different (filtered) versions of the dataset. The results are explained and shown below.

More than 120 reviews.

In order to get a feeling of the QAP regression model, and to get initial results in a reasonable amount of time, we first did the analysis on users who had 120 reviews in total(like and dislike) or more. The results are shown in the figure below.

```
> liked <- read.csv("C:/Users/s169096/Documents/Master JADS/Semester 2.1/SNA4DS/lmatrix_f_divr120plus.csv", sep=",", row.names=1)
> disliked <- read.csv("C:/Users/s169096/Documents/Master JADS/Semester 2.1/SNA4DS/dmatrix_f_divr120plus.csv", sep=",",row.names=1)
> MoviesCor <- sna::netlm(y = disliked, x = liked, mode="graph", nullhyp = 'qapspp', reps = 1000)
> summary(MoviesCor)

OLS Network Model

Residuals:
      0%       25%       50%       75%      100%
-11.496787  -6.328172  -2.264800   3.974733  79.101477

Coefficients:
             Estimate  Pr(<=b) Pr(>=b) Pr(>=|b|)
(Intercept) 8.0746828 1       0       0
x1          0.0633723 1       0       0

Residual standard error: 8.932 on 339074 degrees of freedom
Multiple R-squared: 0.002776    Adjusted R-squared: 0.002773
F-statistic: 943.8 on 1 and 339074 degrees of freedom, p-value:     0


Test Diagnostics:

       Null Hypothesis: qapspp
       Replications: 1000
       Coefficient Distribution Summary:

      (Intercept)        x1
Min      20.139223 -29.111215
1stQ     37.710394  -6.739122
Median   44.611241  -0.239652
Mean     44.921034  -0.007435
3rdQ     51.484472   6.553481
Max      78.481738  29.388775

> MoviesCor <- sna::netlm(y = liked, x = disliked, mode="graph", nullhyp = 'qapspp', reps = 1000)
> summary(MoviesCor)

OLS Network Model

Residuals:
      0%       25%       50%       75%      100%
-18.179288  -5.376102  -0.463702   4.886696  43.755297

Coefficients:
             Estimate    Pr(<=b) Pr(>=b) Pr(>=|b|)
(Intercept) 14.89430499 1       0       0
x1           0.04379977 1       0       0

Residual standard error: 7.426 on 339074 degrees of freedom
Multiple R-squared: 0.002776    Adjusted R-squared: 0.002773
F-statistic: 943.8 on 1 and 339074 degrees of freedom, p-value:     0


Test Diagnostics:

       Null Hypothesis: qapspp
       Replications: 1000
       Coefficient Distribution Summary:

      (Intercept)        x1
Min      284.1323 -30.2007
1stQ     301.1376  -6.7663
Median   306.4277   0.3186
Mean     306.4597   0.1508
3rdQ     311.3878   6.9230
Max      331.2790  27.0493
```

Figure 1

As can be seen from the figure, the QAP regression model explains only 6% of the variation in the like/dislike interaction network for movies when looking at liked movie ratings. When looking at disliked movie ratings, it becomes even less: 4%.

More than 100 reviews.

Next, we decreased the filter threshold for the number of ratings that users have given in the dataset. This increased the amount of datapoints in the dataset. So, we had a dataset with users who had 100 reviews in total (like and dislike) or more, and ran the analysis. The results are shown in the figure below.

```
> liked <- read.csv("C:/Users/s169096/Documents/Master JADS/Semester 2.1/SNA4DS/lmatrix_f_dlvr100plus.csv", sep=",", row.names=1)
> disliked <- read.csv("C:/Users/s169096/Documents/Master JADS/Semester 2.1/SNA4DS/dmatrix_f_dlvr100plus.csv", sep=",",row.names=1)
> MoviesCor2 <- sna::netlm(y = disliked, x = liked, mode="graph", nullhyp = 'qapspp', reps = 1000)
> summary(MoviesCor2)

OLS Network Model

Residuals:
      0%       25%       50%       75%      100%
-12.029353  -6.525862  -2.378062   3.758205  92.231650

Coefficients:
            Estimate   Pr(<=b) Pr(>=b) Pr(>=|b|)
(intercept) 6.91614982 1       0       0
x1          0.09468894 1       0       0

Residual standard error: 8.871 on 634499 degrees of freedom
Multiple R-squared: 0.006349    Adjusted R-squared: 0.006347
F-statistic:  4054 on 1 and 634499 degrees of freedom, p-value:     0


Test Diagnostics:

       Null Hypothesis: qapspp
       Replications: 1000
       Coefficient Distribution Summary:

      (intercept)       x1
Min       27.7449 -48.5160
1stQ      53.0622  -7.3631
Median    61.5625  -0.3786
Mean      61.3501  -0.2169
3rdQ      69.6276   6.9325
Max       95.8010  37.3833

> MoviesCor3 <- sna::netlm(y = liked, x = disliked, mode="graph", nullhyp = 'qapspp', reps = 1000)
> summary(MoviesCor3)

OLS Network Model

Residuals:
      0%       25%       50%       75%      100%
-19.0220686  -5.5297453  -0.5297453   4.8725523  44.4702547

Coefficients:
            Estimate   Pr(<=b) Pr(>=b) Pr(>=|b|)
(intercept) 13.9933485 1       0       0
x1           0.0670496 1       0       0

Residual standard error: 7.465 on 634499 degrees of freedom
Multiple R-squared: 0.006349    Adjusted R-squared: 0.006347
F-statistic:  4054 on 1 and 634499 degrees of freedom, p-value:     0


Test Diagnostics:

       Null Hypothesis: qapspp
       Replications: 1000
       Coefficient Distribution Summary:

      (intercept)       x1
Min       427.6007 -42.8043
1stQ      452.0640  -6.7107
Median    458.2859   0.3391
Mean      458.3345   0.5225
3rdQ      464.4623   8.1611
Max       485.5872  35.6632
```

Figure 2

As can be seen from the figure, the QAP regression model explains only 9% of the variation in the like/dislike interaction network for movies when looking at liked movie ratings. When looking at disliked movie ratings, it becomes even less: 6%.

More than 80 reviews.

Because of the improved results in the last section, we decreased the filter threshold for the number of ratings that users have given in the dataset even more. This again increased the amount of datapoints in the dataset, and we did the analysis on users who had 80 reviews in total (like and dislike) or more. The results are shown in the figure below.

As can be seen from the figure, the QAP regression model explains 14% of the variation in the like/dislike interaction network for movies when looking at liked movie ratings. When looking at disliked movie ratings, it again becomes a bit less: 10,4%.

```
> liked <- read.csv("C:/Users/s169096/Documents/Master JADS/Semester 2.1/SNA4DS/lmatrix_f_dlvr80plus.csv", sep=",", row.names=1)
> disliked <- read.csv("C:/Users/s169096/Documents/Master JADS/Semester 2.1/SNA4DS/dmatrix_f_dlvr80plus.csv", sep=",",row.names=1)
> MoviesCor4 <- sna::netlm(y = disliked, x = liked, mode="graph", nullhyp = 'qapspp', reps = 1000)
> summary(MoviesCor4)

OLS Network Model

Residuals:
      0%       25%       50%       75%      100%
-14.009907  -6.351469  -2.509418   3.503782  93.661771

Coefficients:
            Estimate  Pr(<=b) Pr(>=b) Pr(>=|b|)
(Intercept) 5.6144856 1       0       0
x1          0.1447486 1       0       0

Residual standard error: 8.906 on 1245829 degrees of freedom
Multiple R-squared: 0.01511     Adjusted R-squared: 0.01511
F-statistic: 1.911e+04 on 1 and 1245829 degrees of freedom, p-value:     0


Test Diagnostics:

        Null Hypothesis: qapspp
        Replications: 1000
        Coefficient Distribution Summary:

       (Intercept)       x1
Min       52.90153 -44.87722
1stQ      81.41197  -8.64078
Median    89.45683   0.04482
Mean      89.48835   0.02749
3rdQ      97.86138   8.63319
Max      132.80410  37.65664

> MoviesCor5 <- sna::netlm(y = liked, x = disliked, mode="graph", nullhyp = 'qapspp', reps = 1000)
> summary(MoviesCor5)

OLS Network Model

Residuals:
      0%       25%       50%       75%      100%
-20.7923806  -5.6939698  -0.6500858   4.9323606  45.5148070

Coefficients:
            Estimate   Pr(<=b) Pr(>=b) Pr(>=|b|)
(Intercept) 12.6800858 1       0       0
x1           0.1043884 1       0       0

Residual standard error: 7.563 on 1245829 degrees of freedom
Multiple R-squared: 0.01511     Adjusted R-squared: 0.01511
F-statistic: 1.911e+04 on 1 and 1245829 degrees of freedom, p-value:     0


Test Diagnostics:

        Null Hypothesis: qapspp
        Replications: 1000
        Coefficient Distribution Summary:

       (Intercept)       x1
Min       665.3149 -35.3562
1stQ      690.6394  -8.3288
Median    696.8541   0.3777
Mean      697.3308   0.2584
3rdQ      704.0804   8.9887
Max       730.7926  42.1301
```

Figure 3

More than 60 reviews.

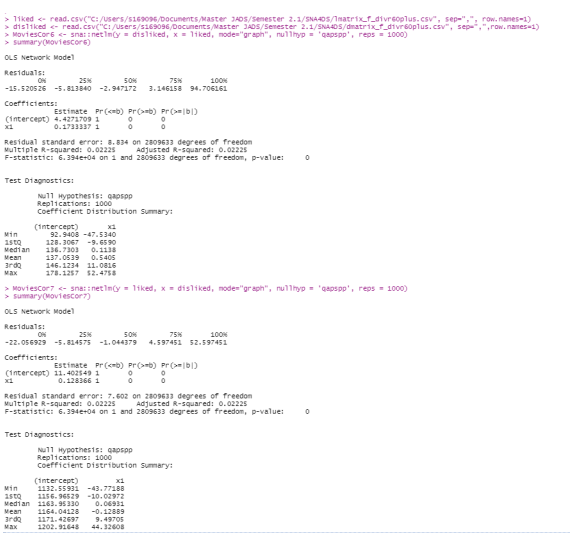
For the last time, we decreased the filter threshold for the number of ratings that users have given in the dataset. This also increased the amount of datapoints in the dataset, and we ran the analysis on users who had 80 reviews in total (like and dislike) or more. The results are shown in the figure below.

As can be seen from the figure, the QAP regression model explains 17% of the variation in the like/dislike interaction network for movies when looking at liked movie ratings. When looking at disliked movie ratings, the model can explain 13% of the variation. Th finally gave us the feeling that we are building somewhat meaningful models.

We even wanted to go a step further and decrease the filter threshold for the number of ratings that users have given to 40. But this would yield a dataset that was too large to analyse, as it costs 25+ hours to run an analysis on it.

Findings in relation to hypothesis.

When relating our findings to our aforementioned hypothesis, we see that it is possible to build models that somewhat can explain the variance in the network in a significant way. However, we can clearly see that more data is not always giving better explanations.



```
> liked <- read.csv("C:/Users/s169096/Documents/Master 2ADS/Semester 2.1/SNA4DS/lmatrix_f_divr60plus.csv", sep=",", row.names=1)
> disliked <- read.csv("C:/Users/s169096/Documents/Master 2ADS/Semester 2.1/SNA4DS/dmatrix_f_divr60plus.csv", sep=",",row.names=1)
> MoviesCor6 <- sna::netlm(y = disliked, x = liked, mode="graph", nullhyp = 'qapspp', reps = 1000)
> summary(MoviesCor6)

OLS Network Model

Residuals:
      0%       25%       50%       75%      100%
-15.520526  -5.813840  -2.947172   3.146158  94.706161

Coefficients:
            Estimate  Pr(<=b) Pr(>=b) Pr(>=|b|)
(Intercept) 4.4271709 1       0       0
x1          0.1733337 1       0       0

Residual standard error: 8.834 on 2809633 degrees of freedom
Multiple R-squared: 0.02225     Adjusted R-squared: 0.02225
F-statistic: 6.394e+04 on 1 and 2809633 degrees of freedom, p-value:     0


Test Diagnostics:

        Null Hypothesis: qapspp
        Replications: 1000
        Coefficient Distribution Summary:

       (Intercept)         x1
Min       92.9408 -47.5340
1stQ     128.3067  -9.6890
Median   136.7303   0.1138
Mean     137.0539   0.5405
3rdQ     146.1234  11.0816
Max      178.1267  52.4758

> MoviesCor7 <- sna::netlm(y = liked, x = disliked, mode="graph", nullhyp = 'qapspp', reps = 1000)
> summary(MoviesCor7)

OLS Network Model

Residuals:
      0%       25%       50%       75%      100%
-22.056929  -5.814573  -1.044379   4.597451  52.597451

Coefficients:
            Estimate  Pr(<=b) Pr(>=b) Pr(>=|b|)
(Intercept) 11.402149 1       0       0
x1           0.123366 1       0       0

Residual standard error: 7.602 on 2809633 degrees of freedom
Multiple R-squared: 0.02225     Adjusted R-squared: 0.02225
F-statistic: 6.394e+04 on 1 and 2809633 degrees of freedom, p-value:     0


Test Diagnostics:

        Null Hypothesis: qapspp
        Replications: 1000
        Coefficient Distribution Summary:

       (Intercept)         x1
Min     1132.55931 -43.77188
1stQ    1156.96629 -10.02972
Median  1163.95330   0.06931
Mean    1164.04128  -0.12889
3rdQ    1171.42697   9.49705
Max     1202.91648  44.32608
```

Figure 4

Also, it very interesting to see that liking movies says more about which movies you might dislike, than vice versa. An explanation for this might be that we saw in the data that there were more positive ratings than negative ratings, which could explain that disliking defines a more specific taste that is better detectable.

ERGM. (about 1000) – 2.5 POINTS

- Present your results appropriately (plots, tables. . .) and discuss your findings in plain English

- Discuss the meaning of your findings in relation to your hypothesis. (half of the points evaluated in this other part)

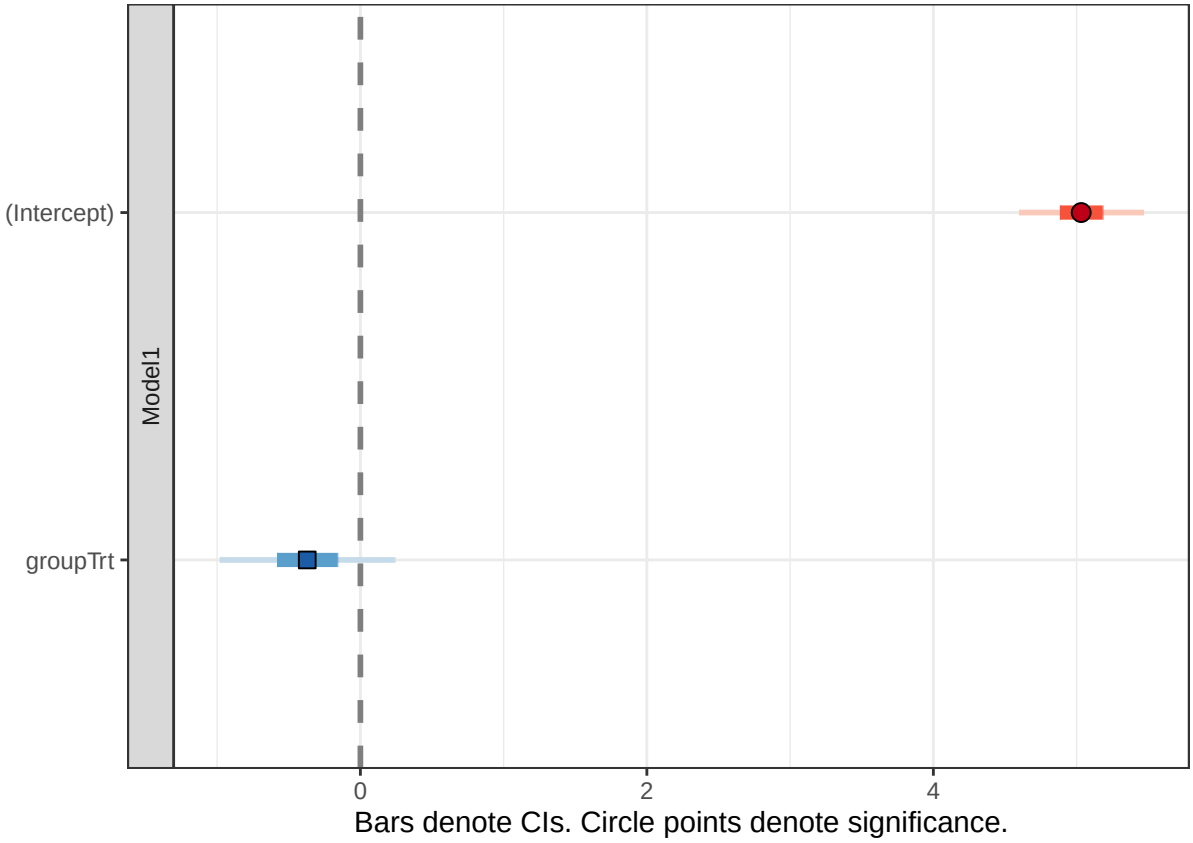
Option 1:

	Model 1
(Intercept)	5.03 ***
	(0.22)
groupTrt	-0.37
	(0.31)
R^2	0.07
Adj. R^2	0.02
Num. obs.	20

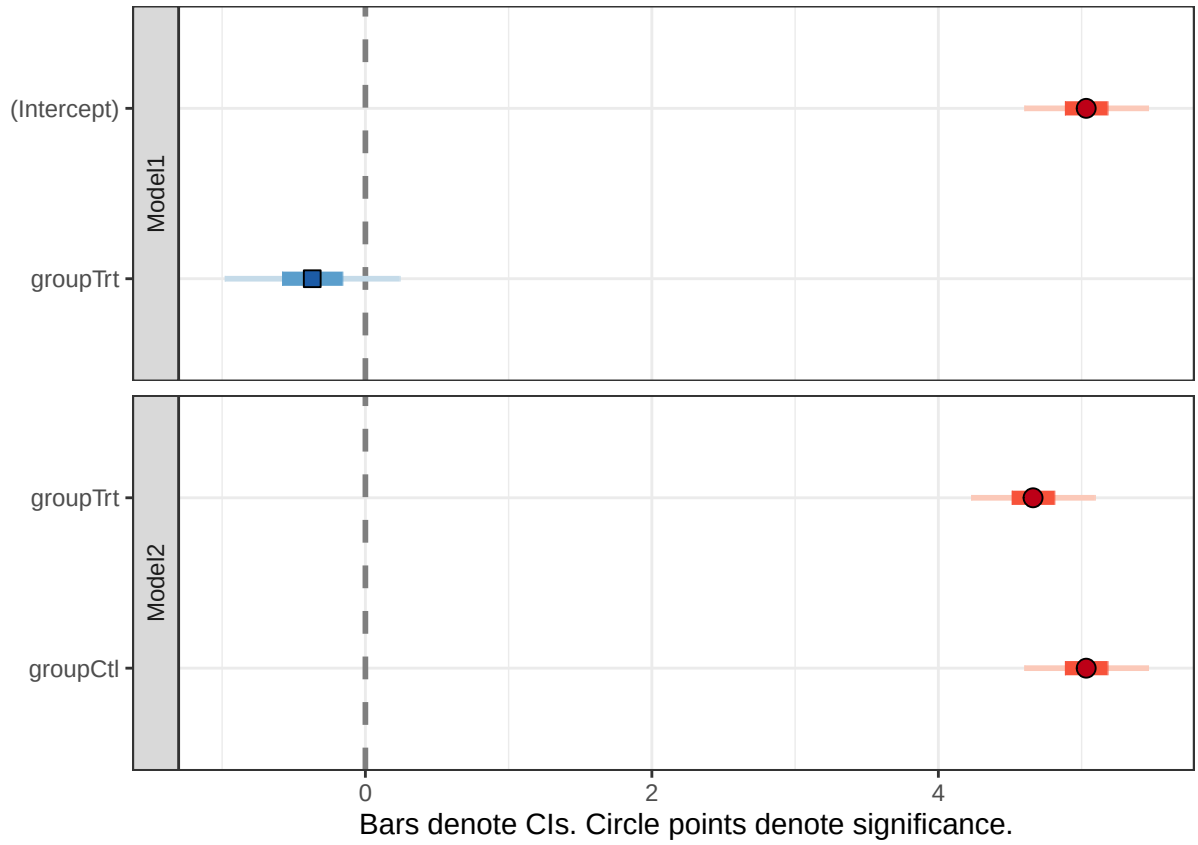
Option 2

	Model 1	Model 2
(Intercept)	5.03 ***	
	(0.22)	
groupTrt	-0.37	4.66 ***
	(0.31)	(0.22)
groupCtl		5.03 ***
		(0.22)
R^2	0.07	0.98
Adj. R^2	0.02	0.98
Num. obs.	20	20

Option 3


(Intercept)
groupTrt
Model1
0
2
4
Bars denote CIs. Circle points denote significance.

Option 4



Conclusion

(about 350 words) – 0.7 POINTS What were your topic and research questions again? (1 sentence)

What did you learn from the two analysis you run? *** most important point to address 0.5 POINTS here

Who benefits from your findings?

What does remain an open problem?

Discussion. What remains an open problem, is that we assume that a frequent reviewer gives a rating to a movie after watching it. However, we don't have information on (re-)viewers that watch a movie but don't give a rating afterwards. This is difficult to overcome as we simply don't have data available on this matter.

Can you give suggestions for future work in this area? For future research it could be beneficial to dive more into the literature and find out more about what is already known in order to give a stronger reasoning behind the explanatory variables used in the analysis. Also, it would be interesting to look into more recent data as streaming services are widely used nowadays and the variation of users is maybe larger now.

References

Amershi, S., Cakmak, M., Knox, W. B., & Kulesza, T. (2014). Power to the people: The role of humans in interactive machine learning. *Ai Magazine*, *35*(4), 105–120.

Aust, F., & Barth, M. (2020). *papaja: Create APA manuscripts with R Markdown*. Retrieved from https://github.com/crsh/papaja

Bell, R. M., & Koren, Y. (2007). Lessons from the netflix prize challenge. *Acm Sigkdd Explorations Newsletter*, *9*(2), 75–79.

Guillory, A., & Bilmes, J. A. (2011). Simultaneous learning and covering with adversarial noise. In *ICML*.

Narayanan, A., & Shmatikov, V. (2006). How to break anonymity of the netflix prize dataset. *arXiv Preprint Cs/0610105*.

R Core Team. (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Takács, G., Pilászy, I., Németh, B., & Tikk, D. (2008). Matrix factorization and neighbor based algorithms for the netflix prize problem. In *Proceedings of the 2008 ACM conference on recommender systems* (pp. 267–274).