**Project Report**

1. **Description of the data**

In this report we will be analyzing the MovieLens 1M movie reviews dataset provided by GroupLens Research. The data contains over 1,000,000 ratings from 6,000 users on 4,000 movies. The ratings range from 1 to 5 where a rating of 5 means they loved the film. Additional information on the users (specifying the user’s age, gender, and occupation) and movies (specifying the film’s year of release and the film’s genre(s)) is provided alongside the ratings data. The data is readily available from the GroupLens.org website, but it does not come in a format that is compatible with the social network analysis packages of **R**.

In order to structure the data into an acceptable form, we used python (specifically, the sqlite library) to load the ratings, users, and movie data into a database. From here we could extract any information we needed and output the data in any form we specified (ideally, in edge-list form). Our first few attempts at extracting data resulted in datasets that were much too large for **R**. Thus, we decided that we would focus on a particular subset of movie ratings: horror movies, specifically those that were rated between Jan 1, 2001 and Jan 1, 2003 and received a movie rating of 4 or higher. We also extracted the user information of the movie reviewers that were in this subset. In this manner, we created a more manageably sized dataset that was compatible with **R**.

Once we load the data into R, we transform our network into a bipartite graph, where users represent one mode and movies the other. That is to say, an edge exists between a user node and a movie node if the user has rated that movie. The final dataset, therefore, contains 589 user nodes, 223 movie nodes, 2545 edges and the following node attributes:

**Attributes of Horror Movies Node:**

* Movie title

**Attributes of User Rating Node:**

* User ID
* Gender (M, F)
* Age group (under 18, 18-24, 25-34, 35-44, 45-49, 50-55, 56+)
* Occupation (academic/educator, artist, clerical/admin, college/grad student, customer service, doctor/healthcare, executive/managerial, farmer, homemaker, k-12 student, lawyer, other/not specified, programmer, retired, sales/marketing, scientist, self-employed, technician/engineer, tradesman/craftsman, unemployed, writer)

For our analysis, our main focus is on the induced graphs of this network: the user-to-user projection and the movie-to-movie projection. To clarify, a connection exists between users if they have rated the same movies, and similarly, a connection exists between movies if they were rated by the same user. The characteristics of the user-user and movie-movie projections are as follows: 589 user nodes / 34045 edges and 223 movie nodes / 7046 edges, respectively.

1. **Analysis**

For the user-user network, we would like to know the following: Do users tend to form ties with users that are similar to themselves? Do users form natural communities? And if so, who are the important nodes in these communities? Can we model the user to user network to differentiate the effects between user gender and age group? Can we predict which users are likely to review the same movie next?

* 1. **Homophily**

We begin by answering the first question: do movie reviewers tend to have ties with other users that are similar to themselves (either by age, gender, or occupation)? To quantify this, we will compute the assortativity coefficients for these three categorical variables. The coefficients can take values between -1 and 1, where values close to 1 mean users are highly similar and values close to -1 mean users are very different from each other. In our user-user network the assortativity coefficients for age group, gender and occupation are 0.005232196, 0.004567059, and -0.001657487, respectively. This tells us that users tend to have ties with users that are similar in age and gender but not in occupation. This result is not too surprising, but they are relatively close to 0, which implies that these results might not be too different from a random graph with similar degree distribution whose edges were assigned by chance.

To test whether this is indeed the case, we will perform both a CUG test and a QAP test on the variables. For the CUG test, our null hypothesis is that the assortativity of users (by either age, gender, or occupation) is due to chance edge formation. For the QAP test, the null hypothesis is that assortativity of users is due to the random placement of nodes. The p-value results and the graphical representation of the results are as follows:

CUG test ----------------- QAP tests ------------------------------

Age: Pr(X>=Obs): 0 p(f(perm) >= f(d)): 0

Gender: Pr(X>=Obs): 0.008 p(f(perm) >= f(d)): 0.045

Occupation: Pr(X>=Obs): 0.494 p(f(perm) >= f(d)): 0.195

**Age Gender Occupation**

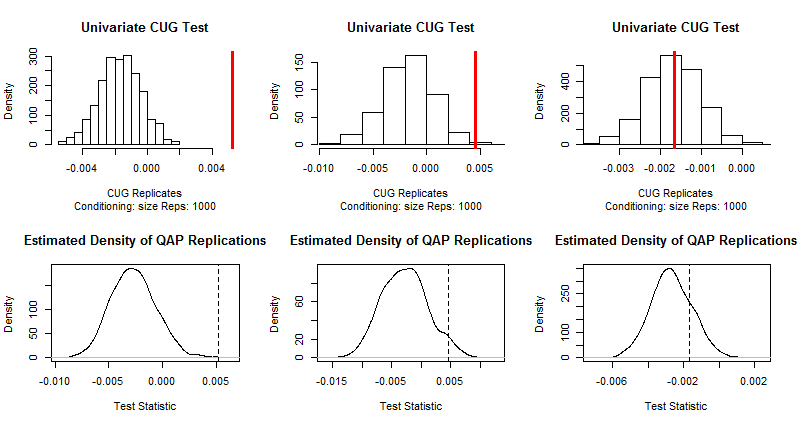


Figure 1

At the 5% level of significance, we can reject the null hypothesis of the CUG test and QAP test for age group and gender but not for occupation. Therefore, it appears that users have ties with other users similar in age and gender. This is good to measure, but ultimately it is not a surprising result…

* 1. **Community Detection**

First, we analyzed if there are different groups of movie reviewers that watched the same horror movie genre. To answer this, we used *leading eigenvector centrality* method to see if we can detect movie reviewers’ communities that watched and reviewed the same horror movie(s). Interestingly the algorithm detected five communities in different size; the following table summarizes communities detected:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Community** | 1 | 2 | 3 | 4 | 5 |
| **Sizes** | 8 | 12 | 14 | 14 | 7 |

We further analyzed to learn more about users’ characteristics by extracting the sub-communities.

**Community 1:** 7 out of 8 users are male in different age-groups ranging from under 18 through 45-49, and different occupation. The female age group was 25-34.

**Community 2:** 100% are male users, 5 users are in 18-24 age group, another 5 are in 25-34 age group, 1 user under 18, and 1 user 35-44 age group - this community is dominated by young male adults in different occupation areas.

**Community 3:** another 100% male dominated community – that is mostly young adults and middles-aged-adults.

**Community 4:** 77% Male users in young adults and middle-aged-adults age range in different occupation; and 33% female users in middle-aged-adults and old adult age range in different occupation. Still male dominated community.

**Community 5:** This community 71% is Female young adults’ users in different occupation areas, and is the only community that is dominated by female users when compared to other communities.

* 1. **ERGM**
  2. **Link Prediction**

1. **Conclusion**

The results …

Limitations of the analysis …

What we might have done with more data or more time … for one, in this report we focused primarily on the user-user projection; we would have liked to have been able to gathered more movie information (director, actors, horror movie subgenre like slasher, gore, supernatural, etc.) .