168 Project

**Preliminaries and Data:**

Netflix currently has 195 million paid subscribers for their services. Our study was done based on a dataset taken from 2005 of 2,649,429 users and 17,770 movies. The rating system in place at the time the data was collected was a one to five-star rating system. For our experiments, we randomly selected 10,000 users and 5000 movies to test which similarity and recommendation methods best predicted the ratings users gave to the films they viewed. In the following sections, we will go into greater detail of the specifics of how this selection of users and movies was divided to predict user ratings.

**Similarity:**

A usual concept in the study and analysis of networks is the idea of similarity. This study of similarity between nodes is used to gather information crucial to understanding the overall trends and relationships in a network. Most often, similarity can be especially used to develop predictor tools of peoples’ behaviors in the study of social networks. Our team decided to use similarity in hopes of predicting the Netflix titles that consumers would be more likely to watch and rate highly based on their previous watch history combined with their ratings of these.

We tested our data against three methods of measuring similarity. An adjusted cosine similarity, and calculation of the Pearson correlation coefficient and Jaccard coefficient were all studied in hopes of finding the most efficient tool in creating an item-based similarity recommendation system. All three methods (and the idea of measuring similarity in general) will be used on a bipartite network between a group of nodes representing Netflix subscribers and a group representing the movies they watch and rate.

**Cosine Similarity:**

There are generally two approaches to measuring similarity in networks, and these are structural and regular equivalence. For our purposes, since we are studying the ratings of the same movies that audience members are watching, we will be using the structural equivalence approach. Cosine similarity takes the most obvious approach to measuring structural equivalence, which is to count the number of shared neighbors between two nodes, and takes it one step further by taking into account the varying degrees of nodes in the network. This ideally rids the experiment of any misinformation due to the number of common neighbors not being scaled according to the overall size of the network.

Although cosine similarity often uses unweighted networks, our study will be using a weighted network to address the different ratings between 1-5 that audience members can give the films they watch. Using the entries of our adjacency matrix, we divide the number of common neighbors by the geometric mean of the degree of nodes, using the following equation:

(insert eqn. 7.35 pg 196)

Our results from calculating the cosine similarity of netflix users was...

**Pearson Correlation Coefficient:**

The Pearson Correlation Coefficient looks at rows of the adjacency matrix by studying the deviation from the average of each row of a given entry. It also looks at the dissimilarity of nodes by taking into account the neighbors of two nodes that are not shared.

(insert eqn. 7.39 pg 197)

For our study, the Pearson correlation coefficient is beneficial by revealing any discrepancies caused by users having some similar ratings for certain movies, but also have a large selection of movies that do not have a shared interest between the two participants.

Results..



****

**Recommendation Evaluation:**

Our process of evaluating these recommendation methods is based off those frequently used for similar tasks, and most notably those developed in “Evaluating collaborative filtering recommender systems” by Jonathan Herlocker et. al and “Evaluating recommendation systems” by Guy Shani and Asela Gunawardana. The evaluation processes outlined in these studies use a double and triple division evaluation system, where “double” and “triple” refer to the number of subsets data is divided into. In a double division evaluation, the data is divided into a training set and probe set, while triple division evaluation takes into account method parameters that optimize the data to result in bias in its performances when run under tests. Our study did not include any method parameters so we will be using the double division evaluation system.

The purpose behind creating a training and probe set is to be able to test estimated recommendations against our own data. Training sets are used as the input data for our evaluation methods described above, and these predictions are then compared to the probe sets to test their accuracy. The level of accuracy is quantified in three forms explained below - ranking, precision, and differentiability

**Ranking:**

The ranking, or ranking score, of a prediction is the most general way of quantifying the accuracy of an evaluation method. Based on the recommendation methods and algorithms computed on the data, we create a list of recommended films for a user in descending order. We then compare this list with the actual films left in the probe set, also placed in descending order based on their given ratings. Every user and movie pair is evaluated against the training set’s prediction, summed together, and then normalized based on the size of the probe set. This gives us the following formula, for our ranking score r:

$r=\frac{1}{\left|\mathrm{E}^{P}\right|} \sum\_{(i, \alpha) \in \mathrm{E}^{P}} r\_{i \alpha}$

**Precision and Differentiability:**

However, at times this general ranking system raises two issues. In reality, only a limited list of films would need to provided as recommendations, and recommendations only provide useful information if they are unique to a user. To address these issues, we used the measures of precision and differentiability to provide alternative metrics for the accuracy of our evaluations.

For precision, we can simply look at the top L results (our study chose L = 20) of a recommendation list, and see how many entries match the probe set of a given user. Quantitatively, we do this for each user i, and then average over all the nodes in the probe set to generate our entire network’s precision.

Differentiability is the only measure of accuracy that compares the evaluations of two users with each other. For this metric, we look at the number of shared entries in the top L results of the recommendation lists for any two users, convert it into a fraction of L, and then subtract from one. Once we find the average over all pairs of nodes in the network, this resulting differentiability tells us how much recommendation lists differ from one another.

**Novelty:**