EE232E - Project 2

IMDb Database Exploration

In this project, we will create and explore a network from the IMDb movie data. The data is available at IMDb Interface. We are going to study the properties of actor/actress network and movies. You can also also find a cleaned up version of the data at this Box link.

1. Download `actors.list.gz`, `actresses.list.gz` (or use the `actor_movies.txt` and `actress_movies.txt` files from the cleaned up data), merge those 2 lists into one file, and remove all actors/actresses with less than 5 (so actors who have acted in four or fewer number of movies) movies; Note that you will have to parse the data in these lists as accurately as possible to extract the entities consistently and create the network. So plan on spending some time in cleaning the data set.

2. Construct a weighted directed graph $G(V, E)$ from the list, while

$$V = \{\text{all actors/actressess in list}\}$$

$$S_i = \{m | i \in V, \text{m is a movie in which i has acted}\}$$

$$E = \{(i, j) | i, j \in V, S_i \cap S_j \neq \phi\}$$

   and for each *directed* Edge $i \to j$, a weight is assigned as $\frac{|S_i \cap S_j|}{|S_i|}$.

3. Run pagerank algorithm on the actor/actress network, look into those who are among top 10, do you know their names? List the top 10 famous movie celebrities in your opinion, what are their pagerank scores? Do you see any significant pairings among actors? Any major surprises, in the sense that well-known actors do not show up in the high pagerank list?

4. Similarly, remove all movies with less than 5 actors/actresses on list, construct a movie network according to the set of actors/actresses, with weight assigned as the jaccard index of the actor sets of 2 movies. Now we have an undirected network instead.

5. Do a community finding on the movie network; use the Fast Greedy Newman algorithm. Tag each community with the genres that appear in 20% or more of the movies in the community. Are these tags meaningful?

6. Add the following nodes into the network, For each of them, return the top 5 nearest neighbors. Which communities does each of them belong to?

   Batman v Superman: Dawn of Justice (2016)

   Mission: Impossible - Rogue Nation (2015)

   Minions (2015)

7. Download the ratings list, derive a function to predict the ratings of the above 3 movies using the movie network. (hint: try to use the ratings of neighbor movies and movies in the same community.)

8. Using a set of features that include the following:

   - top 5 pageranks of the actors (five floating point values) in each movie.
   - if the director is one of the top 100 directors or not (101 boolean values). These are directors of the top 100 movies from the "IMDb top 250". You can also find a list of these movies in the `ratings.list.gz` file.

   train a regression model and predict the ratings of the 3 movies mentioned above. Specify the exact feature set you use and how you compute the numerical values for these features. Compute and state the goodness of fit for your regression model.

9. (**Bonus**) Try predicting the ratings of these movies with a different approach. Construct a bipartite graph, with actors/actresses representing the vertices of one part and movies representing the

vertices of the other part. An actor/actress is connected to all the movies he has played in. Assign a score to each actor based on the ratings of the movies he has played in (it's on you to define an appropriate metric here; natural choices are the average among all ratings, the highest, or average among the top ones). Then predict the ratings of the new movies based on the scores you have assigned to the actors.