DATA620 Assignment 4

Justin Hink

Dataset Chosen

The dataset that I have chosen for this assignment (and thus Project #1) is the MovieLens database which is freely distributed by the GroupLens research group. The dataset contains anonymous ratings of almost 4000 movies from over 6000 users. In the dataset comes in different sizes, with the largest coming in at over 20 million different users. Due to resource constraints, I will only load and analyze a smaller subset of the data (which are also downloadable).

http://grouplens.org/datasets/movielens/

ETL Plan

The data provided on the MovieLens site is not in a pre-made graph format. As such we will need to transform the data. The following steps will be performed.

1) Download raw data files
2) Develop code that transforms the data into a well-known, standardized graph format. At this point, graphML is a leading candidate (based on a recommendation from the iGraph documentation. More on this later). This code will almost certainly be Python (2.7).
3) Once the graphML format has been created, load it into an analysis environment. This will simply be a separate Python program (a notebook will be created as per the project spec).

Analysis Plan

With the data transformed appropriately into a set of Nodes (Movies, Users) and edges we will then be able to perform some rather basic analysis. The following items will be investigated:

1) Calculate centrality metrics for all nodes in the graph
2) Compare centrality metrics for across a number of categorical attributes. The first of which will be movie genres. Others (such as User gender) will also be examined time permitting.