Bipartite Social Network Analysis of Movie Relationship based on Reviewers

Catalin Capota, Scott Page, Tim Sachs

DePaul University CSC-495, Professor Robin Burke

Abstract

This paper describes a bipartite social network analysis of movies relationships based on reviewers derived from a dataset of movie reviews from amazon that contains data from January of 1998 until October 2012. This paper will describe the data, the stages of our analysis and conclude with the findings. We will analyze the network projection to identify key relationships and better understand the relationships between movies developed by avid reviewers.

I. Description of the Data:

The dataset we obtained was provided on the Stanford Network Analysis Project (https://snap.stanford.edu/data/web-Movies.html). The data spans a period of more than 10 years, including all ~8 million reviews from January 1998 until October 2012.

Figure 1: File record layout of the amazon movie review dataset, each field being a new line and records delimited by newline.

```
product/productId (text): asin, e.g. amazon.com/dp/B00006HAXW
review/userId (text):  id of the user, e.g. A1RSDE90N6RSZF
review/profileName (text):  name of the user
review/helpfulness (fraction integer/integer): fraction of users who found the review
helpful
review/score (decimal 0.0-5.0): rating of the product
review/time (Epoch): time of the review (unix time)
review/summary (Text): review summary
review/text (Text): text of the review
```

A program was created to parse this large dataset and extract a subset of fields, this was needed to reduce the time it took to load the dataset into R and maintain a smaller memory footprint.

Figure 2: First Three lines within the generated CSV file extracted from the original dataset.

```
product/productId,review/userId,review/helpfulness,review/score,review/time
B003AI2VGA,A141HP4LYPWMSR,7/7,3.0,1182729600
B003AI2VGA,A328S9RN3U5M68,4/4,3.0,1181952000
```

The Amazon Product ID, further referred as ASIN (*Amazon Standard Identification Number*) was used to establish the relationship between movies along with the User ID to build the Movie-Movie projection. For usability the ASIN number can be used to retrieve the actual movie name, we developed a program to query amazon using the following URL and parse the name.

Figure 3: URL scheme used to query Amazon and sample using the ASIN in figure 1.

http://www.amazon.com/s/ref=nb_sb_noss?url=search-alias%3Daps&field-keywords=**B00006HAXW**

The returned HTML was then parsed to extract the movie name, a CSV file was created that contained the mapping of ASIN number to movie name which was loaded into R and attached as an attribute to the network graph.

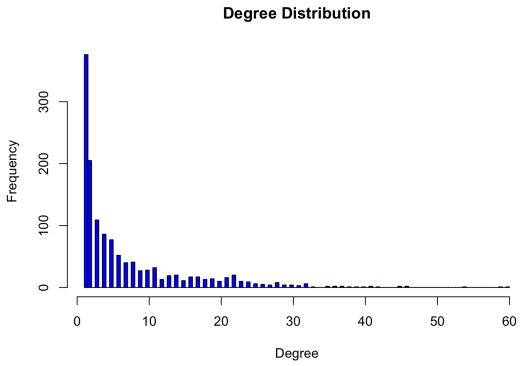
Figure 4: First Four Records Generated CSV file from Amazon.com programatic query and result parsing.

```
ID,ASIN,ProductName
"2371","B0039OEV86","Rush-Hour-Jackie-Chan"
"2748","0783223145","Deer-Hunter-Widescreen-VHS"
"1995","B00003BE03","Godzilla-VHS-Matthew-Broderick"
```

Throughout the processes of understanding the data, visualization and analysis there were significant realizations that modified the size of the dataset, specifically the nodes, edges and structure of the graph. This process will be explained in the analysis section and specific metrics around node and edge counts will be presented then. The basic building blocks and structure of our data was defined above, further changes and enhancements will be described in analysis section, as a summary below are the statistics for our final graphs.

Figure 5: Statistics realized in final graphed dataset of 1998-1999.

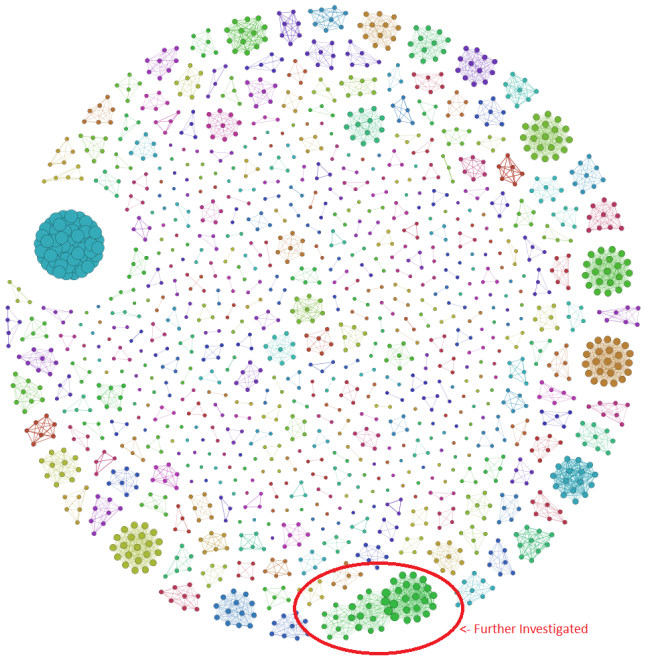
```
Year Range = August 1997 - January 1st 2000.
Nodes = 1322
Edges = 4473
Weakly Connected Components = 49
Average Path Length = 2.396
Average Clustering Coefficient = 0.582
Total Triangles = 12247
Average Weighted Degree = 20.04
Modularity 0.727
```



II. Analysis:

This section of the document will explain the process we used to analyse the dataset in our attempt to better understand how movies are related through reviewers. We specifically tried to find relationships between the movies based on reviewers geared around the movie genre (Oldies, Hit Movies, Scary Movies, etc..) or based on age. Particular interest was centered around which movie clusters develop and what are the main movies that bridge these clusters, acting like transitional movies to either new genres or between drastically different categories.

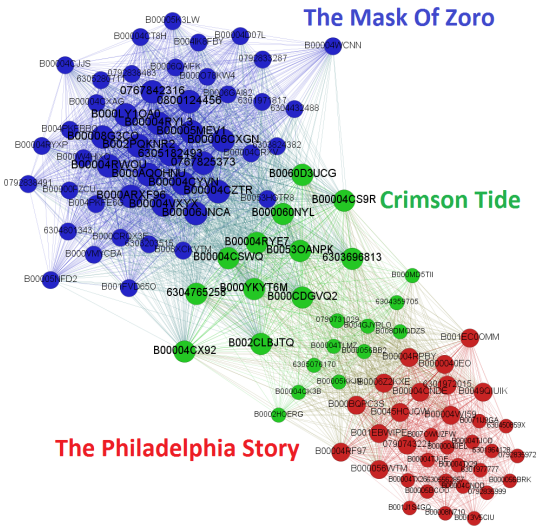
Figure 6: Original Network Visualization for 1998 movie-movie projection with highlighted cluster, used in follow up efforts. Nodes are colored by modularity and sized by connectivity



The original dataset was quite large and we were not able to efficiently work with it within either R or Gephi, we decided to focus on the relationships developed in the initial years and as our understanding grew our intent was to expand this dataset into a much larger time range or compare multiple years. The initial

visualizations focused purely on the 1998 data, producing the following visualization (Figure 6). The movies appeared to be heavily clustered in singletons or small groups. Within a closer look on the bottom portion of the graph an interesting network developed that we further investigated.

Figure 7: Filtered data to only cluster identified in Figure 6, colored by modularity indicating three distinct groupings



Isolating the cluster of interest and running modularity in Gephi revealed three group colorations, however we noticed that although each group had different ASIN's for the highly connected clusters there truly were only three movies contained within the network ("The Mask of Zoro","Crimson Tide" and "The Philadelphia Story"). Looking at the amazon product pages for these movies we realized amazon issues different ASIN's for each type of digital format of the same movie and replicates the reviews within these formats.

With these highly connected components due to the replication of reviews it greatly increased the network sizes in processing, built false clusters and hid the real relationships between different movies. We realized and verified that the dense clusters in our original visualization (Figure 6) were identical movies in multiple formats (DVD, VHS, Digital, Remastering, Alternate Languages, etc..). As a useful side effect of this analysis we are able to identify most popular or produced movies through only the clustering patterns without knowing any details of title or contents.

The next step in our analysis was to create a process for eliminating these differing formats and only retaining one version of the movie to create proper network relationships. We created an R script that eliminates movies from our file based on nodes that share exactly the same neighbors. Essentially only retaining one version of the reviews for a specific movie. Once we removed the edges we were able to expand the graph to the final format illustrated in ***Figure 8***, further analysis was performed with derivative visualizations or alternate views of this graph. It is now possible to see true clusters develop.

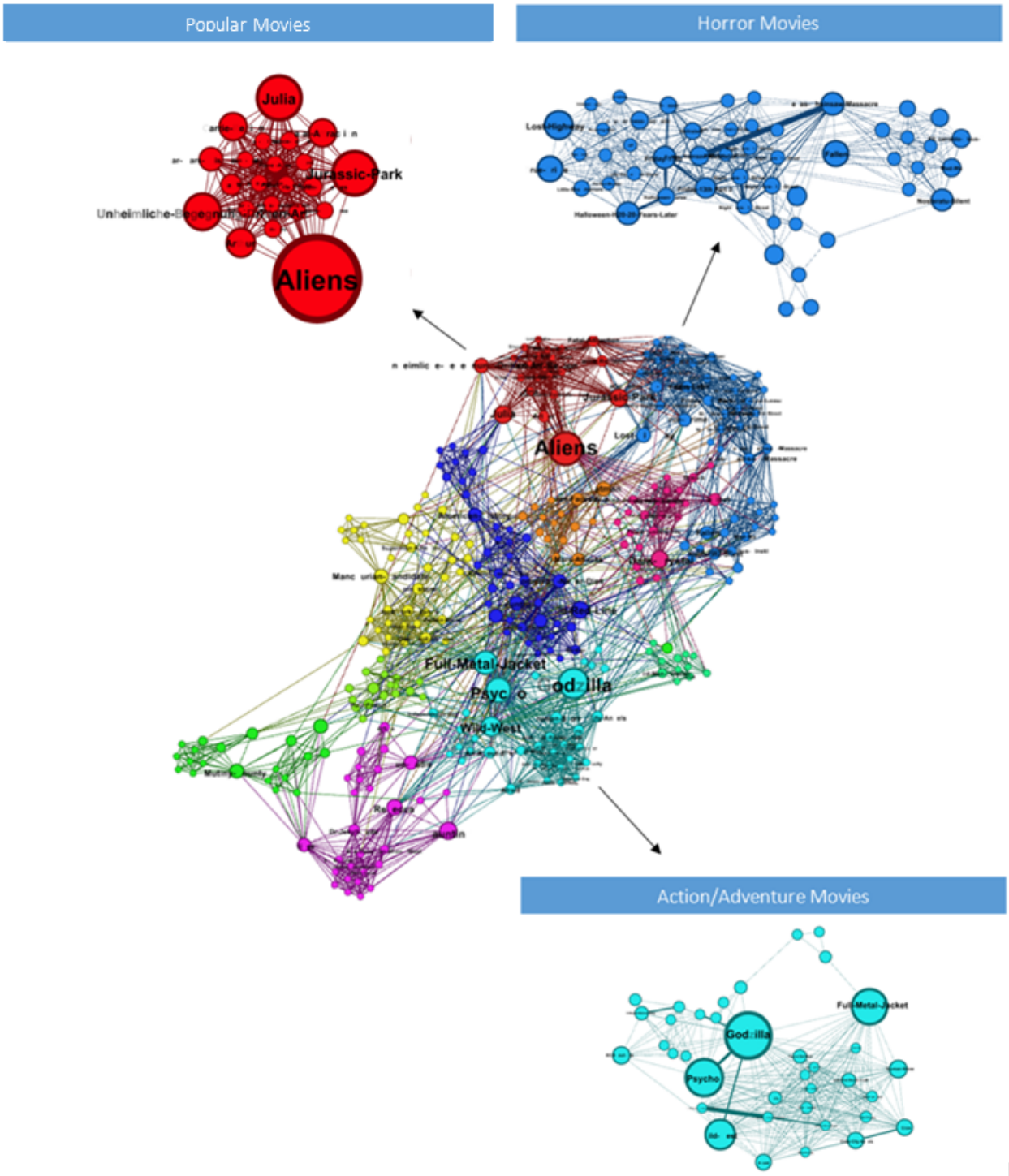


Figure 8: Visualization of the final network dataset, after data was deduplicated, isolate removed and time range expanded to January 1st 2998 - January 1st 2000. Network nodes are sized by the Betweenness and colored by the Modularity.

After digging into the 1998 - 2000 data, we extracted a network from reviews written during 2000. As you can see in figure 9, there were several more reviews written during 2000 than during the previous 2 years, forming a very dense network.

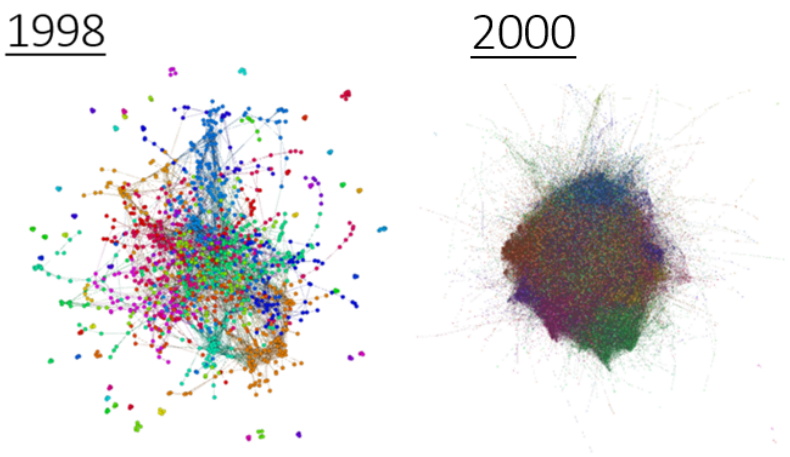
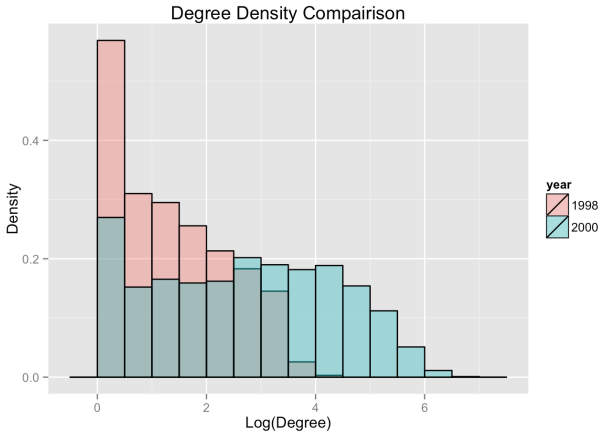
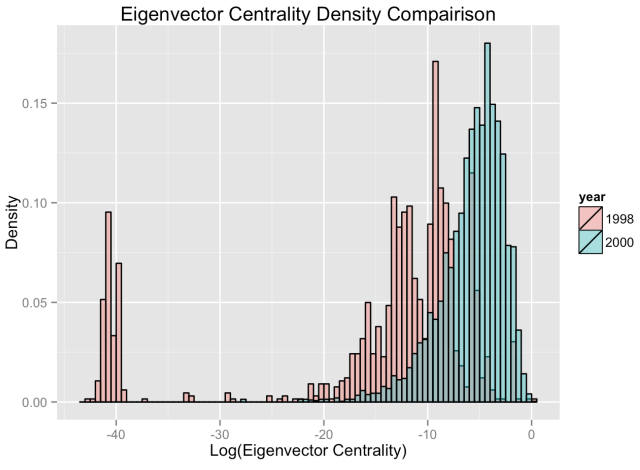


Figure 9 - Comparison plot of 1998 to 2000 vs the year 2000



Due to the significant density of the 2000 data, interpretation of the network through high level stats is necessary. As you can see in the Degree Density Comparison plot the degree of the Year 2000 network is in general significantly higher than the year 1998 - 2000 network. This shift to a more tightly connected movie review network can be explained by the larger number of users on amazon.

Another way of visualizing this shift is with the Eigenvector centrality. In the Eigenvector Centrality Density Comparison plot we can see that the network shifted away from singletons toward higher degree nodes in general.



III. Conclusion:

Through our analysis, we have discovered that the movie - movie projection of Amazon movie reviews yields a network that divides by modularity into movie subnets that share traits in common. The group of movies identified in red from figure 8 form a group including:

- *Alien*
- *Star Wars*
- *Jurassic Park*

These movies are all popular movies, and have high graph strength on deeper inspection. *Alien* was identified in the network with the highest betweenness, this means that it provides a cross over point between between other groups of movies. This makes sense due to its high popularity and potentially means that it produces strong opinions among viewers resulting in numerous reviews from various user groups.

The blue cluster appears to be related to a genre: Horror. Movies such as *Nightmare on Elm Street, Texas Chainsaw Massacre,* and *Friday the 13th* helped us come to this conclusion. As you can see, the Red & Blue groups are located close to one an other on the plot, this means that they likely share some underlying features that attract users from the opposite grouping.

Finally, we looked into the teal grouping of action / adventure movies:

- Godzilla
- Psyco
- Wild West
- Full Metal Jacket

The central location of this cluster within the network represents the wide variety of action adventure movies, that attracts reviewers from numerous other well defined groups.

Limitations

Dealing with the scale of the dataset prevented us from labeling all of the movie nodes or exploring larger time frames. Specifically:

- The primary source of data was larger than would comfortably fit in memory
- The data contained a large number of duplicates and other data errors
- Amazon rate limited our attempts to grab movie names

Dealing with the scale of the data, forced us to break the project into pieces, working in Java for the initial data manipulation and filtering, then in R for the analysis and final transformations, and then finally in Gephi for advanced visualizations.

Recommendations

Our recommendations would be to incorporate the following attributes:

- Review Text
- Movie genres
- Main actors
- Movie publication date
- Movie Budget
- Review score

Using these attributes more thorough analysis could be performed including discovering the associativity of the attributes with the modularity groupings.