Discovering Productive Communities in the IMDB Dataset

Rob Churchill
Boston University

Evimaria Terzi
Boston University

Introduction

The film industry is volatile and unpredictable. Many movies barely break even in the box office, and producers often fail to recuperate their investments. With that in mind, is there any way to discover communities of actors, directors, and writers who produce the most profitable and highest rated movies in the business? And further, do ratings imply revenue?

Goals

We had three main goals in our analysis of the IMDB dataset:

1) Find whether a correlation exists between movie ratings and profits.
2) Identify communities of actors, writers, and directors who work closely with one another on a regular basis.
3) Discover whether certain communities produce higher average profits or ratings than others, and whether a community's ratings imply its profits or vice versa.

The Data

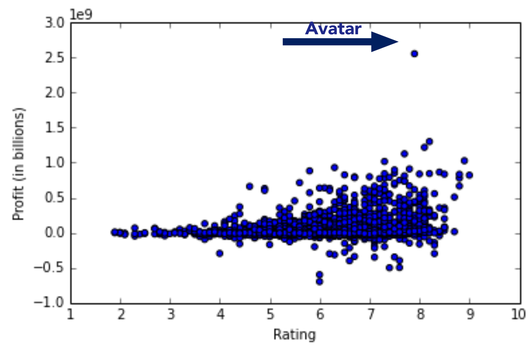
The IMDB dataset is available in pieces all over the internet, but the pieces are often missing a key attribute, like revenues, budget, cast list, or rating. The most pragmatic way to attain the dataset is to scrape.

- Top 500 movies by rating for every year from 2000 to 2014
- Threw out movies whose financial data was not available
- Went from 7500 to 3000 movies
- Most of removed movies were of foreign origin
- Resulting graph of actors more tight knit, nearly every remaining movie was produced in America or Great Britain

1) Rating vs. Profit Correlation

From our dataset of movies, we were able to extract the list of ratings and the list of approximate profits, and find their correlation coefficient. The correlation coefficient was only 0.22, meaning that there is a small, but insignificant relationship between the rating and profit of any given movie.

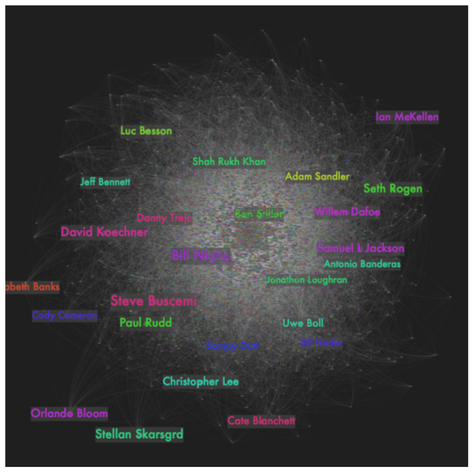
Rating vs. Profit



2) Community Discovery

- Create a graph of relationships between actors
- Edge weight is equal to number of collaborations between two actors
- Spectral Analysis
  - Compute eigenvector of each actor
- Use K-Means with the eigenvectors as data points
- K = 500, smaller leads to one giant community and many, very small communities.
- Note: still possible to have very small communities, so we ignore communities of size less than five

The graph of actors, with the names of highest degree in the graph. Color-coded by community ID



3) Community Analysis

Community statistics were analyzed in the following ways for both ratings and profits:

- **Mean** (dataset benchmarks — rating: 6.31, profit: 53,131,506)
- **Standard Deviation** (dataset benchmarks — rating: 1.05, profit: 149,334,185)
- **Median** (dataset benchmarks — rating: 6.4, profit: 2,276,953)

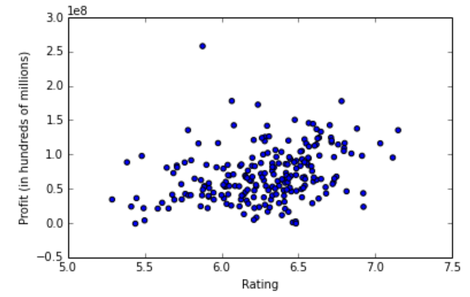
A **productive community** would have:

- High mean and median profit and rating, and a low standard deviation in relation to the entire dataset.

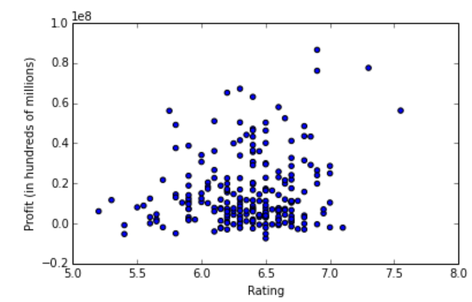
Observations:

- Ratings and profits have no significant correlation in communities.
- The graphs on the right show that many communities have means lifted by a few great movies. When observing the median, the truly high caliber communities show themselves.
- 64 communities were **productive** in respect to rating, 43 were **productive** in respect to profits, and only 13 communities beat the dataset benchmarks on both rating and profit. Even in these communities, there was no significant correlation between rating and profit.

Communities: Average Rating vs. Average Profit



Communities: Median Rating vs. Profit



Conclusion

Although there is no correlation between rating and profit on any level as we have defined, we were able to identify communities in IMDB that were more adept at producing high caliber films, whether it be rating-wise, profit-wise, or both. We can expect that when these communities collaborate in the future, they will produce movies worth seeing, whether it be because of the great ratings, or just because all of our friends are also willing to pay to see it. As for the rest of the communities, the film industry is volatile and unpredictable. Many movies barely break even in the box office, and producers often fail to recuperate their investments.

Future Work

- Use other means of discovering communities
- Incorporate genre into community creation

References

[1] IMDB.com

Acknowledgements

Special thanks to Sanaz Bahargam for giving me the idea in the first place.