MICA Graph Analysis Report 1

Victor R Martinez

October 18, 2016

1 Summary of results

- Gender assignment heuristics provide a ratio of women-men in scripts similar to those found in previous works.
- Following previous literature, male characters occupy more prominent roles than their female counterparts ($t = 2.79$, $p < 0.01$). When inspecting by genre, this result holds for Comedy, Crime, Drama, Adventure, Sport and Fantasy.
- Modularity does not contain enough information to predict a genre.

2 Data and Methodology

Data was 615 movie scripts. Every script was transformed into a graph where nodes represent characters and edges aggregate speaker transitions: if speaker B had a dialogue after speaker A, an edge between A and B was added. Self loops are not considered. Hence, the resulting graphs were directed graphs. If a character had fewer than τ[1] utterances, its node and connections were dropped. Characters could be either male (51%), female (21%) or unknown (28%). Female-male ratios were found in close agreement with the ones in previous work [2]. Each movie was assigned to one or more genre.

2.1 Measures

Following is a small description of the measures used for this analysis.

Degree The number of incoming (in-degree) and outgoing (out-degree) connections of a node.

Betweenness is an indication for the centrality of a node. It is measured as the number of shortest paths that go through a particular node. The higher betweenness, the more important a node is in the communications of a network.

Modularity measures the tendency of a graph to form groups or clusters. It is calculated as the fraction of edges that fall within a given group minus the

[1] For this report, $\tau = 2$

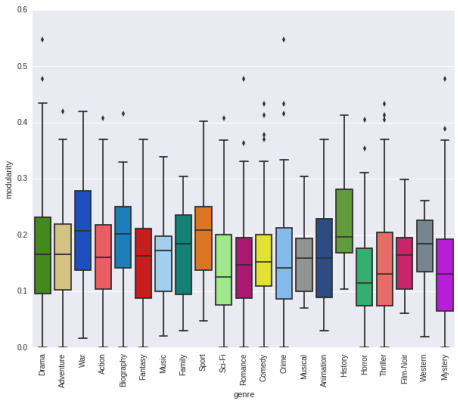

Figure 1: Modularity per genre

expected fraction if edges were distributed at random. It's values lie between $\frac{-1}{2}$ and 1. Networks with high modularity have dense connections between the nodes within modules but sparse connections between nodes in different modules. We used the Louvain Method [1], a greedy optimization method for finding communities that maximize the modularity.

3 Results

3.1 Degree centrality

Characters in movies had from 0 to 29 connections. Most of the characters had 5 connections or less. The difference between male and female degree was not significant at the 95% level ($t = 1.689$, $p = 0.09$). Difference across genres was also explores without any significant difference.

3.2 Betweeness centrality

In most of the scripts, males occupy a more central role than females ($t = 2.79$, $p < 0.01$). This effect was most prominent in Comedy ($t = 5.467$, $p < 0.01$), Crime ($t = 4.379$, $p < 0.01$), Drama ($t = 4.176$, $p < 0.01$), Adventure ($t = 3.596$, $p < 0.01$), Sport ($t = 4.215$, $p < 0.01$), and Fantasy ($t = 2.818$, $p < 0.01$) genres[2].

3.3 Communities and modularity

The number of communities was between 1 and 6, with median 3. On average, the genres with most communities were History, War and Family. On the other hand, Horror, Mystery and Sci-Fy were the ones with fewer communities. All

[2]Controlling for false rate discovery using Benjamin-Hochberg's method

modularity values were positive, ranging between 0 and 0.54 with median 0.154. Figure 1 shows the distribution of modularity across genres as a boxplot. The hypothesis of predicting genre as a function of modularity was explored using one-vs-all logistic regression models. The baseline was set to be the most popular class per genre, which turned out to be quite a high bar for this experiment (95.23% on average). The lowest baseline was found to be Drama with 78.25%. Unfortunately, the logreg model was unable to improve on this score. Further work might want to look for additional variables from the graph.

3.4 Future directions

Create multi-layer directed weighted graphs were edges amount for psycholinguistic norms. Then, for a certain norm (layer), centrality measures will reveal how is that metric traveling across the network.

References

[1] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre. *Fast unfolding of communities in large networks*. J. Stat. Mech. (2008)

[2] Smith, Stacy L., et al. *Gender bias without borders. An investigation of female characters in popular films across 11 countries*. (2014). Retrieved from `http://seejane.org/wp-content/uploads/gender-bias-without-borders-executive-summary.pdf` [October 18, 2016]

[3] Smith, Stacy L., et al. *Gender Inequality in 500 Popular Films: Examining On-Screen Portrayals and Behind-the-Scenes Employment Patterns in Motion Pictures Released between 2007-2012*. Retrieved from `http://annenberg.usc.edu/sites/default/files/2015/04/28/GenderInequalityin500PopularFilms.pdf` [October 18, 2016]