Pick the Right Team and Make a Blockbuster: a Social Analysis through Movie History

Wladston Viana, Ana Paula Couto da Silva, Mirella M. Moro
Universidade Federal de Minas Gerais, Brazil
{wladston,ana.coutosilva,mirella}@dcc.ufmg.br

ABSTRACT

We explore the relation between team performance and its topological network characteristics by creating a social network of movie producers and extracting topological features from each team. Then, we employ machine learning techniques to build a *movie success predictor* based on the extracted features. Results show topological features have predictive power over movie's success and can also be included in other non-topological models to improve accuracy.

CCS Concepts

•Applied computing → **Sociology; •Information systems** → *Data mining;*

Keywords

Movie network, Topological features, Success prediction

1. INTRODUCTION

Team collaboration is present in our society within clusters of movie and show producers, scientists, corporate teams, book editors and even robots [8, 10]. In this context, a crucial problem is how to form teams in order to maximize their performance. For example, a manager from a company may want to rearrange a team to ramp up productivity. This range of interactions among persons that engage in *teamwork* can be modeled through complex networks. Indeed, research using social network analysis reveals relationships between topological features and team performance [2, 14, 15].

Take the film industry, teamwork is practically ubiquitous. Teams are responsible for projects with large budgets that are expected to generate billions in revenue [9], making team performance crucial in both cultural and economic terms. Furthermore, any analysis needs a good dataset. Among all, the film industry has one of the best datasets available: the Internet Movie Database at http://www.imdb.com.

SAC 2016, April 04-08, 2016, Pisa, Italy

DOI: http://dx.doi.org/10.1145/2851613.2851812

Nonetheless, accessing the degree in which isolated features contribute to movie success is extremely difficult. First, movie success can be viewed in different ways: the gross generated, its profits (gross minus budget), its critic and public acclaim and the movie's popularity (the number of people who watched the movie). Success in one dimension does not guarantee success in the others. Finally, many complex factors are involved in movie's success, including the quality of special effects, the effectiveness and range of the movie's marketing campaign, whether it was released on a holiday, the popularity of its featured actors, among others.

Another perspective is analyzing combined features through *social network analysis*. IMDb rich data allows to recover the social network graph formed by movie producing agents; i.e., each movie has a production team corresponding to a connected set of nodes in the graph. The next step is to analyze the whole or parts of the graph. Topological parameters correlated to team's performance include structural holes [2] and the small world coefficient [15]. However, both studies do not consider the *aggregate* effect of many individual topological characteristics of agents on teams, nor the *predictive* capacity of such features regarding team performance.

Indeed, a simple experiment may verify that topological features provide predictive power for movie success. Figure 1 shows results of two sample predictors using the same test/train set. The baseline predictor uses a single training feature (the previous success of movie's team for the parameter being predicted). The test predictor uses the same feature with a topological one: the total number of people who have previously collaborated with the movie's team members. Figure 1 shows a significant gain in R^2 measures[1]. Therefore, even such a simple experiment strongly points towards the relevance of further investigating the effect of topological characteristics in movie team's success.

Here, our hypothesis is: features obtained by aggregating topological metrics of team individuals can predict team's success. Our contributions are summarized as: (i) we introduce a network model and selected metrics to assess movies producing team performance in Section 3; (ii) we define three movie success parameters and use them to group all movies in three performance categories, which are extensively characterized in Section 4; (iii) we propose a method for predicting movie success, which considers only features available *before* the movie release in Section 5; and (iv) we

[1]The R^2 is a statistical measure of how well the regression line approximates the real data points.

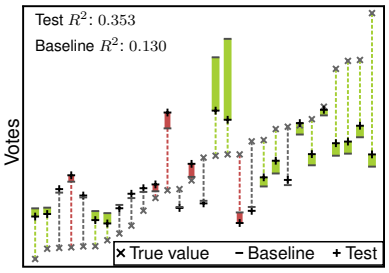

Figure 1: An example for predictive gain by using a single extra topological feature.

conduct a through evaluation analysis on three regression models using five fold cross validation on balanced train/test folds, and perform feature selection and strength analysis in Section 6. Overall, our results show topological features[2] extracted from teams provide predictive power.

2. RELATED WORK

Social studies try to understand how people work together to better achieve their goals [6, 9, 14, 15]. For instance, Stokols et al. [14] explore the "Science of team science" and the processes by which scientific teams organize and conduct their work. Team formation is also an optimization problem by grouping elements in order to most harmonically distribute the different abilities across multifunctional teams [1]. Other studies explore the presence of elements with special characteristics in the team to understand its success; for instance, the impact of extremely famous actors in movie teams [6]. Our work complements such prior studies by focusing specifically in assessing the effect of the social arrangement of the agents in their collaboration networks.

Likewise, social networks studies capture the structure of collaboration graphs [2, 3], topological features from networks to explain team performance [10], and social features to improve predictor's hits [7]. Differently, our work does not focus on interactions limited to a single collaborative work, but on the *whole history* of past collaborations of agents, both in current and previous works.

Similar to us, Uzzi and Spiro [15] evaluated the network of Broadway musical producers and correlated success to its small world coefficient. However, it does *not* explore multivariate analysis of multiple network metrics. Regarding movie success, Ghiassi et al [9] presented a classifier that can predict movie's gross in terms of nine distinct classes. However, they consider a dataset of only 364 movies (1999-2010) and data purchased that is hard to obtain for many movies (e.g., pre-release marketing expenses). We also aim to forecast movie success parameters, but besides financial success, we also want to predict popularity and ratings from movies, while considering social features from movie producers and a larger set of movies in our analysis.

Finally, the IMDb dataset is used in social network research and even in sociology [11, 13]. However, most consider a very small amount of movies in their analyses (e.g., 200 in [11] and 70 in [13]). Moreover, they mostly focus on classification, without deeply exploring social features. Unlike them, we make an extensive analysis of social features from teams by exploring a large set of 12,250 movies.

[2]We use *topological* and *social* features interchangeably.

In social network analysis, having large datasets is critical to yield more robust results. Our dataset considers *all* feature-length cinema movies in the IMDb dataset that received at least 1,000 votes. It contains several movie genres, spanning several decades, and coming from a wide range of countries. This way, our study considers more comprehensive and heterogeneous information regarding teamwork. To the best of our knowledge, we are the first to study the relation between network topology aspects and success considering such a large network of motion pictures producers.

3. NETWORK MODEL AND METRICS

Our goal is to improve movie success prediction by analyzing social features of teams in charge of producing movies. Specifically, we build a graph for movie producers only, without cast and crew (actors, directors, writers, etc.). The reason is twofold: producers are the *core* of the team (making the most important decisions and hiring the others), and adding much more agents would make the graph orders of magnitude larger, possibly exceeding the computing power available and adding uncertainty to the results [4].

Formally, given a set of movies $\mathbb{M}$, each movie $m \in \mathbb{M}$ has a set of producers $\mathbb{P}_m$ (aka m's production team). All producers $p \in \mathbb{P}_m$ are connected nodes in a graph $\mathbb{G}_\mathbb{P}$, which aggregates all producers of all movies in $\mathbb{M}$. Edges in $\mathbb{G}_\mathbb{P}$ are undirected and link producers who have worked together in one or more movie. Each edge has a weight value for the number of previous collaborations between its producers. For each node p, H_p is p's past experience (the number of p's previous productions), and S_p is p's previous success (the mean of success parameters for p's movies as defined in Section 4.2). Also, $\mathbb{G}_\mathbb{P}$ evolves over time to incorporate new data as new movies are produced. Hence, $\mathbb{G}_\mathbb{P}(t)$ is the producers graph considering only movies that were released up to instant t. Finally, for any given movie m, its characteristics are based on the topology from the subset of nodes in $\mathbb{P}_m$. For accessing m's characteristics, the graph is set to the state of the movie's release date, so whatever happened after the movie's release does not affect its features.

With such a model, we are able to represent the whole producing network from IMDb in a graph. However, for computing process limitation, we need to focus on those movies whose analysis will produce significant results, leaving out possible outliers, as explained in Section 4.1.

From the whole graph, we can access social and historical features from movie producers. Specifically, we consider well known ego-metrics that are solely dependent on the node and its relations to the graph [3] (*closeness*, *betweenness*, *clustering*, *square clustering*, *network constrain*[3], *degree*), two known pairwise metrics from social network analysis (*neighbor overlap* and *number of shared friends*), and the new ones *past experience* and *previous success*.

4. DATA, PARAMETERS AND GROUPS

Overall, we want to analyze movie production teams by extracting their social features that can help to explain their success. To do so, we collected the full IMDb dataset[4] in November 2014, with movies from the late 1800's to 2014 from all over the world. IMDb informs full cast and crew,

[3]Index measuring the extent of the bridging of the node (whether the node connects different clusters) [3].
[4]IMDb sources: http://www.imdb.com/interfaces

average and total number of ratings[5], gross, box office, runtime in minutes, genre, etc. Next, we complement the information regarding the network model by summarizing the dataset filtering process, defining the success parameters and grouping the movies in order to facilitate the analysis.

4.1 Dataset Filtering

Building a graph for IMDb is costly and would consider all kinds, sizes and shapes of movies. To get significant results with minimum noise (i.e, minimum outliers and irrelevant data as possible), we prune nodes and their edges as follows.

Producer's Current Activity. Following [15], once a producer has not worked for seven consecutive years, his/her node is removed (likely retired, or even dead).

Movie Release Date. The movie industry only firmed itself with the first Academy Awards (1929), so we divide the dataset in 1930: movies released in [1888–1929] bootstrap the graph creation with starting nodes and edges, and movies from 1930 on are considered for our analysis.

Type of Movie. We focus on feature-length movies (at least 40 minute long) produced for cinema. Over 238,000 movie titles were collected, with over 1.7 million people (246 thousand production teams).

Relevance. Only movies with at least 1,000 *votes* are considered. Then, the dataset was reduced to 19,448 titles (about 8.5% of IMDb) with 39,808 producers[6].

Team Size and Connectivity. Movies with one producer (*one* is not a team) or whose producers are not part of the network giant component are discarded. The final dataset contains 12,250 movie titles with 31,696 producers[7].

4.2 Movie Success Parameters

There is no single definition for movie success. Hence, based on IMDb data, we consider three different numbers as *success parameters*: *economic success* given by the gross income (directly connected with the revenue and how many people payed to watch the movie); *public acceptance* given by the IMDb user rating (how well the title was received by the public); and *movie popularity* given by the absolute number of votes the movie received (number of people who have watched the movie and were interested in evaluating it). Next, we further discuss each of those parameters.

The trust on the movie's *rating* increases with the number of votes, and average ratings for movies with thousands of votes cannot be directly compared to those with only dozen votes. The number of votes also follows a power law, spanning several orders of magnitude. Hence, movie ratings were normalized by their number of votes on the same true Bayesian estimate used by IMDb in its TOP 250 list, i.e., $(\frac{v}{v+t_v}) \times R + (\frac{t_v}{v+t_v}) \times C$, where, for each movie, R is its mean of ratings, v its number of votes, t_v a threshold equals to the least amount of votes of a fully trustable rating, and

[5]IMDb registered users may vote in any title in the database by grading it once. Then, the *rating* of a movie is a weighted average of all votes received by it. A *review* is an opinion provided by a registered user to a movie.

[6]We have also analyzed a dataset for movies with at least 100 votes (65,493 titles representing 28% of IMDb, with 92,128 producers). However, not all network metrics were computable with such larger dataset, but we were able to identify similar results for those that finished processing.

[7]The final dataset is available here: http://goo.gl/0BPffy

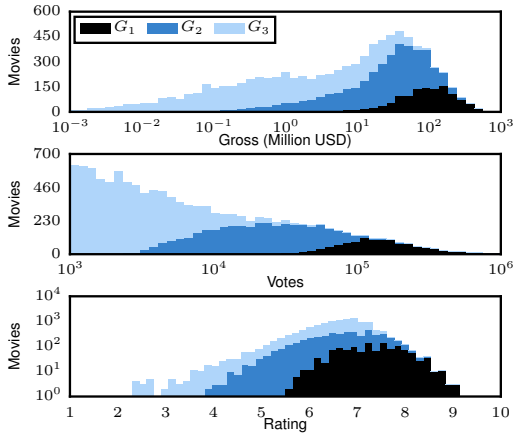

Figure 2: Histogram of movies per success parameter aggregated as the three performance groups.

C the mean vote across all movies (which is provided by IMDb[8] and was equal to 7.0 in November 2014). A value of 2,500 was used at t_v for keeping the movie rating's distribution log-normal (as seen in Figure 2) while also effectively penalizing movies with too few evaluations.

The gross income is available for only 9,210 (70%) movies. Also, the monetary values are usually given in the currency of the country that hosted the movie production and is dated from shortly after the movie's release. To compare monetary values with minimal distortion, the values were normalized: monetary figures were converted to US Dollars using the Historical Currency Converter Web Service[9]; and the corresponding amount in US Dollars was subsequently corrected for inflation by using the CPI Inflation Calculator[10], in October 2014. Movies without gross information are not considered at the analyses that predict gross. Figure 2 shows the distribution of the number of movies regarding the success parameters after normalization.

For analyzing the different degrees of movie success, we have defined three groups (Figure 2). Specifically, we scale the three success parameters in a [0, 1] interval and calculate the simple mean of (the scaled values of) their three success parameters. Finally, we aggregate them into three major groups: G_1 with blockbusters, G_2 with movies of moderate success, and G_3 with movies of low success.

Regarding time distribution, the IMDb website started collecting ratings from movies in the 90s. Nowadays, only the good movies from the past are watched and voted on. We can confirm this hypothesis in the ratings and votes distributions across decades, as illustrated in the bottom part of Figure 3. Recent decades present movies with more votes, and their grades are also more heterogeneous, i.e., when compared to the mean value, their rating values are more disperse than in previous decades. There are also more movies with a much lower box office these days. The distribution of movies in Figure 3 clearly shows those trends, as the number of new movie productions exploded in the last decades.

[8]IMDb Charts: http://www.imdb.com/chart/top

[9]Historical Currency Converter: http://currencies.apps.grandtrunk.net

[10]CPI Inflation Calculator: http://www.bls.gov/data/inflation_calculator.htm

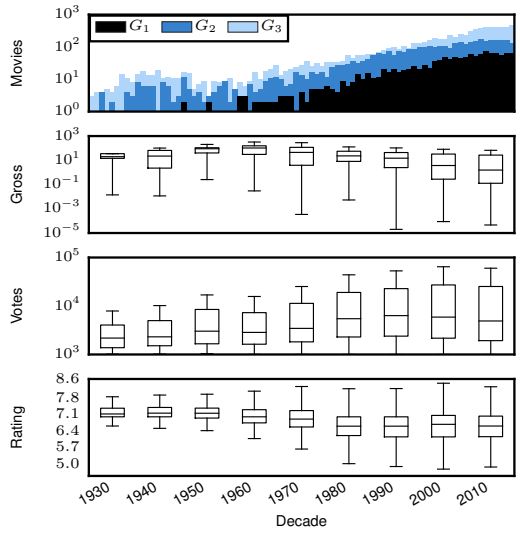

Figure 3: Movie productions and their success parameters through the decades.

5. PROPOSED METHOD

In this section, we explain the setup to accomplish the goal of improving movie success prediction by accessing the importance of topological (social) characteristics from teams. We define our prediction task and regression model (Section 5.1) and formally define their movie's features (Section 5.2). Finally, we explain our methodology to evaluate how topological features predict movie success (Section 5.3).

5.1 Movie Success Prediction Task

The movie success prediction task can be formally defined as follows. Consider the set of movies $\mathbb{M}$ as the feature length movies for cinema as defined in Section 4.1. Each movie $m \in \mathbb{M}$ has a set of features $\mathbb{F}$, and each feature $f \in \mathbb{F}$ provides information about m. For each success parameter (gross, votes and rating), the problem lies in adjusting all k coefficients of a linear combination of all f features so that this linear combination best fits the success parameter being modeled considering all movies in $\mathbb{M}$. Such fitting of coefficients is performed by a regression model, which estimates the relationship between the features and each parameter.

We rely on a multivariate Bayesian Ridge regression model [12] to setup coefficients such that a predictor may forecast each of the three success parameters. Given $\mathbb{F}$, the regressor iteratively adjusts the k features' coefficients to maximize the marginal log-likelihood of the observations.[11]

5.2 Movie's Features

As [9], the prediction task considers only features available *before* the movie release (i.e, the model may be used to estimate the success of movies *beforehand*). Features related to network topology and past history are based on the metrics defined in Section 3. Overall, we consider 121 features spanned over three main aspects of movies: the motion picture itself, the production team topology, and team's past success and experience. They are summarized as follows.

Motion Picture Features. Movie's genre (21 possible values), production continent (6 possible values), runtime length, team size and budget (both genre and continent may present more than one value).

Topological Features. We consider six topological ego-metrics (closeness, betweenness, clustering, square clustering, degree, and network constraint) and two pairwise-metrics (shared friends and neighbor overlap). Then, we employ six math aggregations to get one value for a whole team: arithmetic mean, harmonic mean, median, maximum value, minimum value, and standard deviation. We also consider a node contraction algorithm to achieve one value from a set of nodes in the graph for the ego-metrics [16] (note that it is applied to calculate the ego-metrics only). Moreover, we consider three global metrics to get the network state at time of movie release: global clustering coefficient, average path length and small world coefficient. Hence, we consider 57 topological features (six ego-metrics aggregated in seven ways, two pairwise metrics aggregated in six ways, plus three global ones).

Teams Past. The team past is given by four ego-metrics (previous rating, gross, votes and experience) and one pairwise metric (past experience). The former may be grouped by using the seven aforementioned aggregations (28 features), and the latter by the six mathematical aggregations (6 features), giving 34 features.

Finally, besides each feature individually, combining them is also important. Hence, we added the product of every feature pair as extra input features for the regression model.

5.3 Evaluation Methodology

Having defined our regression model and our feature set, we now detail our methodology for evaluating the predictive power of features that depend on team's social characteristics. Hence, we train and test the same regression models with the same set of movies $\mathbb{M}$, but using different sets of features: only the non-topological ones, only the topological ones and all of them. Then, we use the coefficient of determination R^2 to access how well the regression models are on predicting movie success parameters.

Our evaluation is divided in two parts. First, we analyze all 121 features to identify those with *greater dependence* with movie's success. Then, we remove those features that do not impact the model's predictive accuracy and assess the remaining's predictive power. To do so, we split the movies in three chronological groups: from 2008 to 2013 (3,317 movies = 27%), from 1995 to 2013 (9,775 movies = 52%), and from 1930 to 2013 (12,250 movies = 100%). These intervals provide different levels of chronological heterogeneity to the analysis (approximately the most recent quarter, the most recent half and all data). We use each of these three groups for training and testing the regression models for each of the three success parameters.

Choosing a single split of train/test sets might produce poor results, because the train and test sets could be selected in a biased way. To avoid this problem, we perform *Cross Validation using five folds*. For each cross validation, the movies are randomly split into five distinct groups, and five train/test cycles are performed. Each time, one of the groups is chosen to be the test group, and the other groups are used for training the regression model. In each train/test cycle, the R^2 is computed by amplifying the model to the test set, which gives five R^2 values for each run. We consider

[11] We have also tried two other regression algorithms – Support Vector Regression and Ordinary Least Squares. As the Bayesian Ridge regression model outperforms them, we present only results obtained with it.

Table 1: Correlation for features generated from square clustering (left) and neighbor overlap (right), considering all aggregation metrics and success parameters (V: votes, R: ratings, G:gross).

Aggregator	Distance Corr.			Spearman		
	V	R	G	V	R	G
Std. Deviation	.24	.08	.16	−.23	.06	−.19
Contraction	.17	.10	.08	−.33	.13	−.13
Maximum	.26	.10	.16	−.34	.13	−.20
Minimum	.23	.14	.07	−.37	.19	−.10
Mean	.27	.10	.17	−.38	.15	−.23
Median	.26	.10	.17	−.40	.15	−.20
Harmonic Mean	.26	.13	.12	−.40	.16	−.23

Aggregator	Distance Corr.			**Spearman**		
	V	R	G	V	R	G
Std. Deviation	.09	.13	.16	.04	−.12	−.17
Contraction	.17	.10	.08	−.33	.13	−.13
Maximum	.11	.13	.13	.12	−.14	−.14
Minimum	.21	.09	.15	−.18	.06	−.14
Mean	.19	.06	.25	−.15	.00	−.25
Median	.19	.05	.23	−.13	.01	−.21
Harmonic Mean	.24	.09	.22	−.24	.09	−.20

the mean of these values as the result of the cross-validation execution. Finally, it is important to ensure the special five folds were not randomized in a *biased* way. Hence, when evaluating a model, we execute 30 different cross validation runs and take the mean and confidence intervals ($\alpha = 95\%$) from such runs as the model's final score.

The distribution of votes and gross are highly skewed forming a power law (few movies with most votes and gross). Such skewness might bias the regression model and compromise its accuracy. We have attempted under-sampling the number of movies with lower levels of success in the training sets, but that did not improve regression accuracy. Hence, we generate *balanced* train/test folds. The folds are still generated randomly; however, using balanced sets ensures the equal distribution of movies among each of the three performance groups (G_1, G_2, G_3) in train and test sets.

6. EXPERIMENTAL ANALYSIS

First, we evaluate the dependency of the 121 features and the success parameters (Section 6.1). Then, we compare the baseline regression (non-topological features) against using only topological features and using all selected features (Section 6.2). Finally, we take all results together and derive insights on how to produce a blockbuster (Section 6.3).

6.1 Feature Selection and Strength

Selecting only important features for the regression model is crucial as it reduces the input noise, which in turn increases the prediction accuracy and reduces over-fitting. First, we analyze each feature correlation with the success parameters using the Pearson correlation coefficient [5]. The results were *low* for all features. Then, visual inspection of the scatter plots confirmed that there is no evident *linear* relationship among (practically) all features and the success parameters.

Therefore, we performed a second analysis with: (*i*) the *distance correlation*[12] to detect the strength of non-linear dependence among variables, and (*ii*) Spearman correlation, for monotonic relationships [5]. Table 1 shows the results considering only square clustering and neighbor overlap, in which many features present high degrees of dependence (other metrics provided similar results). We may also visually identify the most informative features by plotting distributions over the three performance groups. Figures 4 and 5 show the clear differences among the group distributions.

Based on these initial analyses, we excluded the features with lowest levels of dependency regarding the three success

[12]Distance correlation is a measure of statistical dependence between two variables. It is equal to zero if and only if the variables are statistically independent.

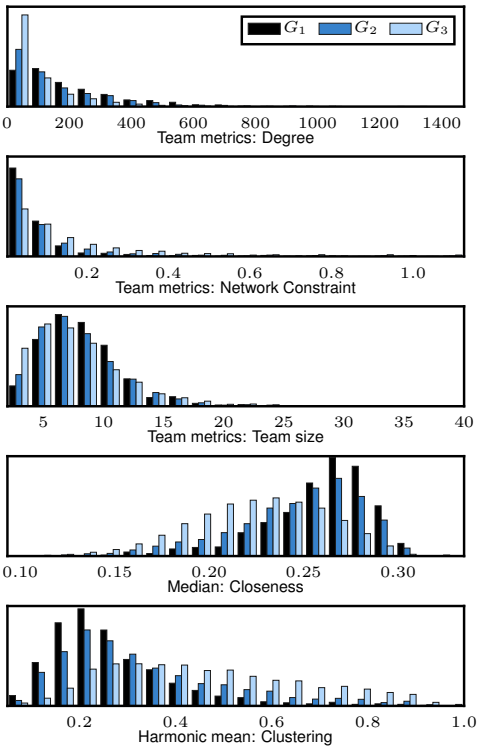

Figure 4: Distribution of topological features for movies in 3 difference performance groups.

parameters. Then, we kept iteratively removing features that did not cause any performance loss, until no further feature could be removed. At the end, only 23 features remained: **19 non-topological**, with 9 genres (romance, comedy, horror, adventure, thriller, mystery, drama, action, and documentary), 3 continents (North America, Europe, Africa), runtime, budget (log transform), previous success (aggregated by node contraction) given by mean ratings and mean votes, previous success (aggregated by mean) given by mean gross and mean votes, and previous experience (aggregated by node contraction) given by number of past joined productions; and **four topological**: degree (aggregated by node contraction), team size, closeness (aggregated by median), and clustering (aggregated by harmonic mean).

Besides identifying the most relevant features, assessing their *strength* by evaluating their coefficients in the regression model is also important. Note that features with a higher coefficient have more impact in the predicted value. Hence, we take an already trained regression model (the one

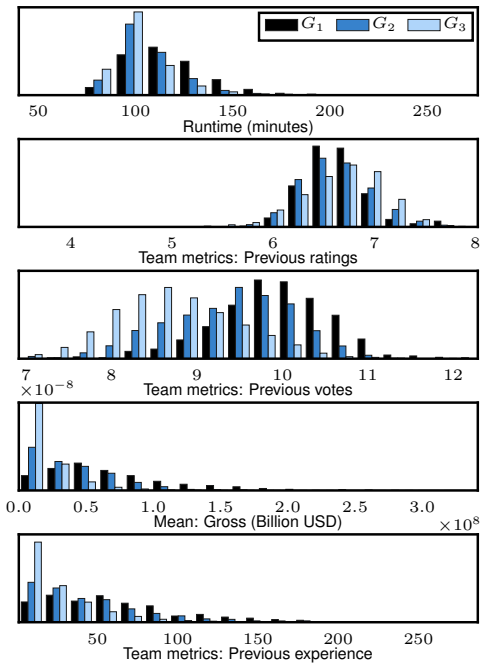

Figure 5: Distribution of non topological features for movies in 3 difference performance groups.

for votes, whose fitting was the best among the three parameters) and sort its coefficients in descending magnitude. The ones with best results were: the team's mean previous gross, mean previous experience, the harmonic mean of clustering coefficients, contracted degree, genres Drama and Documentary, runtime, and continent Africa.

Out of all topological features, the impact of the *clustering coefficient* and *degree* are significantly higher (actual values not presented due space constraints). Also, teams with more ties to producers outside the current team are more likely to succeed, as are teams with low levels of clustering coefficients. These indicate that more homogeneous teams tend to perform worse. So, a very relevant conclusion is assembling heterogeneous teams with many *external connections* is key to form teams with higher success odds. Moreover, such results reinforce the theory of weak ties (or structural holes) that determine the importance of having nodes acting as *bridges* within a successful network [2, 3].

6.2 Prediction Algorithms and Results

We now evaluate the prediction power of the regression models under three perspectives.

Chronological Evaluation. We evaluate the regression models for predicting the success parameters based on each chronological test set (starting in 1930, 1995 and 2008). Table 2 shows the R^2 obtained for each model instance.

First, it is clear that considering a wider range of years in the dataset *decreases* the effectiveness of the prediction model. We speculate that it happens because the organization of movie producing teams is constantly changing. For example, other analyses on the datasets (not illustrated due to space constraint) show an uprising in the clustering coefficient until late 1940's, then a downsizing until mid 1960's, with a very slow growing until 2013.

Second, the prediction accuracy is higher for number of votes, a little bit lower for gross, and very low for normalized ratings. The absolute number of votes a movie receives is the less noisy variable out of the three, as more outside aspects interfere with gross and ratings. Indeed, gross results from the audience that pays to watch the movie in theaters shortly after the release, which is affected by the movie's distribution and advertising (and the spread of movie piracy). Likewise, high ratings come from good cast performance and cultural/emotional response from the audience. Now, the number of votes strictly reveals the public *attention* to the movie, since a person may watch and rate it without necessarily having to pay for and even liking it. Finally, such a broad prospectus of results highlights (once more) that movie success prediction is a *difficult task*.

Nonetheless, for all cases, the prediction model with topological features, past success, past experience and movie's characteristics outperforms the others. Therefore, topological features do have predictive power over the movie's success parameters. Indeed, we may wonder that topological features *alone* could still be used to train movie success regression models with better results than purely guessing.

"Super"-Regressor. We now study the set of 23 features expanded with two success parameters to predict the third one (helpful when having two parameters already computed). To predict number of votes, using all selected features plus movie ratings and gross gives a Bayesian Ridge regression with $R^2 = 0.670$, confidence interval $= 0.0007$ and $\alpha = 95\%$. The same strategy for gross results in $R^2 = 0.573$, confidence interval $= 0.0013$ and $\alpha = 95\%$, and for ratings in $R^2 = 0.356$, confidence interval $= 0.0013$ and $\alpha = 95\%$. Finally, these results hint that an even better prediction model might be obtained for long-term success of movies considering its initial success in different metrics.

6.3 Producing Successful Movies

Figures 4 and 5 show movies that perform better are well determined by their features' ranges. We believe that new teams whose characteristics fall in the same range of values as other successful teams would be more likely to succeed as well. For example, from Figure 4, movies from the G_1 group (blockbusters) are more likely to have a low clustering coefficient [0-0.3] when compared to G_2 and G_3. Moreover, successful teams are likely to have degree larger than 100, network constraints no greater than 0.1, team sizes between 6 and 10 people, and closeness greater than 0.25.

Regarding past experience, teams with a combined value

Table 2: Coefficient of determination (R^2) and confidence intervals for a significance level of 95% obtained with a Bayesian Ridge regression for different configurations of year range of the dataset, and features employed in the regression.

Target	Years	Non Topol.	Topologic	All (14)
Votes	2008–13	.529, ±.0008	.310, ±.0006	.556, ±.0008
	2000–13	.484, ±.0004	.294, ±.0005	.517, ±.0004
	1990–13	.437, ±.0003	.246, ±.0004	.464, ±.0003
Gross	2008–13	.431, ±.0008	.170, ±.0013	.448, ±.0009
	2000–13	.419, ±.0004	.175, ±.0005	.447, ±.0004
	1990–13	.392, ±.0004	.174, ±.0004	.435, ±.0003
Rating	2008–13	.271, ±.0011	.033, ±.0009	.281, ±.0012
	2000–13	.267, ±.0006	.038, ±.0003	.273, ±.0006
	1990–13	.258, ±.0004	.031, ±.0003	.262, ±.0005

of 50+ movies are more likely to produce highly successful movies. Also, teams whose previous movies got an average mean gross of USD 50 million have a higher chance of producing successful movies. The mean rating from movies produced by team members should not deviate much from the global mean (6–7). Finally, one should ideally pick producers whose previous movies received on average a log(votes) of about 9 (about 20 thousand votes).

Predicting success with only data available beforehand is feasible. While considering a large dataset is good, spanning it through many years may *negatively* affect the quality of the regression models. Finally, for all cases, the prediction model with topological features combined with past success and experience have outperformed the others.

7. CONCLUSION

We studied the correlation between movie success and characteristics of movie production teams. Our work is novel for: regarding performance prediction with *aggregate* topological features from teams and considering a whole, big dataset. Nonetheless, a larger and more heterogeneous dataset may *decrease* prediction accuracy. Such fact suggests that movie success prediction might require more sophisticated models, further considering temporal phenomena and topological aspects. Overall, our main conclusions are sum up as: topological organization of movie production teams constantly changes; the "rich-gets-richer" phenomenon happens with movies (i.e., teams with successful past are more likely to produce successful movies); including fresh producers into teams, who act as bridges in the social network, is also crucial; and certain patterns in topological organization of teams are associated with success. This last one has a great impact for all team formation research, since our work represents only a first look at team topology for movie performance. Moreover, if we could generalize such claim to beyond movie production teams, such an impact could be extended to many industries. Then again, it is important to point out that one challenge is to acquire the dataset to support a predicting model. As for future work, we plan to expand the predicting model. Having the whole cast is too challenging computational-wise, so a possibility is to include only the top high paying members and directors as well.

8. REFERENCES

[1] A. Anagnostopoulos et al. Online team formation in social networks. In *WWW*, pages 839–848, 2012.

[2] R. S. Burt. Structural Holes and Good Ideas. *The American Journal of Sociology*, 110(2):349–399, 2004.

[3] R. S. Burt. *Brokerage and closure: An introduction to social capital*. Oxford University Press, 2005.

[4] E. Cohen et al. Computing classic closeness centrality, at scale. In *COSN*, pages 37–50, 2014.

[5] G. W. Corder and D. I. Foreman. *Nonparametric Statistics: A Step-by-Step Approach*. Wiley, 2014.

[6] A. Elberse. The power of stars: Do star actors drive the success of movies? *Journal of Marketing*, 71(4):102–120, 2007.

[7] V. Etter, M. Grossglauser, and P. Thiran. Launch hard or go home!: predicting the success of kickstarter campaigns. In *COSN*, pages 177–182, 2013.

[8] B. Fields et al. Analysis and exploitation of musician social networks for recommendation and discovery. *IEEE Trans. on Multimedia*, 13(4):674–686, 2011.

[9] M. Ghiassi, D. Lio, and B. Moon. Pre-production forecasting of movie revenues with a dynamic artificial neural network. *Expert Systems with Applications*, 42(6):3176 – 3193, 2015.

[10] T. U. Grund. Network structure and team performance: The case of english premier league soccer teams. *Social Networks*, 34(4):682–690, 2012.

[11] D. Kim et al. A user opinion and metadata mining scheme for predicting box office performance of movies in the social network environment. *The New Review of Hypermedia and Multimedia*, 19(3-4):259–272, 2013.

[12] D. J. MacKay. Bayesian interpolation. *Neural computation*, 4(3):415–447, 1992.

[13] A. Oghina et al. Predicting IMDB Movie Ratings Using Social Media. In *ECIR*, pages 503–507, 2012.

[14] D. Stokols et al. The science of team science: overview of the field and introduction to the supplement. *American Journal of Preventive Medicine*, 35(2):77–89, 2008.

[15] B. Uzzi and J. Spiro. Collaboration and creativity: The small world problem. *American journal of sociology*, 111(2):447–504, 2005.

[16] D. B. West et al. *Introduction to graph theory*, volume 2. Prentice hall Upper Saddle River, 2001.