PREVISÃO DE SUCESSO DE TIMES UTILIZANDO TOPOLOGIA SOCIAL

WLADSTON FERREIRA FILHO

PREVISÃO DE SUCESSO DE TIMES UTILIZANDO TOPOLOGIA SOCIAL

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientadora: Mirella Moura Moro
Coorientadora: Ana Paula Couto da Silva

Belo Horizonte
Dezembro de 2015

WLADSTON FERREIRA FILHO

PREDICTING TEAM SUCCESS USING SOCIAL TOPOLOGY

Dissertation presented to the Graduate Program in Computer Science of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Master in Computer Science.

Advisor: Mirella Moura Moro
Co-advisor: Ana Paula Couto da Silva

Belo Horizonte
December 2015

Ferreira Filho, Wladston
X0000x Predicting Team Success Using Social Topology / Wladston Ferreira Filho. — Belo Horizonte, 2015
, 0 f. ; 29cm

Dissertação (mestrado) — Federal University of Minas Gerais
Orientadora: Mirella Moura Moro

Coorientadora: Ana Paula Couto da Silva

1. Social Networks. 2. Team Formation. 3. Regression analysis. 4. Success Prediction. I. Title.

CDU 000.0*00.00

[Folha de Aprovação]

Quando a secretaria do Curso fornecer esta folha,
ela deve ser digitalizada e armazenada no disco em formato gráfico.

Se você estiver usando o `pdflatex`,
armazene o arquivo preferencialmente em formato PNG
(o formato JPEG é pior neste caso).

Se você estiver usando o `latex` (não o `pdflatex`),
terá que converter o arquivo gráfico para o formato EPS.

Em seguida, acrescente a opção `approval={`*nome do arquivo*`}`
ao comando `\ppgccufmg`.

Se a imagem da folha de aprovação precisar ser ajustada, use:
`approval=[`*ajuste*`][`*escala*`]{`*nome do arquivo*`}`
onde *ajuste* é uma distância para deslocar a imagem para baixo
e *escala* é um fator de escala para a imagem. Por exemplo:
`approval=[-2cm][0.9]{`*nome do arquivo*`}`
desloca a imagem 2cm para cima e a escala em 90%.

To my mother, who pushed me through the difficult times during this research, which happened to take place in the hardest part of my life so far. She provided me all the support I could ever need. Thanks Mom, this thesis is dedicated to you with love.

Acknowledgments

Foremost, I would like to express my greatest gratitude to my advisor, Prof. Mirella Moro, for always supporting my research with patience, knowledge and invaluable insight. She gave me the freedom to explore research questions that most interested me, while always guiding me towards producing relevant work. I owe all my academic achievements to her. I thank my co-advisor, Prof. Ana Paula Couto, for helping me reach a higher technical quality in this work.

I would also like to thank my parents for greatly encouraging me to research, my friends Rômulo, Christophe and Leonardo for pushing me and cheering me up in the hardest times. I thank Prof. Fabricio Benevenuto and Prof. Krishna Gummadi, who together with my advisor Mirella set me on this path.

I thank my fellow colleague Pedro Santos, for working with me and being a great companion in the earliest phase of the research. Finally, I'd like to thank Dr. Ronan Rêgo, for providing me the means to keep working when adverse moments arose. I also thank my friend Raimondo Pictet for helping me review the final work.

"Do. Or do not. There is no try."
(Master Yoda)

Resumo

O trabalho em equipe está presente em várias atividades importantes na sociedade, como por exemplo na produção de filmes e shows, de trabalhos científicos, nos esportes e em escritórios. Na maioria dos casos, os agentes que trabalham em uma mesma indústria se organizam em uma rede social, e times são formados não de forma aleatória, mas no contexto dessa rede. Em trabalhos colaborativos, é necessário que os agentes escolham suas equipes, idealmente seguindo uma estratégia que maximize a probabilidade de formação de um time de sucesso. No presente trabalho, considerando a tarefa de prever o sucesso de um time inserido em uma rede social, avaliamos o poder preditivo de características puramente sociais de times em comparação a outros fatores não-sociais sabidamente correlacionados com o sucesso de equipes. Ao contrário de outros trabalhos propostos para prever o sucesso de times ou avaliar a influência de fatores sociais no sucesso de equipes, nossa abordagem considera múltiplos fatores sociais, um grande número de times e uma vasta rede social. Nossa abordagem consiste em utilizar dados históricos de co-produção de filmes para montar um grafo representando a rede social de agentes produtores de filmes, e extrair características sociais e não sociais dos times de produção de cada filme utilizando esse grafo, para em seguida avaliar o poder preditivo dessas características em relação ao sucesso dos filmes. Apresentamos uma caracterização da rede de produção de cinema ao longo das décadas e uma avaliação do poder preditivo dos fatores topológicos estudados, juntamente com sugestões de como esses resultados podem ser utilizados na composição de novos times. Nossos resultados mostram que algumas características topológicos de times ajudam a tarefa de previsão de sucesso, complementando as demais características que já se sabiam ser relacionadas ao sucesso de times.

Palavras-chave: Redes Sociais, Formação de Equipes, Análise de Regressão, Previsão de Sucesso.

Abstract

Teamwork is present in virtually all important activities in society, such as movie and show production, scientific research, sports and corporate offices. In most cases, agents working in the same industry organize themselves in a social network, and teams are formed in a non-random way according to this network. In collaborative work, agents must choose their team, preferably according to a strategy that maximizes the odds of forming a high-achieving team. In this thesis, considering the task of forecasting success of a team in the context of a social network, we assess the predictive power of purely social characteristics from teams in comparison to other non-social characteristics knowingly correlated with team success. In contrast to previous works on forecasting or evaluating the impact of social characteristics in team success, our approach considers multiple social characteristics and a large number of different teams in the context of a large social network. Our approach uses historical co-production data from movies to assemble a graph representing the social network of movie producing agents. It extracts several topological and non-topological characteristics from movie-producing teams according to this graph. Finally, it assesses the predictive power of topological characteristics being studied and presents insights on how such results can be used to compose better teams. Our results show that some topological characteristics from teams help forecasting team success, complementing non-topological characteristics already known to correlate with higher-achieving teams.

Palavras-chave: Social Networks, Team Formation, Regression analysis, Success Prediction.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Team collaboration is ubiquitous in our society, notably within movie and show producers, scientists, corporate teams, book editors, and even robots [??????]. In this context, a crucial problem is *how to form teams as to maximize their performance*. For example, a soccer coach would benefit from composing a team with higher winning odds; a university or department dean would prefer to fund research teams with higher potential of producing breakthrough results; and a manager from a company may require to rearrange a team in order to ramp up productivity.

The way people form teams is largely influenced by their professional contacts and past collaborations, i.e., ways in which agents have access to new information and opportunities. Most teams do not exist in isolation, but within a social context amidst other teams. Therefore, knowledge of the collaborative network from people who work in a given context is key information for the team formation problem.

A social network of agents that collaboratively work in a context emerges when combining connected co-workers from many different collaboration instances. This social network formed by agents engaged in *teamwork* can be mathematically modeled by a graph: a graph G is an ordered pair $G = (V, E)$, in which V is a set of vertices or nodes and E is a set of edges or lines between two elements of V. Considering the context of movie teams, such graph can easily be obtained from a movie dataset with records informing cast and crew. Here, each record features a production team that actively worked together, and therefore corresponds to a connected set of nodes.

In fact, research using social network analysis in these types of graphs reveals interesting relationships between topological features from the network and team performance [???????]. Known topological metrics correlated to team performance involve the study of structural holes [?] and the effect of *structural coefficients* of the entire network [?]. However, both studies do not consider the *aggregate* effect of many

individual topological metrics of agents forming teams, nor the *predictive* capacity of these features with respect to team performance.

Using social network analysis to access the impact of combined social features in team performance, rather than analyzing social features in isolation, is a novel perspective that can yield new evidence on how social elements influence team performance [**?**]. Moreover, multivariate analysis works best when large datasets are employed. Among all, the film industry has one of the largest and most detailed datasets available: the IMDb (Internet Movie Database at `http://www.imdb.com`), making it an ideal candidate for the case analysis. With such a rich dataset, we can perform multivariate predictive analysis to identify patterns in topological features related to team success.

Finally, teamwork in the film industry is ubiquitous. Teams are responsible for projects with large budgets that are expected to generate billions in revenue and impact spectators worldwide [**?**]. Hence, improvements to movie success forecasting techniques that result from this novel multivariate research are potentially relevant both in cultural and economic terms.

1.1 Motivation

A simple experiment may point towards the relevance of further investigating the effect of topological characteristics in team success. Figure **??** shows results of two sample predictors using the same test/train set. The baseline predictor uses a single training feature: the previous success of the movie's team members with respect to the parameter being predicted. The test predictor uses the same feature and an extra topological one: the total number of people who have previously collaborated with the movie's team members. Figure **??** shows a significant gain in R^2 measures[1].

Nonetheless, movie success forecasting is not a trivial task because many complex factors are involved: 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, and so on [**??**]. Additionally, movie success can be viewed in different ways: the box office (gross generated revenue), its profits (box office minus budget), its critic and public acclaim, or the movie's popularity (the number of people who watched the movie). Moreover, success in one dimension does not guarantee success in the others. This leads to the proposal of the following hypothesis explored in the present work:

In a multivariate predictive analysis of movie success, does using many

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

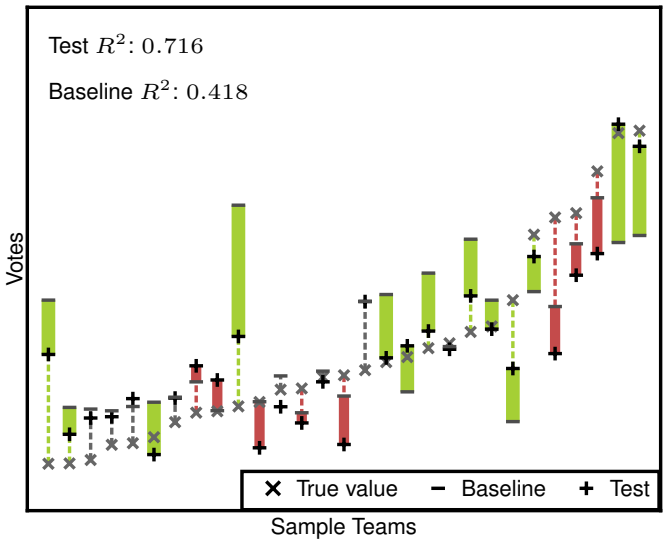

Figure 1.1: An example for predictive gain by using a single extra topological feature. Green and red bars represent prediction error reduction and increase, respectively.

topological features from teams lead to a more accurate analysis?

1.2 Objectives

In this thesis, we investigate such hypothesis by analyzing improvements in movie success prediction models based on multiple social features. To this end, a social network model is defined from movie production collaborations gathered from IMDb. Movie success parameters are normalized and social data from teams is calculated in many different ways. The prediction accuracy of models using different features from movies and its production teams are also compared.

1.3 Contributions

Our main contribution is a new movie success predictor model for movie popularity with an R^2 score improvement. With this result, we achieve two main research goals: improving movie success forecasting and high-achieving team formation. The contributions of this work are summarized as follows.

- We introduce a network model and selected metrics to assess the performance of movie producing teams. We analyze the whole set of movies from IMDb and

filter it according to well defined features: producer activity, release date, type of movie, relevance, team size and connectivity.

- We define three movie success parameters based on IMDb data (economic success, public acceptance and movie popularity) and use them to group all movies in three performance categories, which are extensively characterized.

- We propose a method for predicting movie success. The prediction task considers only features available *before* the movie release. Moreover, an important novelty is considering topological features, such as structural metrics and constraints related to the production teams' social network.

- Finally, we conduct a thorough evaluation analysis with a complete methodology. It is based on evaluating three different regression models using fivefold cross validation on balanced train/test sets. We also perform feature selection and strength analysis, and identify 23 features that can efficiently predict movies success. Our results show that topological features[2] extracted from teams provide predictive power.

This thesis is organized as follows. Chapter **??** presents basic information about the dataset employed in this work, along with key concepts from Social Network analysis, Machine Learning and Statistics. Chapter **??** discusses related work. We present our prediction model, dataset characterization and strategies for filtering and adjusting its data in Chapter **??**. We then describe our experimental methodology and analysis in Chapter **??**. Finally, we present our discussion, future work and conclusions in Chapter **??**.

[2]We use topological and social features interchangeably.

Chapter 2

Fundamental Concepts

In this chapter, we present fundamental concepts and review the related work that is mentioned or employed later on. Section **??** overviews techniques and metrics from Social Network Analysis. Then, Section **??** describes the main characteristics of the dataset studied, the IMDb. Finally, Section **??** presents techniques from Machine Learning that are used in our analysis.

2.1 Social Network Analysis

Social Network Analysis (SNA) emerged in the 20th century through the combination of network and graph theories as a fundamental approach to understanding social structures. SNA inherits from *graph theory* the description of these structures as sets of *nodes* connected by *ties* (or *edges*). SNA typically involves different metrics to evaluate *connections*, *distributions* and *segmentation* of a given network.

Metrics can be calculated from a node in the network, or from the network as a whole, to quantitatively characterize network aspects. For instance, centrality metrics such as closeness (introduced by **?**) and betweenness (introduced by **?**) could capture the importance or relevance of the node in the graph. Social metrics providing information that is related only to a single node are referred to as *ego-metrics*. Node clustering metrics [**?**] and network constraints [**?**] could indicate how heterogeneous the pool of connections of the node is and thus its access to new information. Node degree (i.e., the number of ties connecting it to other nodes) indicates how large the network of a node is, possibly indicating the amount of support the node receives in its productions.

Besides these *ego-metrics*, there are those to characterize the nature of the relationship between a pair of nodes. For instance, the *number of shared neighbours* is the

number of nodes to which both nodes in the pair are connected, whereas the *neighborhood overlap* is the ratio of the number of shared friends to all nodes connected to the pair. These metrics provide an indication of how socially close nodes in a pair are.

Finally, there are network metrics that refer to characteristics of the networks as a whole. Important metrics of this type include the *clustering coefficient* (measures the degree to which nodes in the network tend to cluster) and the *average shortest path length* (the average of the shortest path lengths between all possible pairs of nodes in the graph). The *small world coefficient* is the ratio of the global clustering coefficient to the average shortest path length. This measure is related to the connectivity and cohesion among elements in a network. The more a network exhibits characteristics of a small world, the more connected the agents are to each other and connected to agents who know each other through past collaborations [**?**].

Networks represented in graphs can be one-mode, with all nodes in the network of the same type, or two-mode, with two distinct types of nodes. Two-mode networks in which no ties exist connecting nodes from the same type (i.e., all ties connect two nodes from distinct types) are *bipartite networks*. For instance, a bipartite graph can represent an online bulletin board in which nodes are either users or discussions topics, and ties exist only between users and topics to indicate whether a user participated in a topic. A very famous class of two-mode networks are *collaboration networks*, in which one type of node represents human made artifacts (books, research papers, movies, Broadway musicals) and the other represents people who worked in creating the artifact. The network studied here is a two-mode network in which nodes are either movies or movie producers, and edges indicate that a movie producer worked in a movie.

It is also possible to transform a two-mode network into a one-mode network **?**. In this method, one type of node is chosen and a new network is created only with the chosen type of node. In the new (projected) network, ties are created between nodes that are connected to a common node in the original network. For collaboration networks, when projecting the original network for the authors, the projected network represents authors with ties linking authors that worked together.

Two-mode networks, such as collaboration networks, are rarely analyzed without prior transformation into one-mode networks, because the network metrics are not defined for or adapted to two-mode networks. However, projected networks always lose information from the original network, and calculating social metrics from projected networks requires special attention.

For instance, clustering coefficients from projected one-mode networks can be numerically high only due to the two-mode nature inherent to the original network. For

accurately obtaining the network measures from a bipartite network, a more complex analysis considers random graphs, as described by **?**. Also, a new clustering technique, called squared clustering, was developed by **?** to better identify the clustering effect in networks projected from bipartite networks. These techniques (comparison to analog random graphs and square clustering) are used in the present work in order to extract social information more accurately.

Most measures in bipartite networks are for single nodes. However, we need network measurements from *teams*, which are groups of nodes. In these cases, one possible approach is node contraction: a group of nodes is replaced by a newly created single node [**??**]. Edges reaching the initial nodes then reach the new node. If there exist multiple edges from the same origin reaching the new node, they are combined into a new single edge. In this case, the new edge has the sum of the weights of the previous edges. Extra information (such as past success and experience) carried by the individual nodes and edges which are now combined can be joined into the created node or edge. Finally, all the aforementioned ego-based metrics can be calculated on such a *supernode*. Note that the node contraction technique cannot be used for aggregating pair-wise metrics among all pairs in a group of nodes, since it aggregates all information into a single node.

2.2 The IMDb Dataset

The IMDb, an US-based company owned by Amazon.com, provides free access to its famous dataset of cinema, TV movies and series for research purposes at its website: `http://www.imdb.com/interfaces`. IMDb's data is modularly grouped by information type (such as "business", "biography", "plot"). Each data module is provided in a compressed CSV (Comma Separated Values) text file. Hence, reconstructing the required parts of the original dataset requires only a subset of the data modules.

The original dataset can be built as a SQL database from its constituent data-modules using the `imdbpy2sql` script, available from the IMDbPY tool[1]. This software includes many features for accessing and querying the built SQL database.

The IMDb dataset describes movies and their full cast and crew in extensive detail, including movie alternate names, associated keywords, references to other movies, gross (revenue), box office, runtime in minutes, genres[2], plot, production budget, audio

[1] IMDbPY: `http://imdbpy.sourceforge.net/`

[2] One or more out of 21 possible genres can be assigned to a movie. The complete IMDb genre list is available at `http://www.imdb.com/genre/`.

and color schemes, production countries and certificate ratings for various countries[3]. It also contains extra information about the companies that worked in several aspects of the movie such as special effects, sound mix and distribution. It lists movie producers, writers, directors, editors, soundtrack composers, actors and actresses, makeup personal, and even stunt actors. For each of these people, more information such as full name, short biography, birth date, birth city and height is sometimes available.

Anyone can register at IMDb and vote for any title in the database by grading it in a scale of [1–10] once. IMDb datasets also include movie *ratings* as weighted averages of all votes[4] and the absolute number of *votes* movies received. Most movies in the IMDb database are from virtually unknown productions, i.e. they received very little or no user ratings and reviews. More than half the movies in the database received less than 30 votes, whereas all professionally produced movies typically receive several thousands *votes*.

Here, the full IMDb dataset was downloaded in November 2014. At that time, it contained information of movies from the late 1800's to 2014, from all over the world, totalling approximately 1.4 gigabytes of data. Movies from 2014 were recently added to the dataset, and could have missing or unstable data. Therefore, movies from 2014 were not considered.

2.3 Machine Learning and Regression Analysis

Regression Analysis is a technique to estimate relationships among variables. Its goal is to build *regression models* that exploit such relationships to accurately derive values for one or more output variables according to given input variables. This concept is the founding block for most predictive analysis techniques, including Machine Learning [**?**].

Machine Learning is the development of algorithms that learn and make predictions from data. Rather than relying on static instructions, Machine Learning algorithms use existing datasets to train *regression models* that ultimately allow for accurate prediction and decision making. As a type of Regression Analysis, it involves evaluating target variables as functions of data features modeled as input parameters [**?**]. There are several regression techniques that can be used to accomplish such a goal, including *linear* regression techniques, the simplest being Ordinary Least Squares

[3]Each country has a legal entity responsible for issuing movie certificates. In the United States, it is the MPAA (Motion Picture Association of America: `http://www.mpaa.org/film-ratings/`).

[4]Ratings are different from *reviews*, which are more elaborate opinions provided by registered users.

(OLS). OLS consists in adjusting coefficients of a linear function of the input in order to minimize the square error of the model.

However, there are characteristics that might cause models based on OLS to perform poorly. These include excess noise in the target variable, incorrect feature selection, inclusion of too many irrelevant features, presence of non-linear dependencies between the input features and the predicted variable, and presence of extreme outliers. To handle these cases, many different regression techniques were proposed, each having particular applications where it performs best.

For instance, the Ridge Regression [?] is a variation of OLS to better handle the presence of outliers. It minimizes the sum of the square error, plus an extra term that brings the model to a more desirable state with the aid of a Tikhonov Matrix, using L2 regularization [?]. Support Vector Machines (SVM) [?] is yet another type of regression model indicated for recognizing non-linear relationships.

Another alternative is using models based on Bayesian Inference, such as the Bayesian Ridge [?] that uses a probabilistic model. Therefore, instead of minimizing the error or square error, these models are iteratively adjusted to maximize the model's observed likehood or log-likehood. Such a behavior results in regression models that handle greater amounts of noise in the output variable [?].

Even when choosing the most appropriate regression technique with respect to the data being analyzed, problems can still arise due to overfitting. This happens when the model is excessively adjusted to recognize the input data. It then performs really well for the training data but does not generalize to data not considered by the training [?].

To control overfitting and therefore produce a model that generalizes across diverse types of data, data is typically split into a *train set* and a *test set*. The train set adjusts the regression model to fit the training data, whereas the test set is never used to adjust the model. The test set is solely used for assessing how well the model generalizes to data it has not seen before [?].

However, doing a single train/test routine may result in biased outcomes due to the accidental choice of a train/test data split that does not generically and accurately represent the data. For this reason, the ideal method is not to perform a single train/test evaluation, but *multiple* evaluations of the model using cross validation. In this case, data is split among k folds (typically 5 to 10), and k train/test splits are performed. Each time, $k - 1$ folds are used for training and the remaining fold for validation. The final performance of the model should be derived from the mean performance from all k evaluations. [?]

The R^2 measure is the standard to understand how well the model fits the target

variable, i.e., how well the model can predict the target variable. It is the square of the correlation coefficient between the values predicted by the model and the real values. The R^2 measure is also known as the coefficient of determination, as it directly represents the total percentage of variation of the predicted variable that is explained by the model. An R^2 value of zero means the model fits the target variable no better than the mean of the target variable, whereas an R^2 value of one means the model can perfectly explain all variation in the target variable.

Most machine learning techniques require that all features are normalized in either the [0,1] interval or to have unit variance and zero mean. This is also true for Binary Features. These special types of features are defined after characteristics that can either be present or absent in the object being modeled. For instance, each movie in a set can either be associated with the genre "Comedy" or not. In this case, a Binary Feature can be used to input this information into regression models: a feature "Genre:Comedy" is defined, and it assumes the value 1 if the movie is associated with the aforementioned Genre, and 0 otherwise [**?**].

Chapter 3

Related Work

In this Chapter we go over previous work related to our goals. Specifically, Section **??** presents the literature related to social network analysis, team success analysis and prediction. Section **??** presents previous research efforts with goals similar to ours. Then, Section **??** presents the key points in which our contributions differ from previous work.

3.1 General Research on Team Performance Analysis and Social Network Analysis

? was one of the first to explore metrics from large collaboration networks. He demonstrated how to construct a social network from bibliographic data and how to extract interesting characteristics from the social graph by using complex networks metrics. He showed that different scientific communities form small-world networks and are highly clustered, and proposed a method for estimating tie strength. The technique used in this thesis to extract a social network from collaboration data closely follows Newman's work.

After the foundations for analytically extracting characteristics from social networks, and with the wider availability of social network data and processing power, this research area expanded across many academic disciplines, inspiring innovative work. **?** used the IMDb dataset and social network analysis to explore gender inequality by tracking differences between social features from actors and actresses. He found some network metrics are link to better chances of career advancement, and that women are more likely to drop out of their careers in comparison to men.

In another interesting application of social network analysis, **?** also used IMDb

movie data for estimating the relevance of a movie according to the network of cross references movies received. They considered the inclusion in the US Library of Congress National Film Registry as the ground truth for higher significance, and the contributions are more accurate in predicting whether a movie is highly relevant than the combined opinion of movie experts. On the other hand, we examine movie success in a broader context that includes a combination of popularity, financial success and public acceptance.

Studies in different social contexts try to understand how people work together to better achieve their goals. **?** explore the "Science of team science", focusing on the processes by which scientific teams organize and conduct their work. Specifically, they explore how teams connect and collaborate in order to achieve breakthroughs that would not be attainable by either individual or simple additive efforts.

Likewise, others identify correlations between social characteristics and team performance. **?** performed social network analysis in football teams and found that network characteristics in the graph that models the way players pass the ball is related to team winning odds. Unlike those work, here we focus on the social characteristics of teams in relation to their connections external to the team, instead of team's internal links.

Although there is a vast pool of previous work connecting social structure and productivity, few of them investigate social features in the context of the team formation problem. Moreover, the team formation problem is well established and concerns strategies and algorithms for efficiently forming teams in the best possible way. Many of previous work on this problem focus on efficiently selecting elements with different skill sets in order to form multi-functional teams.

For example, **?** present a technique for efficiently forming teams with a minimum set of combined skills. They also use the IMDb dataset to gather sample data to an emulated problem in which teams need to be formed from directors with a minimum combined skill set. Also, the genres of movies they have previously acted serve as the skill.

? also provide techniques for forming teams with a best possible multiple skillset. **?** form teams to work on given projects by matching nodes that have the specific knowledge related to the keywords present in the project's description. Deterministic team formation extends even to autonomous robots: **?** present techniques for autonomous robots (working in disaster zones) to dynamically form teams to better perform the task of locating humans in need of rescue.

Differently from these works, we contribute to the team formation problem not by efficiently finding elements to assemble a team that must satisfy a given restriction, but by identifying *social patterns* present in teams that perform better in the context

of producing successful movies. Moreover, other approaches on team formation tackle it as an optimization problem by grouping elements in order to most harmonically distribute the different abilities across multifunctional teams [?]. Other studies explore the presence of elements with special characteristics in the team to understand its success. For instance, ? studies the impact of extremely famous actors in movie teams. We complement such prior studies by focusing in accessing the effect of the social arrangement of the agents in the network, rather than the effect of individual characteristics of team members.

3.2 Team Success and Network Topology

The main focus here is to use topological network characteristics from teams to improve team success forecasting. Hence, we now go over approaches that look for relations between performance and social characteristics.

? study the network formed by the collaboration among 16 countries and show the small world coefficient of the whole network is correlated to more patent registrations from those countries. ? study the collaboration among 1,106 companies and find the small world metric is correlated to knowledge creation inside companies and innovation. Similarly, ? show the more Wikipedia[1] editors with higher social capital (taking part in a cohesive and centralized cluster) working in a wikipedia article, the faster the article reaches higher quality classifications. ? find specific kinds of network ties among open source developers are correlated with the development of more popular open source projects.

In a large study of Flicker, ? show nodes with higher degree have more influence than nodes with lesser degree, regardless of the node's credibility rankings. In the context of scientific research, ? show social capital can be estimated with metrics from social networks analysis, and researchers with more social capital publish research that has more impact and reachability.

Higher performance is often related to social capital deriving from increased access to Structural Holes, which in turn come from a privileged position occupied by the node within the social network ??. Burt claims competitive advantage is mainly derived from this phenomena and provides several ways in which individuals and business can occupy positions of less network constraint [?]. Indeed, ? investigated the interaction among 17 real state agents working on the same office and found that agents with higher network constraints had significantly worse performances in comparison to other agents.

[1] Wikipedia: `http://www.wikipedia.org`

Similar to us, **?** evaluated the network of Broadway musical producers (choreographers, writers and directors, not the cast) and found the artistic and financial success of such a network as a whole is correlated to its small world coefficient. The authors analyzed many network metrics and found some of those were correlated to success. Nonetheless, all these studies do not explore multivariate analysis of many network metrics. Also, when considering the aggregation of network metrics, previous research only employs the simple mean value among the agents' metrics, which can lead to bias because drastically different distributions can have the same mean value.

Regarding the prediction of movie success, **?** propose to mine public opinion and trends in order to predict it. However, their experimental dataset considers only 200 chosen movies (i.e., with clearly distinct features that may produce biased results) and their analysis is limited to the classification in four different success bins. Likewise, **?** propose to predict IMDb movie ratings based on textual comments from social media (Twitter and YouTube). They also consider a limited set of 70 movies for experimental evaluation.

? present 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), which contains abstract information such as the value the movie production has from its famous actors, or whether the special effects and the movie's competition are "High", "Medium" or "Low". They also include data purchased from a media research company that is hard to obtain for many movies (e.g., pre-release marketing expenses).

3.3 Final Considerations

We aim to forecast movie success parameters. However, besides financial success, we also plan to predict popularity and ratings from movies by considering social features from movie producers and a larger set of movies in our analysis. The aforementioned studies (and others that relate network topology and team performance) consider very small datasets. Unlike them, we make an extensive analysis of social features from teams by exploring a large set of 12,250 movies.

In social network analysis, considering such larger datasets is critical in order to yield more robust results. Our dataset is assembled from the whole set of feature-length cinema movies in the IMDb dataset that received at least 1,000 votes. It is composed by 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 than any of the previous work.

Furthermore, we also include features used by these studies in our analysis, whenever data was available in the IMDb dataset. Some data (such as genres) is easily available, whereas others (such as marketing expenses) are hard to obtain. Social features are used in addition to all other features in a very similar way as the prediction task presented by **?**.

To the best of our knowledge, we are the first to study the relation between network topology aspects and success by considering such a large network of motion pictures producers.

Chapter 4

Movie Data Processing and Success Prediction

This chapter presents the methodology necessary to predict movie success and a characterization of the dataset. We start by describing the steps for filtering, processing and interpreting the dataset along with characterizations in Section **??**. Then, we present our new methodology to correctly interpret the dataset into movie parameters suitable for analysis in a prediction model in Section **??**–this includes movie success parameters and social parameters from movie-producing teams. In Section **??**, we characterize the dataset in terms of such parameters. In Section **??**, we discuss feature interaction. Finally, in Section **??**, we present and describe the prediction model and methodology employed to assess how movie characteristics are related to movie success.

4.1 Dataset Processing

This section presents all pre-processing steps necessary to extract the proper information required for our analysis. We present the network model and strategies for filtering irrelevant information from the dataset, as well as extracting and normalizing movie success parameters. We then group the movies into three performance groups and present a characterization for the adjusted success parameters and performance groups.

4.1.1 Network Model

Assessing topological information from movie production teams requires to build a graph for modeling the interaction network of movie producing agents as accurately

as possible. Specifically, we construct a graph for movie producers only, leaving out the rest of the crew and cast (actors, directors, writers, etc.). The reason is that producers are the *core* of the team: they make the most important decisions and hire the rest of the crew, ultimately claiming most responsibility for the success of the movie. Furthermore, adding more types of agents to the model would make it orders of magnitude larger, exceeding the computing power available and possibly adding uncertainty to the results [**???**].

As discussed in Section **??**, most network measures are only defined for one-mode networks, so our network model cannot be directly based on a bipartite network. As presented in Section **??**, IMDb provides co-production data from movies that is directly mappable to a bipartite collaboration network. Therefore, one initial step in pre-processing the IMDb dataset is to derive an one-mode network from it. To this end, we employ the methodology proposed by **?**.

Formally, given a set of movies $\mathbb{M}$, each movie $m \in \mathbb{M}$ has a set of producers $\mathbb{P}_m$ (also called 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 $\mathbb{M}$. Moreover, edges in $\mathbb{G}_\mathbb{P}$ are undirected and link producers who have worked together in one or more movie. The number of previous collaborations between two producers is expressed by a weight value in the edge connecting them. For every node p, H_p is p's past experience (i.e. the number of p's previous productions), and S_p is p's previous success (i.e. the mean of success parameters for p's movies). Finally, $\mathbb{G}_\mathbb{P}$ evolves over time to incorporate new data when new movies are produced. Therefore, $\mathbb{G}_\mathbb{P}(t)$ represents the connections among producers considering only movies that were released up to instant t.

For any given movie m, its characteristics are based on the topology from the subset of nodes in $\mathbb{P}_m$. For assessing m's characteristics, the graph is set to the state of the movie's release date; hence, whatever happened after the movie's production does not affect its features. This is achieved by processing movies in chronological order. By doing so, once a movie is processed, the graph only contains information from the already processed, previously released movies.

Besides edges and nodes, an auxiliary data structure is annexed to $\mathbb{G}$ to accommodate necessary non-topological information from movies. As presented in Section **??**, the IMDb dataset provides an incredibly rich pool of information, including collaboration data from producing agents. Some of this auxiliary, pre-release information is extracted from the dataset and stored separately from the graph: movie budget (adjusted and normalized as is movie gross), movie duration in minutes, genre (such as Action, Drama, Sci-Fi, etc.) and production countries. This information is also used by **?** in their pre-release movie success prediction analysis.

With such a model, we are able to represent the entire producing network from IMDb in a graph. However, because processing power is limited, we need to focus on those movies whose analysis will potentially produce more significant results.

4.1.2 Dataset Filtering

Building a graph for the entire IMDb dataset is very costly and would include all kinds, shapes and sizes of movies. In order to get more significant results with minimum noise (i.e. minimum outliers and irrelevant data), we prune some nodes and their edges. Moreover, it is important to improve the accuracy of our graph model in representing how movie production parties interact, even if information must be removed from the graph. Data is filtered before the experiments according to the following parameters.

Producer Current Activity. According to **?**, once a producer has not worked for seven consecutive years, his/her node should be removed (we note that such person is likely to be retired or deceased).

Movie Release Date. The movie industry only cemented itself with the establishment of the first Academy Awards (1929) [**?**]. At that time, sound technology for movies emerged, people begun to consider movie theaters as an entertainment option, and professional movie production took off. Considering this, we split the dataset in 1930: movies released in [1888–1929] are used to bootstrap the graph creation with starting nodes and edges, and movies released from 1930 on are considered for our analysis. We note that evaluating social metrics on a network with very few nodes and edges may produce distorted results [**?**]. Therefore, using early movies to create a starting graph generates a well defined network in order to properly analyze the movies from 1930 onwards. Movies from 2014 on are disregarded because they were very recently added to the database, and therefore might contain unreliable (i.e. unstable) key information for the analysis, such as ratings or gross.

Type of Movie. We focus on feature-length movies (at least 40 minutes long[1]) produced for cinema. This includes documentaries, but leaves out TV productions and short movies as they present different goals and styles. It is also debatable if their ratings and gross can be compared to results obtained in the cinema. Over

[1]In order to be eligible for receiving an award from the Academy of Motion Picture Arts and Sciences (`http://www.oscars.org/`), a movie must fit their definition of "feature-length", then having 40 or more minutes of total runtime.

238,000 movie titles were collected, involving over 1.7 million different people, of which 246 thousand are producers.

Relevance. We are not interested in evaluating amateur productions, but professionally produced movies (see Section **??**). Therefore, only movies with at least 1,000 *votes* are considered. Thus, the dataset was reduced to 19,448 titles (about 8.5% of IMDb) and 39,808 producers[2].

Team Size and Connectivity. Movies with just one producer (which does not characterize 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[3], as collected in November, 2014.

4.1.3 Movie Success Parameters

As discussed in Chapter **??**, there is no single definition for movie success. Therefore, we have evaluated the data available at IMDb and decided to consider three different measures as *success parameters*. First, *economic success* is given by the gross income, which is directly connected with the revenue and represents how many people were interested in paying to watch the movie. Second, *public acceptance* is given by the IMDb user ratings and indicates how well the title was received by the general public. Finally, *movie popularity* is given by the absolute number of votes the movie received and represents the number of people who have watched the movie and were interested in evaluating it.

Figure **??** illustrates the correlation between these three variables. Note that the relationship presented in Figure **??**(a) between *economic success* and *popularity* confirms the positive correlation found by **?** for these parameters. Figure **??**(b) presents the relationship between movie's *public acceptance* and *popularity*. Note that this graph has a logarithm scale. Most movies have average *public acceptance* levels paired with low *popularity*, and movies in the higher spectrum of *popularity* have an above average *public acceptance*. Figure **??**(c) presents the relationship between movie's *public acceptance* and *economic success*. It clearly shows that movies with a higher *economic success* tend to deviate more from the average in terms of *public acceptance*.

Since movie *popularity* is positively correlated with *economic success*, and popular movies have *public acceptance* values that deviate more from the average (as shown

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

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

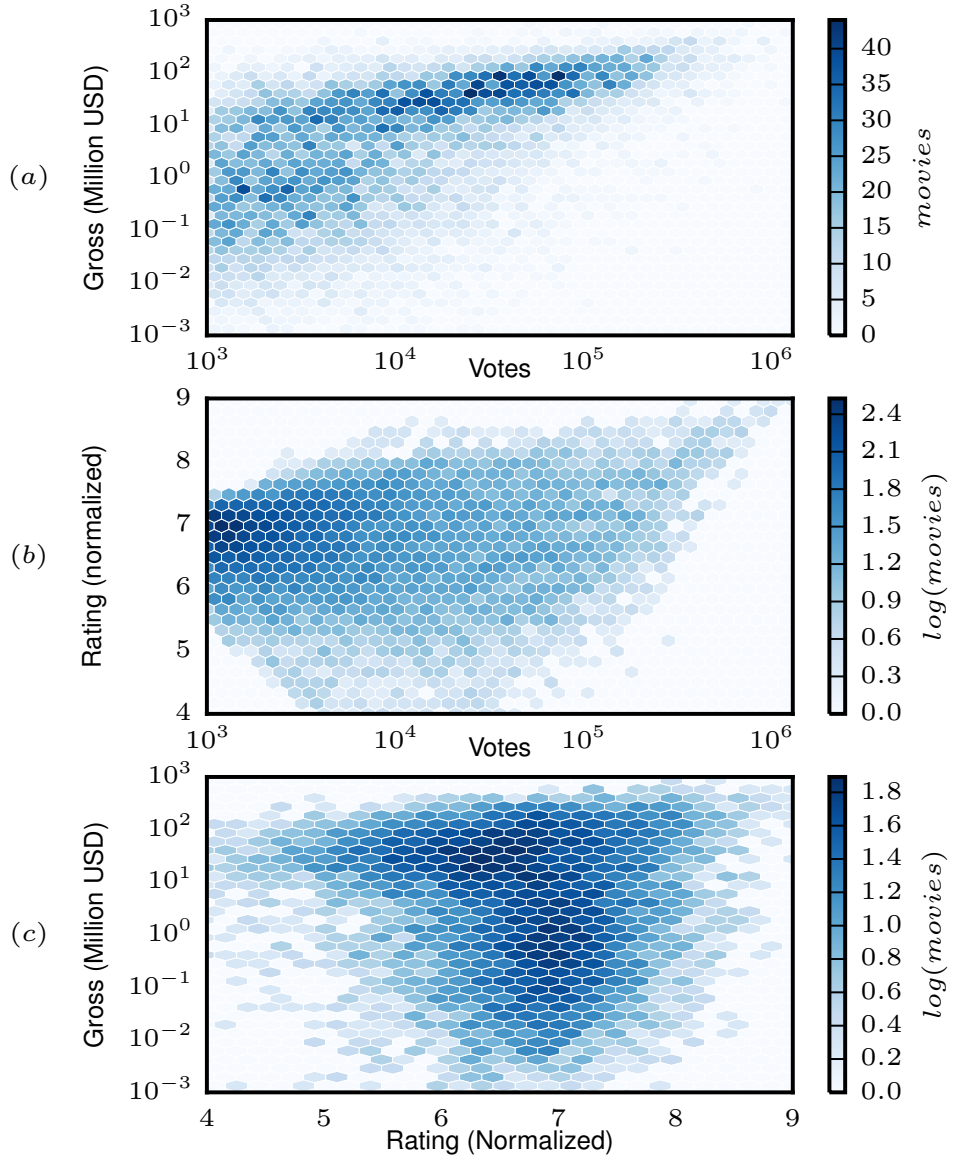

Figure 4.1: Relationships between movie's success parameters

Figure **??**(a) and (b)), we conclude that all three graphs are in accordance to each other. Movies with a higher economic success draw more attention, and therefore have a more diverse pool of opinions and evaluations. This translates to more heterogeneous values of *public acceptance*.

Next, we discuss strategies for normalizing each of such success parameter.

4.1.3.1 Average Rating

The trust on the values for average rating increases for movies that received more ratings (more votes), because a small number of biased evaluations can significantly change the movie's overall evaluation. Very popular movies have dozens of thousands of votes, and thus the bias generated by unfair evaluations is reduced. Therefore, average ratings for popular movies cannot be directly compared to those of movies that received only a few thousand or even less evaluations. Hence, average ratings were normalized by the number of votes received on the same true Bayesian estimate **?** used by IMDb in its TOP 250 list[4], according to the following equation:

$$\text{Weighted Rating} = \left(\frac{v}{v+t_v}\right) \cdot R + \left(\frac{t_v}{v+t_v}\right) \cdot C, \tag{4.1}$$

where for each movie, R is the mean of its ratings, v is the number of votes it received, t_v is a threshold equal to the least amount of votes of a fully trustable rating, and C is the mean vote across all movies. The value of C is provided by IMDb[5] and was equal to 7.0 at the time the dataset was collected (November, 2014). A value of $t_v = 2,500$ was chosen in order to ensure log-normal rating distribution (as seen in Figure **??**(c)). This ensures that movies with few evaluations (and therefore less trustable mean ratings) have their mean ratings pushed towards the global mean vote, effectively compensating for the higher probability of the movie having a biased deviation from the mean that is not due to the movie itself. For instance, there are several obscure productions with only a few hundreds of evaluations that have a mean rating that is as high as the most widely acclaimed movies. However, its unlikely that these movies have the same level of public acceptance.

4.1.3.2 Gross Income

Gross income information was only available for 70% of movies (9,210 instances). 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. Values for gross income and budget were converted to US Dollars using the Historical Currency Converter Web Service[6].

Since the dataset contains movies produced in many different decades, US dollar gross values from distinct moments in history cannot be fairly compared due to infla-

[4]IMDb Top 250 Movies: `http://www.imdb.com/chart/top`
[5]IMDb Charts: `http://www.imdb.com/chart/top`
[6]Historical Currency Converter: `http://currencies.apps.grandtrunk.net`

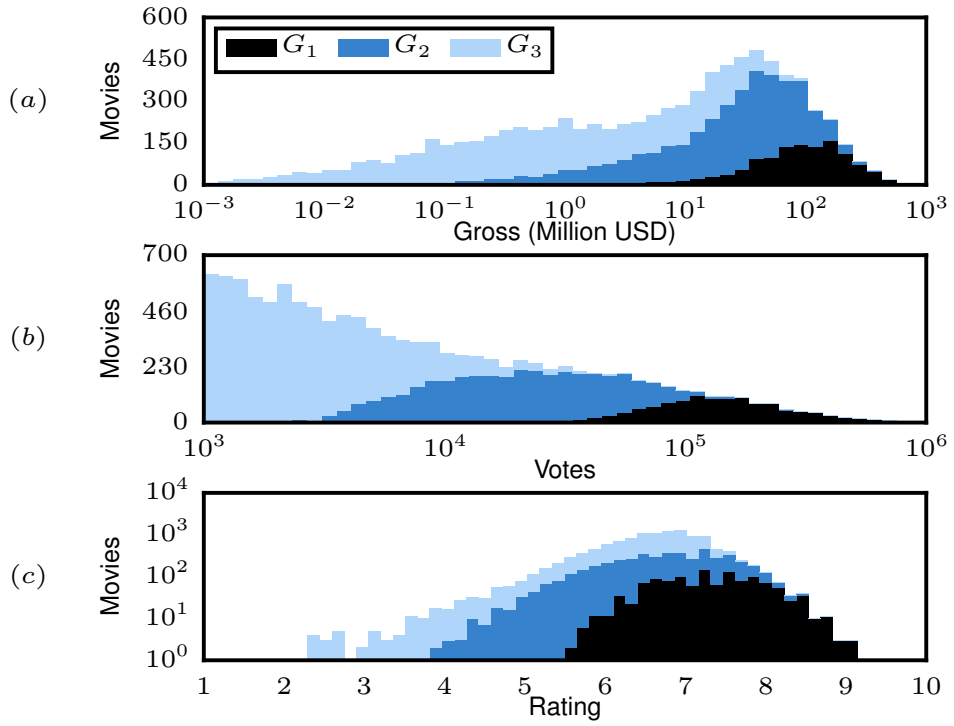

Figure 4.2: Histogram of movies per success parameter aggregated as the three performance groups.

tion. Therefore, the amounts in US Dollars were subsequently corrected for inflation using the CPI Inflation Calculator[7], an online feature provided by the Bureau of Labor Statistics, in October 2014.

The movies without a valid historical exchange record for US Dollars (only six instances) had their gross information discarded. Movies without gross information are not considered for analyses involving gross. Finally, Figure **??**(a) shows the movies' distribution of gross income, which follows the shape of a double-normal distribution.

4.1.3.3 Number of Votes

Figure **??**(b) shows the number of votes a movie receives follows a power law, spanning several orders of magnitude. For this reason, these numbers were logarithmically adjusted by following the methodology proposed by **?**. This adjustment is made because several regression models are impaired when predicting data that spans several orders of magnitude.

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

4.1.4 Dataset Characterization

We now present the distribution of the extracted success features from movies through decades. Figure **??**(a) shows the histogram of movie productions throughout the decades. It clearly shows the number of movie productions exploded in the last decades (note that its y-axis is in log-scale). We also observe a few changes in the historical distributions of movie's success parameters.

There have been more movies with inferior box office in recent years (Figure **??**(b)). We argue that, in the past, movie productions were more expensive to fund and few people could afford a commercial movie production. Now, it is cheaper to produce a movie and to sell it to a specific (niche) audience. Such movies can accumulate a smaller gross but still be able to cover its production budget.

In recent decades, movies have more votes (Figure **??**(c)). Also regarding movie's ratings, Figure **??**(d) shows clearly that older movies have significantly higher average ratings than recent movies. Ratings are also more heterogeneous: when compared to the mean value, their ratings are more dispersed than in previous decades.

We argue that this phenomenon is not due strictly to movies produced in the past having a higher quality, but also to selection bias. People either choose a new movie to watch or an old classic with good references: there is little (or no) motivation to watch unimportant or unsuccessful movies from the past. It is expected that only good movies from the past are watched and voted on today. Furthermore, the IMDb website started collecting ratings from movies in the nineties. Movies released before that have ratings that are subject to this selection bias.

This selection bias also happens because people are constantly motivated by heavy ad campaigns to watch newer movies, even if they are unsure whether the movie is worth watching.

4.1.5 Movie Grouping by Performance

For better understanding and analyzing the different degrees of movie success, we scale the normalized values from the three success parameters (number of votes, gross income and average rating) in the $[0, 1]$ interval. Subsequently, we calculate the simple mean of (the scaled values of) their three success parameters. Then, we take the resulting value and divide the movies that were considered for the experiment by breaking the distribution into 10 slices (each slice contains movies whose average mean success falls into one specific range[8]). For example, slice 0–10% contains the top 10% most successful

[8]We reaffirm that only movies that pass all constraints defined in Section **??** take part in this analysis.

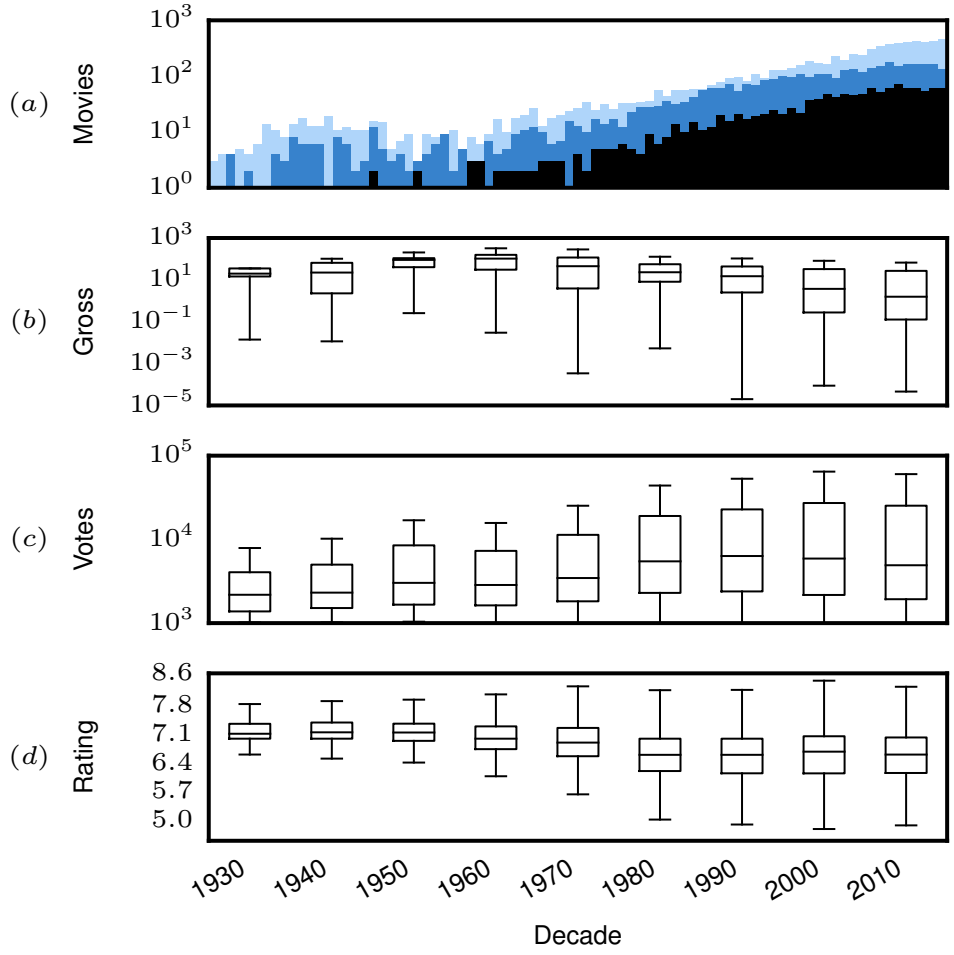

Figure 4.3: Movie productions and their success parameters through the decades.

movies, while slice 10–20% contains movies whose success parameter values fall between the 10^{th} and 20^{th} percentile of the success distribution. Table **??** shows the number of movies and the total number of movies per slice as well. Overall, we aggregate the slices into three major groups: group G_1 with blockbusters, group G_2 with movies of moderate success, and group G_3 with movies of low success.

Going back to Figure **??**, it shows the distribution of the number of movies with respect to the success parameters after normalization. Their distributions are drastically different among the tree groups, specially regarding gross (Figure **??**(a)) and number of votes (Figure **??**(b)). In Chapter **??**, we further discuss the different characteristics among these groups.

Table 4.1: Grouped distribution of success.

Slice	# of Movies	Mean Success	Mean Rating	Mean Gross	Mean Votes	Group
0–10%	1225	0.764	0.738	0.865	0.690	G_1
10–20%	1225	0.665	0.663	0.808	0.523	
20–30%	1225	0.607	0.647	0.766	0.409	G_2
30–40%	1225	0.563	0.638	0.735	0.315	
40–50%	1225	0.529	0.641	0.704	0.243	
50–60%	1225	0.506	0.644	0.687	0.186	
60–70%	1225	0.485	0.649	0.676	0.131	
70–80%	1225	0.468	0.638	0.665	0.100	G_3
80–90%	1225	0.447	0.610	0.628	0.102	
90–100%	1225	0.394	0.599	0.482	0.101	

4.2 Defining Features for Movies

This section presents the features extracted from each movie, which can be divided into two main groups: features related to characteristics from the movie producing team and features related to the movie itself. Table **??** presents an overview of all features.

4.2.1 From Movie Characteristics

We include 32 features that are unrelated to social characteristics from the movie production team composition: the movie's genres, production countries, runtime length, and the global state of the interaction graph model (Section **??**) at the time of the movie's release. These characteristics have also been previously explored to predict movie success [**?**] and influence [**?**]. Next, we describe the methodology for extracting these features.

Movie genre is categorical data: the IMDb dataset may assign one or more genres to movies. Therefore, we used 21 binary features to represent this categorical data, each one representing whether the movie belongs to a specific genre.

The IMDb dataset also provides a list of countries associated with the production of a movie. Since there are about 200 different countries, it is not efficient to encode country production into features directly. Therefore, the UN list for countries and continents was used to transform the movie's list of related countries into this list of related "geographic regions": Africa, Asia, Europe, Latin America, North America and Oceania. Again, binary features were used, each encoding whether any country of the

Table 4.2: All 121 features considered in the experiment.

(a) **Features obtained via aggregation**

Feature Type	Metric Name	Number of Aggregations	Number of Features
Ego Metrics	Closeness	7	70
	Betweenness	7	
	Clustering	7	
	Square clustering	7	
	Degree	7	
	Net. constraint	7	
	Prev. rating	7	
	Prev. gross	7	
	Prev. votes	7	
	Prev. experience	7	
Pairwise Metrics	Past experience	6	18
	Shared friends	6	
	Neighbour overlap	6	

(b) **Features obtained via One-Hot Encoding**

Feature Type	Characteristic	Number of Groups	Number of Features
Movie Characteristics	Genres	21	27
	Continents	6	

(c) **Simple Features**

Feature Type	Feature Name	Number of Features
Global Metrics	Global clustering coefficient	3
	Average path length	
	Small world coefficient	
Movie Characteristics	Runtime in minutes	3
	Production team size	
	Production budget (normalized)	

given continent took part in producing the movie. We note that movies can be jointly produced by teams in different countries, or by teams residing in a single country.

Runtime length and production team size are directly represented as integer features. Runtimes range from 40 minutes to 450 minutes[9]. Production team sizes range

[9]"Sátántangó" by Béla Tarr (1994) has a runtime of 432 minutes.

from 2 to 19[10].

Finally, we calculate topological characteristics concerning the state of the whole network (following **?**): its clustering coefficient, the average path length, and its small world coefficient.

4.2.2 From Movie Production Teams

Here, we present features derived from the set of social characteristics and past history from movie producing teams, which are based on the aspects presented in Table **??**. In addition to each producer's track record, the table contains well known topological metrics regarding social network analysis, which capture different social aspects of nodes in a social network graph.

We consider six topological ego-metrics (closeness, betweenness, clustering, square clustering, degree, and network constraint) and two pairwise-metrics (shared friends and neighbor overlap). These metrics are defined for singular nodes or pairs of nodes, and there is no standard methodology for extracting such topological information from sets of two or more nodes that constitute a team. To address this problem, we aggregate topological information from nodes in teams in two different ways: (*i*) we aggregate values from metrics calculated from single nodes (or pairs of nodes) in the team, and (*ii*) we temporarily contract all nodes from the team into a single node (using node contraction, Chapter **??**) and calculate metrics in the contracted node[11].

In the first approach, aggregating the different metric values in a single way (e.g., considering only the simple mean of the values) is prone to information loss, since many distinct distributions can have the same mean value. Therefore, in addition to the mean, five other statistical aggregation functions are used. This way, more information about the distribution of the values is preserved. The seven methods for statistical and non-statistical aggregation are briefly presented in Table **??**.

Moreover, we consider three global metrics to get the network state at time of movie release: the global clustering coefficient, the average path length and the 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).

[10]Production team of "Knuckleball!" (2012).

[11]Node contraction cannot be used with pair-wise metrics, since it aggregates all information into a single node

Table 4.3: Summary of all metrics used for team characterization.

(a) **Ego metrics**

Closeness	Captures how close a producer p_k is from all other producers reachable from it in $\mathbb{G}_{\mathbb{P}}$.
Betweenness	Fraction of all shortest paths, computed using breadth-first search, connecting pairs of producers that pass through a particular producer p_k.
Clustering	Fraction of pairs of producer p_k collaborators who have also collaborated with one another.
Square clustering	Fraction of possible cycles of size 4 that exist at the node [**?**].
Network constrain	Index measuring the extent of the bridiging of the node (whether the node connects different clusters [**?**]).
Degree	Total number of partners of a producer p_k.
Past experience	Number of prior movies produced by the node.
Previous success	Mean success metrics for prior movies produced by the node, defined for all three success parameters extracted from movies.

(b) **Pair-wise metrics**

Shared friends	Number of nodes connected to both nodes in the pair.
Neighbour overlap	Rate of shared friends and total nodes connected to the pair.
Shared experience	Number of prior movies jointly produced by the pair of nodes.

(c) **Global team and graph metrics**

Global clustering	Measure of degree to which the producers in $\mathbb{G}_{\mathbb{P}}$ tend to cluster together.
Average shortest path	Average across all values of shortest path for all pairs of producers in $\mathbb{G}_{\mathbb{P}}$.
Team Size	The total size of full production team.

The node's track record is given by four measures related to its past success and experience (average previous rating, gross, votes and number of previously produced movies) and one pairwise metric (number of movies previously jointly produced by the pair). The former may be grouped by using the seven aforementioned aggregation

Table 4.4: Summary of aggregation techniques used to generate features.

Aggregator	**Description**
Arithmetic mean	Indicates the central tendency or typical value of a set of numbers by using the sum of their values.
Harmonic mean	Indicates the central tendency of a set of rates.
Median	Number separating the higher half of the team producers metrics from the lower half.
Minimum, Maximum	Lowest and highest values of the team producers metrics.
Standard Deviation	Amount of variation or dispersion of the team producers metrics.
Node Contraction	Combines multiple nodes in a graph into one, so aggregate information can be retrieved from the super node.

techniques, yielding 28 features. The latter may be grouped by the six mathematical aggregation functions, generating six more features. At the end, this part considers 34 features.

With these characteristics, we capture a complete composition of the team into numerical and analytical features, such as:

- whether they are composed of amateurs, experienced producers, or mixed, by looking at the teams' past experience metric, aggregate in terms of mean, minimum, maximum and standard deviation;
- whether there is a producer who has had a spectacular track record in the team, by looking at the teams' previous success metric, aggregated in terms of mean and harmonic mean;
- whether the team has a central position in the whole network, by looking at team's betweenness metric, aggregated using the team's *supernode*;
- whether they are strongly or weakly connected to other producers in the industry, by looking at the degree and network constraint metrics, aggregated by team's *supernode*.

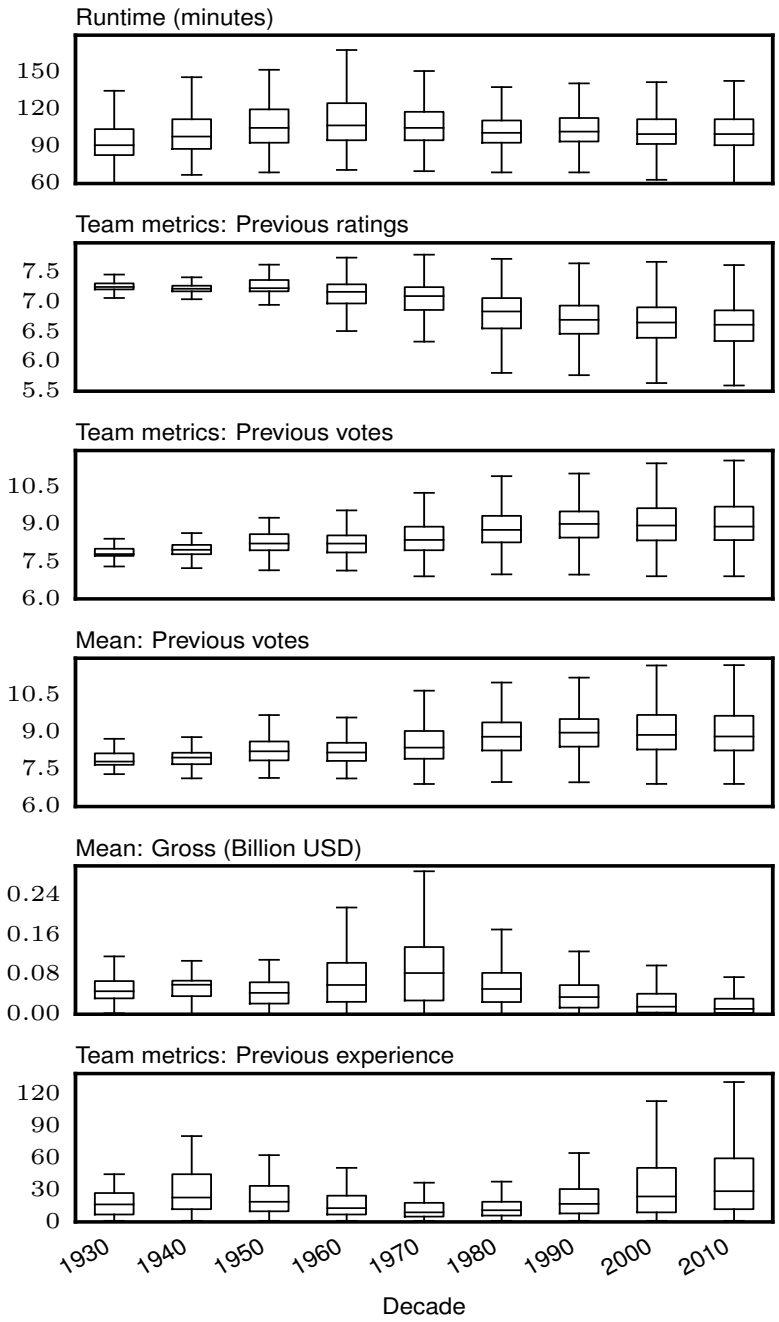

Figure 4.4: Distribution of non-topological features throughout the years.

4.3 Feature Characterization

We present distributions from the selected features across the decades in Figures **??** and **??**. Features have very different distributions throughout the years, indicating that the movie producing network is dynamic and constantly evolving, albeit in

a slow pace[12]

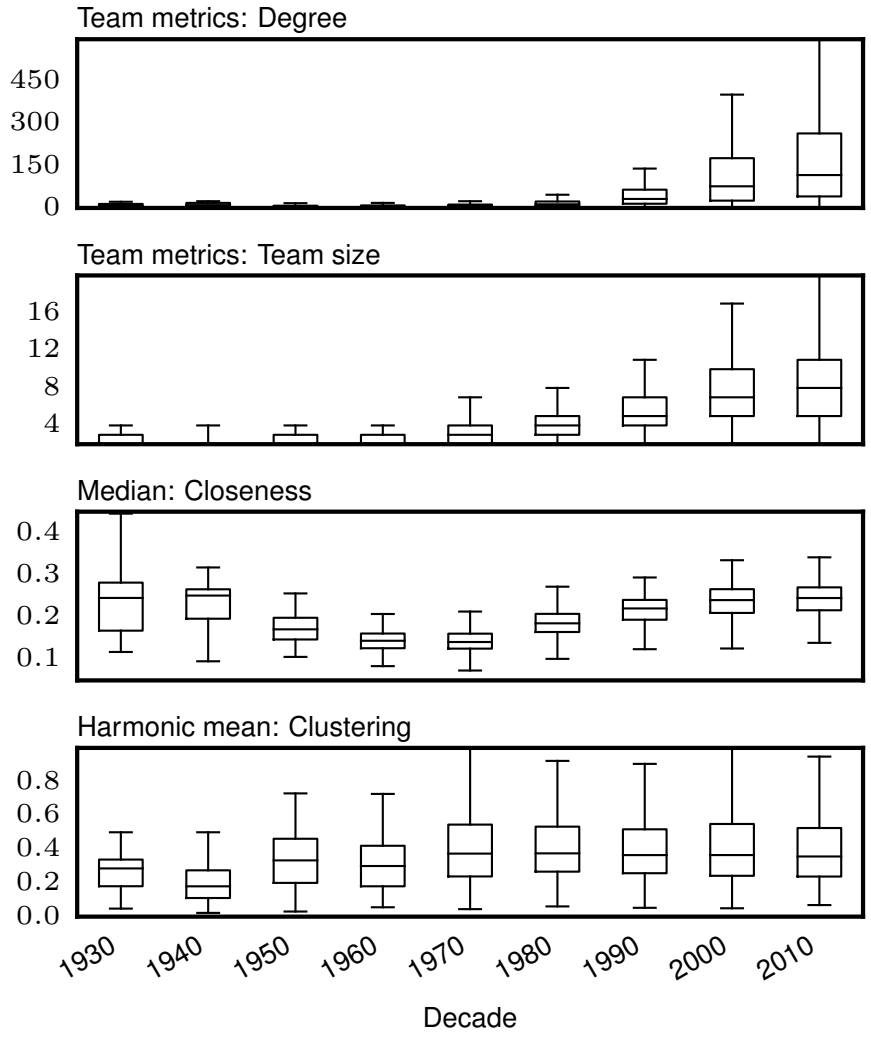

Figure 4.5: Distribution of topological features throughout the years.

We can see from these characterizations that different performance groups have different distributions for the features being extracted. This is an important indicator that the extracted features can indeed provide information on movie performance.

Analyzing the features and their distributions also yields interesting insights into the network. Figure **??** shows the evolution of the small world coefficient from the whole graph thorough the years. We can see a clear disruption in the organization of the network in the 40s and 50s, that was probably caused by World War II. The efforts

[12]Figures **??** and **??** (presented in detail in Chapter **??**) present the distribution of selected features in the whole set of data, but in the form of color-coded histogram bars according to the movies performance group.

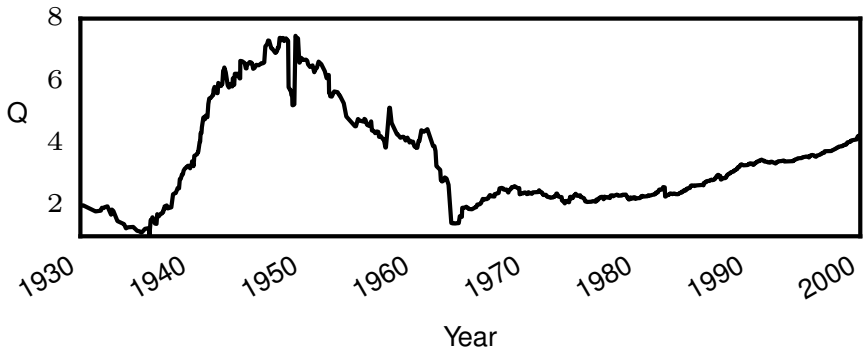

Figure 4.6: Evolution of the small world coefficient for the IMDb network cut considered in this research.

on the war and infrastructural recovery that followed it could have captured resources that were previously available for movie production. Also, the war destruction could have taken away many filming sets, equipments, and even lives of actors and producers.

We observe that the post-war rise in the network's small world coefficient was not as intense as in pre-war years. We hypothesise that this reflects a greater division of the world, as movies producers from opposing cold war countries would not be able to cooperate. Also, during the Cold War, producers and directors (mostly in the USA) were persecuted as communists, that lead to a decrease in production. Moreover, in recent years the cost of movie production went down, what could have resulted in a greater pool of movie producers spread around the world, leading to the existence of more teams that are not closely knit in a central core.

4.4 Feature Interaction

The interaction between features must also be considered: for instance, even if a feature is seemingly unrelated to the target variable, when seen in combination with other features, patterns may emerge. In order to study the interaction among features, we created movie scatter plots where dot colors are set according to movie performance groups, and axes represent features' normalized values.

Specifically, Figure **??** shows the formation of colored clusters. The colors encode team's performance, which indicates that certain combination of features can be used to help identifying team's performance. To take advantage of this phenomena, we also included the product of every feature pair as new features. There are dozens of possible pairs such that analysing every pair individually is not practical. Machine Learning

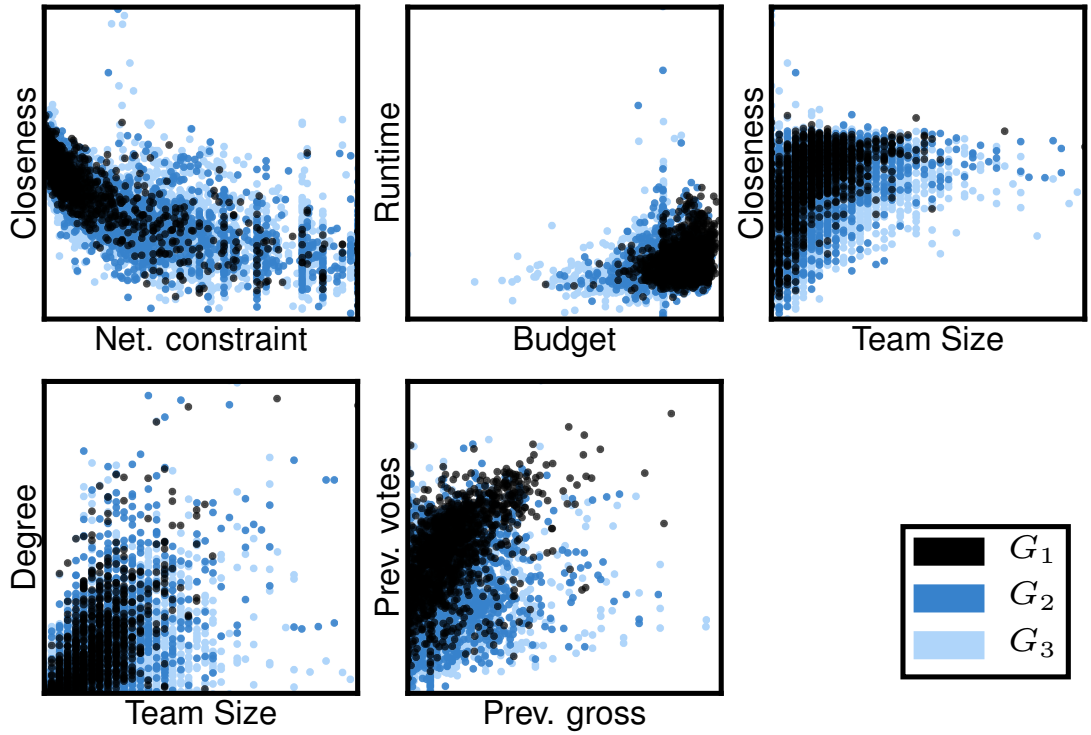

Figure 4.7: Joint scatter plot of movies per features. The shade of a point represents its performance group. Dark clusters form in the graph, meaning iteration between features should be considered.

techniques are employed to automatically identify pairs that add the most relevant information. Figure **??** presents some of the relevant pairs for illustrative purposes.

There are two ways to calculate feature interaction: normalizing the features to the [0–1] interval (using a simple linear interpolation) and then calculating their product, or doing the same in reverse order. From experimental evaluation, both methods produce similar results, but normalizing before calculating the product performs slighting better. We assume this happens because multiplying variables of the same magnitude mimics better the way in which variables interact.

4.5 Movie Success Prediction Model

In this section, we present the regression model chosen for experimentation and provide a general overview of the proposed prediction model.

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 with at least one producer

connected to the giant component (as defined in Section **??**). 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 in such way 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.

Overall, we consider a set of 121 features spanned over three main aspects of movies: characteristics of a motion picture itself, topological characteristics from its production team, and the team's past success and experience (Section **??**). As for **?**, our model only uses features available *before* the movie release and thus may be used to make predictions (see Algorithm **??** in Chapter **??**)

We rely on a multivariate Bayesian Ridge regression model to predict each movie success parameters described in Section **??**. This is a linear regression model that uses Bayesian Inference and is similar to the Ridge regression model. This regression model is well described in the literature (Chapter **??**). It was chosen because it better handles features with lots of noise (i.e., considerable unexplained variation is observed when modeling the data). Two other regression algorithms were also tried: the Support Vector Regression (SVR) and the Ordinary Least Squares (OLS, Section **??**). However, comparing the resulting coefficients of determination (R^2) shows the Bayesian Ridge regression model outperforms them both.

Most linear regression techniques, including the one used in this study, work best when their input features and target features are normalized and scaled to the same magnitude. Therefore, all of our features are normalized, as well as our target variables. There are two major techniques to achieve this. The first is scaling the values to fit within the [0, 1] interval. The second is to transform the data so that its mean equals zero and variance equals one. Both normalization techniques were attempted, and zero-one scaling provided better results.

Chapter 5

Success Prediction Model Analysis

In this chapter, we present our methodology and experimental results related to the prediction model presented in Chapter **??**. We begin with the methodology in Section **??**, which presents our approach for assessing the predictive model's accuracy according to different types of input features. Then, in Section **??**, we present our approach for analyzing all 121 features from Table **??** in order to select the ones that show a *greater dependence*[1] with movie success. Finally, in Section **??**, we present and discuss results, showing they confirm the main hypothesis of this work, which is: *In a multivariate predictive analysis of movie success, using many topological features from teams leads to a more accurate analysis.*

5.1 Methodology

Choosing a single split of train/test sets might produce poor experimental results, because the train and test sets could be selected in a biased way [**?**]. To avoid this problem, we perform *K-fold Cross Validation* with $K = 5$. The selection of K was made such that it effectively controls the bias while it generates large folds that are more heterogeneous. 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 coefficient of determination[2] is computed by evaluating the model using the respective test set. This gives five R^2 values for each run. We consider the mean of these values as the result of the cross-validation execution.

[1] Dependence is the statistical relationship between two sets of data.

[2] Throughout this chapter, the coefficient of determination (denoted as R^2) is used to assess how well the different regression models are predicting movie success according to their given inputs.

For each evaluation, it is important to ensure that the special five folds were not randomized in a biased way. Hence, when evaluating a model, we take results from 30 different *5-fold cross validation* executions. We then calculate the mean and confidence intervals ($\alpha = 95\%$) from such runs as the model's final score.

If the distribution of votes and gross are highly skewed, forming a power law (few movies with most votes and gross, see Figure **??**), there is a high probability that either no or an insufficient amount of high-performing movies will be present in several randomized folds. Such skewness might favor the formation of biased train/test sets. This could impair the model's ability to recognize higher performing movies and thus compromise its accuracy.

To account for this phenomenon, we attempted under-sampling the number of movies with lower levels of success in the training sets. However, following experimental evaluation showed this does *not* improve regression accuracy. Hence, we choose to generate *balanced* train/test folds. Using *balanced* folds, train and test sets of movies are grouped in a way that ensures groups have similar distributions of movies among each of the three performance groups (G_1, G_2 and G_3 as defined in Table **??**). Moreover, all train/test folds are still generated in a random fashion.

Finally, Algorithm **??** provides an overview of all major steps and processes involved in the experiments. Briefly, it shows all steps starting from the raw data all the way to a trained movie success prediction model and its performance.

5.1.1 Predictive Evaluation

The methodology for evaluating the prediction model consists in splitting 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%). We use each group for training and testing the regression models for each of the three success parameters. Table **??** summarizes the three different experiment sets.

To evaluate how the features that depend on team social characteristics add predictive power to the model, experimental trials are performed with the same set of movies $\mathbb{M}$, but using different sets of features. Specifically, we evaluate the regression model three times by considering: only the non-topological features, only the topological features and all features. Finally, we use the coefficient of determination R^2 to assess how well the regression models are predicting movie success parameters.

Algorithm 1: Movie Prediction Task

```
input  : Raw IMDb Movie Data
output: Movie success prediction model
𝔾 ← [ ] ;                                   /* producers' graph */
movie_map ← [ ] ;                          /* movies with features */
filter_by_votes(𝕄, minimum_votes);
sort(𝕄, release_date);
foreach m ∈ 𝕄 do
    𝔾.update(ℙm) ;                          /* add movie producers */
    if any(ℙm ∩ 𝔾.giant_component) then
        x ← m.success_parameters;
        f ← calculate_features(m);
        movie_map.add(m, x, f);
    end
end
normalize_features(movie_map);
add_product_of_features(movie_map);
models ← [ ] ;                              /* regression models */
for i ← 1 to 30 do                                    /* 30 runs */
    R2,model ← ramdom_cross_validation(movie_map);
     models.add([R2, model]);
end
return {models.mean, models.confidenceInterval};
```

Table 5.1: Three test sets for evaluating prediction accuracy.

Year Range	Number of Movies	Proportion of complete set
2008–2013	3,317	27%
1995–2013	9,775	52%
1930–2013	12,250	100%

5.2 Feature Selection

Selecting only the most relevant features for the regression model is crucial because it reduces input noise [**?**], which in turn increases the prediction accuracy and reduces overfitting. To do so, we visually inspect and compare feature distribution of different movies performance groups (G_1, G_2, G_3) and use statistical dependency metrics[3].

First, each feature's correlation with all success parameters (economic success given by gross, public acceptance given by ratings and movie popularity given by number of votes) were calculated using the Pearson correlation coefficient [**?**]. The

[3]Statistical dependency metrics are measures that can be calculated from sets of data in order to access information about their dependence.

Table 5.2: Correlation for different features considering all aggregation metrics and success parameters (V: votes, R: ratings, G: gross).

(a) **Features generated from square clustering**

Aggregator	Distance Corr.			Spearman			Pearson		
	V	R	G	V	R	G	V	R	G
Std. Deviation	.24	.08	.16	−.23	.06	−.19	−.23	.07	−.14
Contraction	.17	.10	.08	−.33	.13	−.13	−.13	.06	−.06
Maximum	.26	.10	.16	−.34	.13	−.20	−.26	.09	−.15
Minimum	.23	.14	.07	−.37	.19	−.10	−.15	.08	−.04
Mean	.27	.10	.17	−.38	.15	−.23	−.26	.09	−.16
Median	.26	.10	.17	−.40	.15	−.20	−.25	.08	−.16
Harmonic Mean	.26	.13	.12	−.40	.16	−.23	−.20	.10	−.08

(b) **Features generated from neighbor overlap**

Aggregator	Distance Corr.			Spearman			Pearson		
	V	R	G	V	R	G	V	R	G
Std. Deviation	.09	.13	.16	.04	−.12	−.17	.06	−.12	−.15
Maximum	.11	.13	.13	.12	−.14	−.14	.12	−.13	−.13
Midrange	.09	.10	.16	.02	−.08	−.18	.05	−.80	−.16
Minimum	.21	.09	.15	−.18	.06	−.14	−.18	.07	−.13
Mean	.19	.06	.25	−.15	.00	−.25	−.15	.00	−.24
Median	.19	.05	.23	−.13	.01	−.21	−.17	.02	−.23
Harmonic Mean	.24	.09	.22	−.24	.09	−.20	−.24	.08	−.21

correlation levels were *low*, with absolute values ranging between 0 and 0.57 for all features. Then, visual inspection of the scatter plots confirmed that there is no evident *linear* relationship among all features and the success parameters.

Therefore, we performed a second analysis with: (*i*) the *distance correlation*[4] to detect the strength of non-linear dependence among variables, and (*ii*) the *Spearman correlation*, for monotonic relationships [**?**]. Table **??** shows the results considering only features derived from *square clustering* and *neighbor overlap*, in which many features present high degrees of dependence (other metrics provided similar results). Pearson coefficients were very similar to Spearman coefficients, but they were also included in the tables for completeness. Negative numbers for Spearman indicates inversely related variables.

Based on this preliminary analysis, the features with lowest levels of dependency

[4]Distance correlation is a measure of statistical dependence between two variables. This measure is zero if and only if the variables are statistically independent [**?**].

with respect to the three success parameter were filtered out. Cross-validation trials showed that removing such features actually *improved* model accuracy. Features that did not cause any performance loss were iteratively removed, until no further feature could be removed. At the end of this procedure, only **23 features remained**. Of those features, 19 are non-topological: nine genres (romance, comedy, horror, adventure, thriller, mystery, drama, action, and documentary), three continents (North America, Europe, Africa), duration time, 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. The remaining four are topological: degree (aggregated by node contraction), team size, closeness (aggregated by median) and clustering (aggregated by harmonic mean). Note that although only 4 topological features were selected for the prediction task, a total of 57 topological features are calculated for each movie (see Table **??**).

Further adding any feature to this set, such as more binary features for genres and continents (or any of the others feature from Table **??**), either makes no noticeable effect or impairs the model's accuracy. An overview of all 23 selected features is presented Table **??**.

Further confirming this analysis, a visual representation of how informative a feature is with respect to movie performance is possible by plotting the feature's distributions next to the three different movie performance groups. The more different the shapes and values of the distributions, the more they provide information that can be used to predict performance. Therefore, Figures **??** and **??** respectively show different distributions for informative topological and non-topological features. Both figures confirm the presented features provide useful information.

Table 5.3: Set of features selected to be used in the regression model. The topological features are listed last.

Based on	Feature Type	# of features	Description
Genre	Binary	9	One for each of the genres: "Romance", "Comedy", "Horror", "Adventure", "Thriller", "Mistery", "Drama", "Action" and "Documentary".
Continent	Binary	3	Binary features for each of the continents: Africa, Europe, North America.
Runtime	Integer	1	Movie's runtime in minutes.
Budget	Log Transform	1	The logarithm of the movie's budget adjusted adjusted to present US Dollars.
Previous success	Ego-metric aggregation: node contraction	2	Mean ratings and mean votes for movies previously produced by one or more agents.
Previous success	Ego-metric aggregation: mean	2	Mean of node's past productions gross and number of votes.
Previous experience	Ego-metric aggregation: node contraction	1	Number of production previously made by one or more agents from the team.
	Topological Features		
Degree	Ego-metric aggregation: node contraction	1	Number of distinct peers from nodes, disregarding nodes in the team being analyzed.
team being analyzed.			
Team Size	Simple Feature	1	Number of nodes in the current team.
Closeness	Ego-metric aggregation: median	1	Median of closeness metric for nodes in the team.
Clustering	Ego-metric aggregation: harmonic mean	1	Harmonic mean of clustering coefficient for nodes in the team.

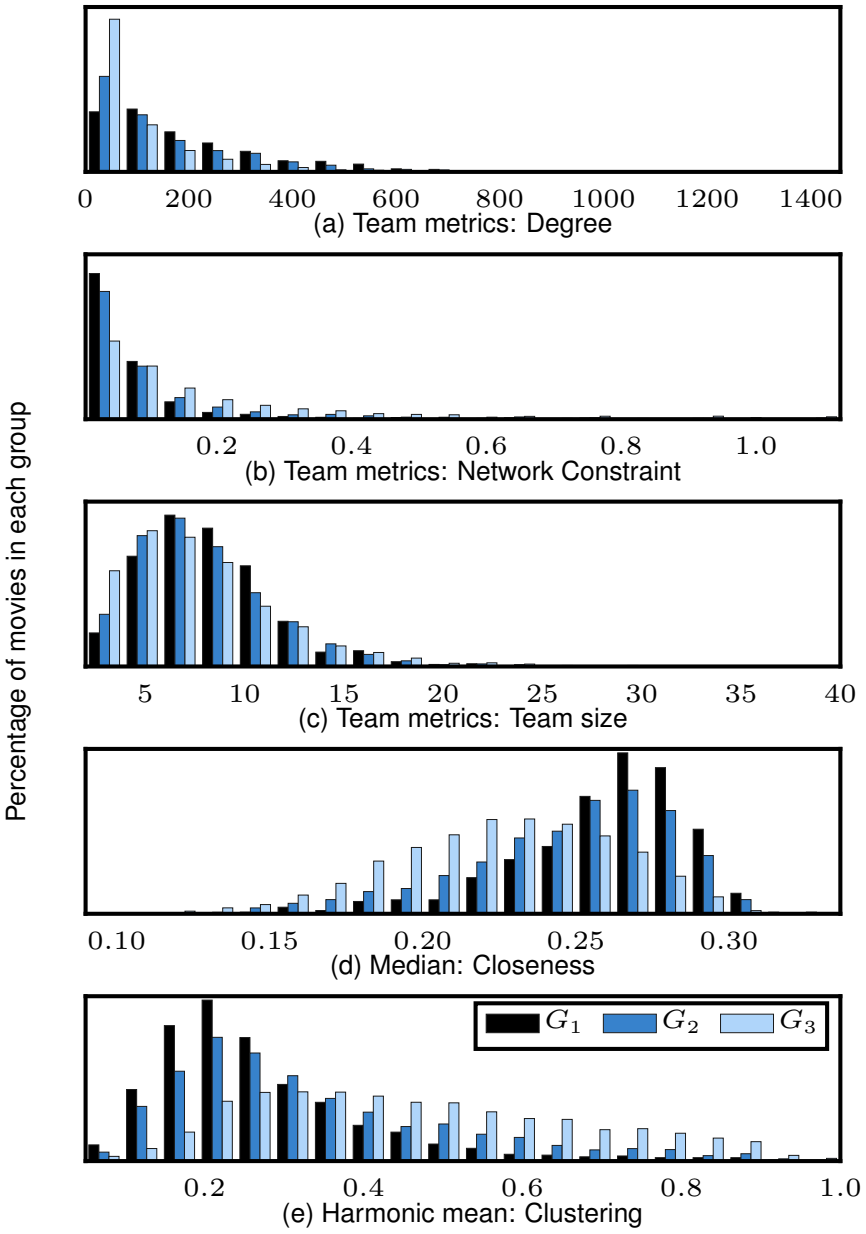

Figure 5.1: Distribution of topological features for movies in the three different performance groups. The x axis represents the portion of movies from the given group (encoded by color) having the specified value.

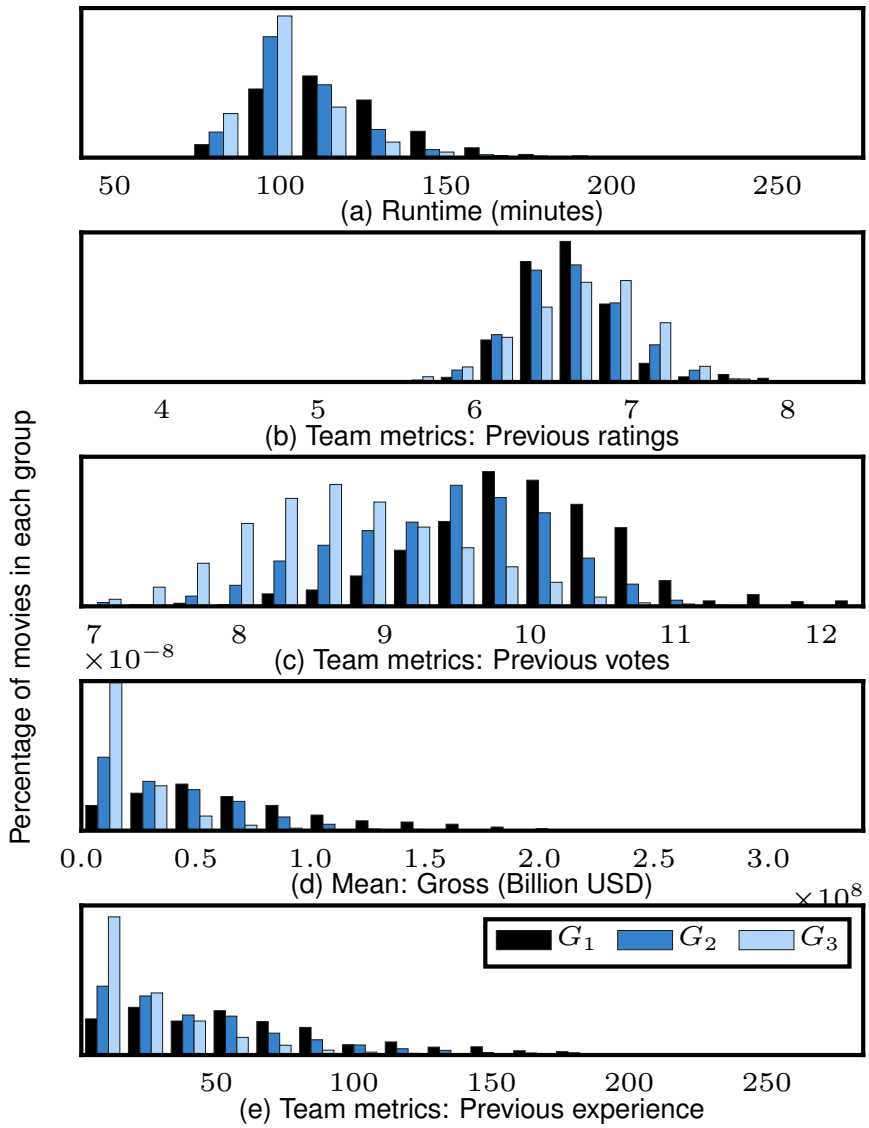

Figure 5.2: Distribution of non topological features for movies in the three different performance groups. The x axis represents the portion of movies from the given group (encoded by color) having the specified value.

5.2.1 Producing Successful Movies

In this section, we assume new teams showing topological characteristics previously associated to successful movies have a higher probability of producing successful movies. Under this assumption, we present how our findings can help to chose agents for a new movie producing team, such that it has improved success odds.

Figures **??** and **??** show movies that perform better are well determined by their

features ranges. We believe 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 **??**, movies from the G_1 group (blockbusters) are more likely to have a low clustering coefficient (0–0.3 interval) when compared to G_2 and G_3. Moreover, successful teams are likely to have degree larger than 100, network constraints smaller than 0.1, team sizes of 6–10 people, and closeness greater than 0.25.

We observe that teams with a combined past experience 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. We also see that the mean rating from movies produced by team members should not deviate much from the global mean (6–7). Ideally, one should pick producers whose previous movies received in average about $8,000$ votes.

5.3 Results

In this chapter, we present an experimental evaluation for our prediction model. First, the prediction model is evaluated with three different groups of features (Section **??**). The most and least successful movies are presented along with their team's key topological features (Section **??**). We then present the movies of different levels of success for which the predicted values had greatest and smallest prediction errors (Section **??**). Finally, we conclude this section by taking all results together and giving insights into ways of producing a blockbuster (Section **??**).

5.3.1 Prediction Model Evaluation

Here, we evaluate the three instances of regression models for predicting each of the three success parameters based on each chronological test set (Table **??**). Table **??** presents R^2 measures for each model instance. The model that only uses non-topological features is taken as baseline because it only uses features already explored by previous works (e.g., genre, runtime, and budget).

The results show 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 evolving. Hence, forms of organization that were related to best results in one chronological context might not generalize across all history. As an evidence, Figure **??** showed a rise of the clustering coefficient until late 1940's, then a decrease until the mid-1960's, with a very slow growth until

Table 5.4: 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	Gain
Votes	2008–2013	.529, ±.0008	.310, ±.0006	.556, ±.0008	5.10%
	2000–2013	.484, ±.0004	.294, ±.0005	.517, ±.0004	6.82%
	1990–2013	.437, ±.0003	.246, ±.0004	.464, ±.0003	6.18%
Gross	2008–2013	.431, ±.0008	.170, ±.0013	.448, ±.0009	3.94%
	2000–2013	.419, ±.0004	.175, ±.0005	.447, ±.0004	6.68%
	1990–2013	.392, ±.0004	.174, ±.0004	.435, ±.0003	10.97%
Rating	2008–2013	.271, ±.0011	.033, ±.0009	.281, ±.0012	3.69%
	2000–2013	.267, ±.0006	.038, ±.0003	.273, ±.0006	3.37%
	1990–2013	.258, ±.0004	.031, ±.0003	.262, ±.0005	1.55%

2013 (Section **??**). This result strengthens the importance of splitting the data into different profiles and analysing then separately.

Also, the model performs very differently when predicting each success parameter: it performs better for number of votes, a little worse for gross (R^2 values are 19.42% smaller than the best model's) and worse still for normalized ratings (R^2 values are 49.46% smaller than the best model's). Here, we argue that the absolute number of votes a movie receives is the less noisy variable out of the three. In other words, more external factors interfere with gross and ratings. For instance, gross is directly connected to the audience that *pays* to watch the movie in theaters shortly after the release. That can be affected by the movie's distribution efficiency and advertising reach (and the spread of movie piracy). Ratings way be heavily affected by the performance of a specific actor and the cultural/emotional response from the audience. The absolute number of votes measures the public *attention* captured by the movie, since a person may watch and rate it without necessarily having to pay for it or like it. Finally, such a broad range of results highlight (once more) that movie success prediction is a difficult task. It is especially hard when assessing gross and ratings due to the morphing complexity of production team organization in the movie industry.

Nonetheless, for all cases, the prediction model that considers topological features (along with past success, past experience and movie characteristics) outperforms the others. This makes clear topological features do have predictive power over movie success parameters. Indeed, we see that topological features *alone* could still be used to train movie success regression models with better results than pure guesswork.

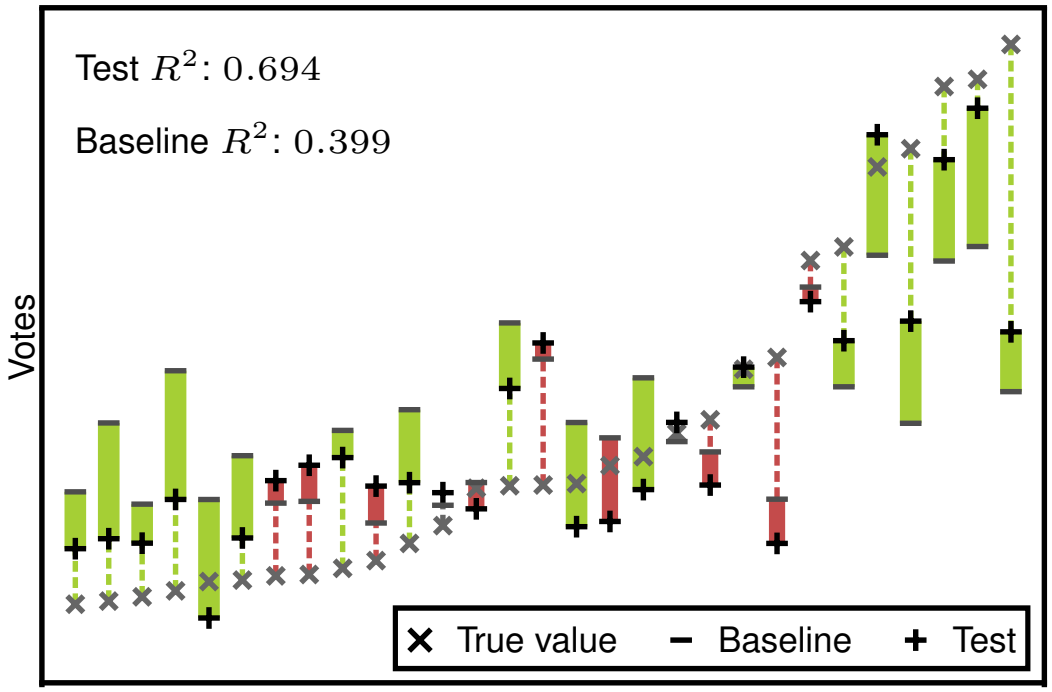

Figure 5.3: Performance comparison of regression models without (baseline) and with (test) topological features, using the same movie train/test split. Green bars for when test is better than baseline, and red bars otherwise.

5.3.1.1 Topological Features Influence

In order to complement our discussion, Figure **??** is a graphic comparison between using only non-topological features (baseline) and using all features (test). This evaluation considers a random sample of 29 movies[5]. Both models were trained and tested with the chronological dataset [2008–2013]–which produced the best results (Section **??**). The y axis is scaled to the $[0, 1]$ interval. In order to improve readability, movies in the test set are ordered by their predicted variable magnitude (ground truth), i.e. the movie's actual $\log(\textit{number of votes})$.

This extra analysis emphasizes the relevance of considering network topology. In fact, we may claim that it ushers the development of more sophisticated prediction models in which the *topology* of teams is also considered. Moreover, new tools could unveil new knowledge on how precisely the network topology impacts team success in broader contexts.

[5]This is just an **illustrative** sample of how our predictor compares to the baseline. The sample has only 29 movies for clarity reasons.

Table 5.5: Coefficient of determination (R^2) and confidence intervals for a significance level of 95% obtained with a Bayesian Ridge regression. It shows the best results obtained only using pré-release information side by side with results obtained when also using other of the movie's success parameters as new features.

Target	Years	Without extra features	With extra features
Votes	2008–2013	.556, ±.0008	.670, ±.0007
	2000–2013	.517, ±.0004	.687, ±.0004
	1990–2013	.464, ±.0003	.669, ±.0004
Gross	2008–2013	.448, ±.0009	.574, ±.0009
	2000–2013	.447, ±.0004	.613, ±.0006
	1990–2013	.435, ±.0003	.614, ±.0005
Rating	2008–2013	.281, ±.0012	.438, ±.0014
	2000–2013	.273, ±.0006	.479, ±.0007
	1990–2013	.262, ±.0005	.480, ±.0006

5.3.1.2 "Super"-Regressor

In this final evaluation, we study how including success parameters into the set of 23 features (Table **??**) can improve prediction accuracy. Specifically, given one success parameter, we include the other two in the feature set. For instance, this would be helpful when having two parameters already computed (even if partially) and trying to predict the third one. Considering that we want to predict the 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 $\alpha = 95\%$. Using the same strategy, we can predict gross with $R^2 = 0.573$, confidence interval = 0.0013, and $\alpha = 95\%$. For ratings, the results are $R^2 = 0.356$, confidence interval = 0.0013, and $\alpha = 95\%$. A side by side comparison of the best performing model against the one with these extra features is shown Table **??**. Using just one extra feature yields results that are worse when two extra features are used. However, even when just one extra feature is used, the results are still better when compared to the best perming model that does not use pre-release information. Finally, these results hint an even better prediction model might be obtained for long-term success of movies considering their initial success.

5.3.2 Best and Worst Movies

For better illustrating the analysis, we present the best and worst movies according to the normalized mean of the success factors. Tables **??** and **??** respectively show the movies with the highest and lowest scores, including examples of low-scoring popular

Table 5.6: Worst movies by normalized average of ratings, gross revenue and number of votes. The last two are low scoring movies that are also popular (received more than 6 thousand votes).

Movie Title	Score	Degree	Team Size	Clustering
Waxwork II: Lost in Time (1992)	.247	32	2	.296
Fuk sau che chi sei (2010)	.252	3	2	1.00
Nature Calls (2012)	.259	77	5	.322
Perro come perro (2008)	.264	17	7	.796
Playback (II) (2012)	.269	309	12	.332
Jack & Diane (2012)	.269	95	11	.419
Angels Crest (2011)	.291	71	5	.371
Che? (1972)	.299	15	4	.185
S. Darko (2009)	.306	109	7	.402
Captain America (1990)	.331	97	5	.200

Table 5.7: Best movies by normalized average of ratings, gross revenue and number of votes. The last three are high-scoring movies that are also not very popular (received less than 5 thousand votes).

Movie Title	Score	Degree	Team Size	Clustering
The Dark Knight (2008)	.978	343	10	.275
The Lord of the Rings (2003)	.952	77	5	.251
Inception (2010)	.949	17	7	.176
The Lord of the Rings (2001)	.946	309	12	.252
The Lord of the Rings (2002)	.941	95	11	.228
Star Wars (1977)	.94	71	5	.264
The Dark Knight Rises (2012)	.938	15	4	.176
Pulp Fiction (1994)	.936	5	10	.195
The Shawshank Redemption (1994)	.936	109	7	.241
The Matrix (1999)	.934	97	5	.218
Friendly Persuasion (1956)	.599	9	4	.180
They Drive by Night (1940)	.599	17	2	.263
Hubble 3D (2010)	.580	3	3	.208

movies and high-scoring unpopular movies.

These results confirm our findings: for bad movies, all values of team harmonic mean of clustering coefficient are below 0.280, whereas good movies obtained higher values. Degree and team size values for movies are comparable across the two tables. However, as seen in Sections **??** and **??**, these metrics play a very important part in the regression model if combined with other features. This might indicate that these features provide predictive power once combined with other features.

5.3.3 Prediction Hits and Misses

In this section, we dissect the prediction results by calculating the difference between predicted values and real values of success parameters for each movie. Specifically, we predict movie's success parameters 30 different times (using the same methodology previously presented), and store the predicted results. We use these predictions to compute the mean prediction error and confidence interval associated with success parameters for each movie.

By using these values, we focus our analysis on the movies in which our model predicts success parameters with highest and lowest errors, i.e., our prediction misses and hits. To provide a throughout overview of such movies, we present a separate table for displaying errors from prediction results in each of the three success parameters. Each table is also divided in two parts, one shows nine movies with high prediction error (misses), and the other shows nine movies with low prediction errors (hits). Also, in each table, rather than showing the movies having the overall highest and lowest errors, we select nine movies having success values from distinct ranges: 0–0.2, 0.2–0.3, 0.3–0.4 and so on until the range 0.9–1.0. Considering all movies that would fit the range, the movie having the highest (or lowest) prediction error is displayed on the table. In other words, the movies in these tables were selected by picking the highest (or lowest) prediction error for the movies having different ranges of the success parameter being observed in the table.

According to this organization, Tables **??**, **??** and **??** show the predictor's hits and misses for gross, number of votes and ratings. All values (but degree) are normalized for clarity. Each table is ordered by the level of success of the movie according to the table's success parameter (i.e., Table **??** ordered by gross, Table **??** ordered by vote, and Table **??** by rating), from the weakest to the strongest. Two topological characteristics are also highlighted: clustering and degree (which could explain some of the hit and misses, Section **??**).

We now take a closer look at some of those movies starting with *Erased* (2012) in Table **??**. It had an estimated budget of 12 million euros and was produced by an experienced team of producers. Such features lead our predictor to assign a high gross value to it. However, the movie gross extracted from IMDb database is very low (only 147 British pounds), resulting in a high prediction error. Hence, either it is a really unexpected outcome, or there is a typo in the gross data.

On the other hand, consider the movie *Hubble 3D* (2010), for which the model heavily under-predicted the gross income. Its producers had very little past experience or past success, and they were very poorly connected to other producers in the network.

Table 5.8: Errors in predictive results for gross.

(a) **Predictor Misses**

Title	Clustering	Degree	Error	Gross
Erased (2012)	.290	352	+.489	.148
Loosies (2011)	.451	232	+.416	.293
Leap Year (2010)	.118	574	+.518	.370
Ballast (2008)	.303	109	+.457	.475
In the Land of Blood (2011)	.252	204	+.324	.550
The Objective (2008)	.433	224	-.235	.668
You're Next (2011)	.446	43	-.304	.783
Hubble 3D (2010)	.857	3	-.366	.843
Slumdog Millionaire (2008)	.174	303	-.284	.904

(b) **Predictor Hits**

Title	Clustering	Degree	Error	Gross
Emergo (2011)	.836	5	+.298	.147
London River (2009)	.344	74	+.137	.241
The ABCs of Death (2012)	.718	149	-.003	.399
From Within (2008)	.854	6	-.002	.468
Cheung Gong 7 (2008)	.382	76	.000	.532
Cocktail (2012)	.447	62	.000	.620
Haywire (2011)	.096	562	.000	.785
Predators (2010)	.202	211	.000	.846
Ice Age: Dawn of (2009)	.797	7	+.001	.923

This indicated the movie would be a good candidate for a flop. However, it was one of the first feature-length IMAX productions, whose work was overseen by IMAX itself. With such innovative aspects, it drew masses to movie theaters.

Now, consider the movie *Ice Age: Dawn of the Dinosaurs* (2009) as one big hit. Even though it has a very high clustering coefficient, the model still predicted its gross almost to the point (error of +.001). Such result shows its team's past success (in gross) was high enough to counterbalance the team's high heterogeneity. Likewise, we note that among the hits, the highest errors are within the low performance group (G_3 with *Emergo* and *London River*).

Table **??** presents some interesting results while predicting movie popularity (number of votes). For instance, the movie *Intouchables* (2011) was produced in France by an unexperienced producing team whose past productions were unpopular or obscure movies. None of the producers had ever produced a single movie that had reached

Table 5.9: Errors in predictive results for votes.

(a) **Predictor Misses**

Title	Clustering	Degree	Error	Votes
The Incident (2011)	.152	161	+.358	.132
The Factory (2012)	.147	554	+.431	.254
Recep Ivedik (2008)	.540	6	-.297	.372
Black Dynamite (2009)	.929	10	-.393	.479
ParaNorman (2012)	.820	7	-.505	.570
Yip Man (2008)	.709	24	-.546	.663
Moon (2009)	.565	104	-.561	.742
Intouchables (2011)	.699	10	-.633	.813
Django Unchained (2012)	.079	984	-.115	.905

(b) **Predictor Hits**

Title	Clustering	Degree	Error	Votes
Buck (2011)	.511	68	.000	.119
Zweiohrküken (2009)	.312	82	.000	.221
Wrecked (2010)	.272	421	+.001	.337
Cyrus (2010)	.201	167	-.001	.460
Don't Be Afraid of (2010)	.169	348	.000	.502
The Last Airbender (2010)	.230	211	.000	.642
Fast Five (2011)	.190	222	-.003	.753
X-Men: First Class (2011)	.205	361	-.009	.833
The Dark Knight (2008)	.202	343	-.035	.996

such a mass success. However, this movie has a widely acclaimed high-quality storyline. So, even though it had a relatively low gross, barely covering its budget, the movie name spread after its release, reaching a wide level of popularity.

For the hits on this table, we note the high performance group has only movies from big franchises: *Fast Five*, *X-Men: First Class* and *The Dark Knight*. Given that franchises tend to keep their producing teams fairly large and well connected, it also provides more interesting team-based features for the predictor to hit the mark.

Table **??** presents the predictor hits and misses for ratings, whose errors are more evident than the previous parameters. The main reason here could be cultural factors. For instance, movies like *Jonas Brothers* (2009) and *Hanna Montana & Miley Cyrus* (2008) which are about teenage pop singers, received extremely low ratings. Nonetheless, these movies were produced by experienced and well connected teams with the support of a huge producing company (Walt Disney Pictures), have extremely

Table 5.10: Errors in predictive results for ratings

(a) **Predictor Misses**

Title	Clustering	Degree	Error	Rating
Jonas Brothers: The 3D (2009)	.466	149	+.551	.137
Hannah Montana & Miley (2008)	.880	16	+.484	.228
Zeiten ändern Dich (2010)	.221	148	+.339	.341
Utomlyonnye solntsem 2 (2010)	.763	24	+.314	.481
Mausam (2011)	.564	26	+.209	.509
Cars 2 (2011)	.600	4	+.296	.622
Zombieland (2009)	.130	272	-.301	.790
How to Train Your Dragon (2010)	.278	183	-.257	.855
The Dark Knight (2008)	.202	343	-.253	.961

(b) **Predictor Hits**

Title	Clustering	Degree	Error	Rating
Dragonball Evolution (2009)	.281	105	+.422	.174
Far Cry (2008)	.298	165	+.240	.295
BloodRayne: The Third (2011)	.304	141	+.123	.367
It's Alive (2009)	.134	850	+.012	.487
Livide (2009)	.321	33	.000	.598
Scusa ma ti chiamo amore (2008)	.076	18	.000	.641
No One Killed Jessica (2011)	.327	112	+.001	.717
Gangs of Wasseypur (2012)	.459	92	-.005	.889
The Dark Knight Rises (2012)	.198	372	-.168	.908

popular actors/singers and made significant box offices. It is possible that the ratings came from the kids' parents and not the kids themselves, who would probably rate them as an award quality movie. Such cultural aspects are very hard (if not impossible) to detect from movie topology or other concrete movie features considered here.

The hits on Table **??** are probably the hardest to explain. Nonetheless, as for the misses, we postulate the cultural aspects involved in these movies could also explain such results. Moreover, the profiles of the three top performance movies are completely different: *No one Killed Jessica* is an India-produced crime-drama-thriller; *Gangs of Wasseypur* is an R-rated Indian 320-minute movie featured at Cannes Film Festival; and *The Dark Knight Rises* ends Christopher Nolan's Batman trilogy.

5.3.4 Final Considerations

In this chapter, we have confirmed that it is indeed possible to predict success by considering only data available *before* the movie is released. Also, considering a large dataset for the prediction task is good; however, 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, confirming this work main hypothesis.

Besides identifying the most relevant features, it is also important to assess their *strength* by evaluating their coefficients in the regression model. Note that features with a higher coefficient have more impact in the predicted value. Hence, we take an already trained regression model (the one for votes, whose fitting was the best among the three parameters) and sort its coefficients in descending magnitude. The coefficients with the highest magnitudes are: the team's mean previous gross, mean previous experience, the harmonic mean of clustering coefficients, contracted degree, genres Drama and Documentaries, runtime and continent Africa.

Out of all topological features, the impact of the *clustering coefficient* and *degree* are significantly higher than the others (their actual values are informed in the illustrative samples on the hits and misses analyzes in Section **??**). Also, teams with more ties to producers outside the current team are more likely to succeed, as are teams with the lowest levels of clustering coefficients. These indicate that more homogeneous teams tend to perform worse. In other words, this is a very important result and suggests that assembling heterogeneous teams with many *external connections* is key in forming a team with higher success odds. It is important to notice that 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 [**???**].

Chapter 6

Conclusion and Future Work

In the context of forming high-achieving collaboration teams and predicting whether a collaboration team will produce successful work, we present a predictive analysis of movie producing team's success. This analysis employs features and characteristics from previous work, as well as novel topological characteristics derived from the agent's social ties. We found certain patterns in topological organization (such as low network constraint and high closeness) of teams are associated with success. This has profound implications in team formation research. This work is a first glance at how topology of teams affects movie performance. However, we hope this claim may be generalized beyond movie production teams.

Such association between network topology of collaboration teams and success could be used in many practical applications. For instance, it could be used to guide strategies for helping managers responsible for team formation to assemble teams with higher performance. Hence, they could optimize their revenue capacity and social impact. The results could also be used to develop tools that automatically analyze online enterprise social networks. This could be achieved by assessing how far teams on the network are from the best performing configurations. Future algorithms could suggest changes in the network in order to achieve such configurations. Nonetheless, one fundamental challenge is to acquire the dataset to support a predictive model.

This study revealed how topological and non-topological characteristics of movie production teams are related to movie success. There has been prior work on the impact of topological characteristics on movie performance and success. In such a context, this work is novel on aggregating topological features from teams. Moreover, this was performed on a very large dataset (whereas previous work considers just a small fraction of movies). However, we also show a larger and more heterogeneous dataset can decrease prediction accuracy. This suggests that movie success prediction

may require more sophisticated models that further consider temporal phenomena and topological aspects.

Moreover, we found the topological organization of global movie production teams is steadily evolving throughout the years. Our study confirms the "rich get richer" phenomenon with movies: teams with a better success past are more likely to produce successful movies. However, our study also confirms the importance of including fresh producers into teams, who act as bridges in the social network.

In the future, we plan to improve the network model by developing more effective ways to work with the network graph. Doing so could allow considering the movie's full cast and crew for example. Another interesting direction would be to apply this methodology for analyzing other types of collaborative networks (such as the one formed by research paper publishers or by corporative knowledge workers performing teamwork). This would allow confirming that the social structure of these networks change profoundly with time, and that social metrics from their collaboration effort can help explaining team success.

The prediction model could also be improved in many ways, such as by using more advanced means for feature selection. Also, more complex machine learning tools, such as artificial neural networks, which are able to detect non-linear relationships, could be used to possibly further improve the prediction accuracy. The model could be extended to predict other relevant factors such as forecasting which movies are going to be Oscar winners. The model could also be extended by considering new features, such as the presence of famous actors.

Finally, the initial results of this dissertation were published in **?**. The main results were published in **?**.