

# Data-driven Blockbuster Planning on Online Movie Knowledge Library

Ye Liu\*, Jiawei Zhang<sup>†</sup>, Chenwei Zhang\*, Philip S. Yu\*<sup>‡</sup>

\*Department of Computer Science, University of Illinois at Chicago, IL, USA  
{yliu279, czhang99, psyu}@uic.edu

<sup>†</sup>Department of Computer Science, Florida State University, FL, USA  
{jzhang}@cs.fsu.edu

<sup>‡</sup>Institute for Data Science, Tsinghua University, Beijing, China

**Abstract**—In the era of big data, logistic planning can be made data-driven to take advantage of accumulated knowledge in the past. While in the movie industry, movie planning can also exploit the existing online movie knowledge library to achieve better results. However, it is ineffective to solely rely on conventional heuristics for movie planning, due to a large number of existing movies and various real-world factors that contribute to the success of each movie, such as the movie genre, available budget, production team (involving actor, actress, director, and writer), etc. In this paper, we study a “Blockbuster Planning” (BP) problem to learn from previous movies and plan for low budget yet high return new movies in a totally data-driven fashion. After a thorough investigation of an online movie knowledge library, a novel movie planning framework “Blockbuster Planning with Maximized Movie Configuration Acquaintance” (BigMovie) is introduced in this paper. From the investment perspective, BigMovie maximizes the estimated gross of the planned movies with a given budget. It is able to accurately estimate the movie gross with a 0.26 mean absolute percentage error (and 0.16 for budget). Meanwhile, from the production team’s perspective, BigMovie is able to formulate an optimized team with people/movie genres that team members are acquainted with. Historical collaboration records are utilized to estimate acquaintance scores of movie configuration factors via an acquaintance tensor. We formulate the BP problem as a non-linear binary programming problem and prove its NP-hardness. To solve it in polynomial time, BigMovie relaxes the hard binary constraints and addresses the BP problem as a cubic programming problem. Extensive experiments conducted on IMDB movie database demonstrate the capability of BigMovie for an effective data-driven blockbuster planning.

**Index Terms**—Knowledge Base Discovery; Blockbuster Configuration Planning; Data-driven Application

## I. INTRODUCTION

The movie industry attracts great interests from both movie investors and the public audience because of its high profits and entertainment nature. Attracted by the huge market, lots of investors are inquiring about identifying high-gross movies to invest in. Besides recognizing profitable movies, it is rewarding to provide a reasonable and promising planning for a new movie at its developmental stage, which has been greatly ignored in previous works due to the complexity of various factors, including the movie genre and production team (actor, actress, writer and director).

The booming movie industry has accumulated thousands of previous movies as well as their gross statistics, which may

serve as a movie knowledge library to help achieve better results for future movie planning. Therefore it is no longer efficient to rely on conventional heuristics for comprehensive movie planning [1]. Data-driven movie planning methods are in great need to exploit the accumulated knowledge to support the decision-making process when planning for a new movie. The data-driven planning has shown a huge success on the well-known TV series “House of Cards”, produced by Netflix, using the data collected from viewer<sup>1</sup>.

Generally, popular movie genres and renowned movie stars are the favorable choices during the planning so as to maximize the gross. But remuneration of the movie stars and movie’s available budget also need to be considered in the movie planning. Meanwhile, a seamless collaboration among team members is the premise of high gross. For example, it will always be easier for directors to continue working on a new movie with the movie genre and production team members that they are acquainted with. And the old acquaintances can always have a tacit understanding and easy to arouse spark when they cooperate in their new movies.

**Problem Studied:** In this paper, a research problem, namely the “Blockbuster Planning” (BP) problem, is introduced. Given an online movie knowledge library which consists of the existing movie information, we plan the movie configuration including genre and production team for a new movie under a pre-specified budget. We note that although there are occasions where a low budget production with unknown stars becomes a hit, we focus on the common cases involving known persons with available data. The objective of an optimal planning is to achieve: (1) the maximized gross, and (2) the optimized acquaintance among the movie configuration factors.

The BP problem studied in this paper is a novel research problem, and few existing methods can be applied to solve it. The BP problem significantly differs from related works, such as (1) *movie gross prediction* [2], (2) *viral marketing* [3], [4], (3) *team formation* [5], [6]. (1) The *movie gross prediction* problem [7] studied in existing works merely focuses on inferring the movie gross while the BP problem aims at providing the optimal planning of various movie factors which can lead to the optimal gross for investors. (2)

<sup>1</sup><https://thenextweb.com/insider/2016/03/20/data-inspires-creativity/>

BP and the *viral marketing* problems [4] are both planning problems that aimed at maximizing certain target objectives, but they are solving totally different problems in distinct scenarios: *a) viral marketing* problems are usually studied in online social networks based on certain information diffusion models, while the BP problem is studied in the online movie knowledge libraries instead; *b) viral marketing* problems aim at maximizing the infected users, while BP's objective is to maximize the movie gross; *c) instead of selecting the optimal users in viral marketing* problems, the BP problem aims at planning for an optimal movie factor configurations. Recently, a variation of the LT model named PNP [8] is proposed for the movie design problem. The objective of PNP is very similar to our work except that PNP aims to attract most of the target users but our model aims to achieve the maximum gross under the given budget. (3) Different from conventional *team formation* problems [5], where team members are planned for the entrepreneurial team project base on satisfying skill qualification and minimizing the communication cost of the team members, our method also aims to maximize movie gross.

The BP problem is challenging to solve due to:

- *Unknown Movie Success Factors*: What are the contributing factors in the success of a movie? Few research works have ever been studied this problem, and relevant movie factors are still unknown.
- *Movie Gross/Budget Function*: How much gross (budget) can a movie make (require), given a configuration of the movie success factors? A proper estimation of the movie gross and budget will be required for studying the BP problem.
- *Movie Configuration Acquaintance Function*: How to compute the acquaintance scores among the movie configuration factors? A function that can measure acquaintance properly is needed in defining the BP problem.
- *NP Hardness*: Based on our analysis, we demonstrate that the BP problem is actually an NP-hard problem, and no solution exists that can solve it in polynomial time if  $P \neq NP$ .

To solve the aforementioned challenges, a new movie planning framework “Blockbuster Planning with Maximized Movie Configuration Acquaintance” (BigMovie) is proposed in this paper. With a thorough analysis of an online movie knowledge library dataset, IMDB, a set of factors affecting movie success are identified. The effectiveness of these extracted factors are validated in Section IV. The acquaintance scores of the movie configuration factors can be calculated based on an acquaintance tensor constructed with the historical collaboration records which is discussed in great detail in Section V. The BP problem is formulated as a constrained optimization problem with hard binary constraints, which aims at maximizing the inferred gross function as well as the acquaintance measure. We further demonstrate that BP is at least as difficult as the *Knapsack problem* and the *Maximal Clique problem*, which renders the BP problem to be NP-hard as well. By relaxing

the hard constraints, we introduce an approximation solution to resolve the problem in polynomial time. For the experimental result, we can see BigMovie outperforms the competitors. In addition, at the end of the paper, the case study is provided, which demonstrate that by using BigMovie, a lucrative movie planning can be achieved.

## II. PROBLEM FORMULATION

In this section, we will first define several important concepts used in this paper, and then provide the formulation of the BP problem.

### A. Notation

At the beginning of this section, we will first define some notations used in this paper. Throughout this paper, we will use lower case letters (e.g.,  $x$ ) to denote scalars, lower case bold letters (e.g.,  $\mathbf{x}$ ) to denote column vectors, upper case bold letters (e.g.,  $\mathbf{X}$ ) to denote elements of matrices, upper case calligraphic letters (e.g.,  $\mathcal{X}$ ) to denote sets, and bold-face upper case letters (e.g.,  $\mathbf{X}$ ) to denote matrix and high-order tensors.  $T$  is used to represent the transpose of a vector (e.g.,  $\mathbf{x}^T$ ).  $\|\cdot\|_1$  is the  $\ell_1$ -norm of vector (e.g.,  $\|\mathbf{x}\|_1$ ).

### B. Terminology Definition

**Definition 1. Online Movie Knowledge Library**: An online movie knowledge library can be represented as an undirected graph  $G = (\mathcal{M}, \mathcal{C}, \mathcal{E}, \mathcal{A})$ , where node set  $\mathcal{M} = \{m_1, m_2, \dots, m_n\}$  denotes the set of  $n$  movies in the library and  $\mathcal{C} = \{c_1, c_2, \dots, c_l\}$  is the set of  $l$  production team members. The node set  $\mathcal{C}$  can be divided into  $\mathcal{C}^t \cup \mathcal{C}^s \cup \mathcal{C}^w \cup \mathcal{C}^d$ , which denote the set of actors, actresses, writers and directors, respectively. Link  $\mathcal{E}$  represents the relationship between movie production team and movies. For instance, link  $((c_i, m_j) \in \mathcal{E})$  indicates participation of a production team member  $c_i$  in a movie  $m_j$ . And set  $\mathcal{A}$  denotes the attribute of node set  $\mathcal{M}$ . For the movie  $m_i$ , the relative attribute is  $\mathcal{A}_{(m_i)} = \mathcal{A}_{(m_i)}^g \cup \{a_{(m_i)}^b, a_{(m_i)}^g\}$ , where  $\mathcal{A}_{(m_i)}^g$  is the genre list of movie  $m_i$ ,  $a_{(m_i)}^b$  is the budget of movie  $m_i$  and  $a_{(m_i)}^g$  is the gross of movie  $m_i$ .

**Definition 2. Movie Configuration**: Each movie  $m_i \in \mathcal{M}$  in the online knowledge library will have an unique configuration, covering movie production team (involving actor, actress, writer and director), movie genre, etc, which can be represented as vector  $\mathbf{x}_{(m_i)} = [\mathbf{x}_{(m_i)}^t, \mathbf{x}_{(m_i)}^s, \mathbf{x}_{(m_i)}^d, \mathbf{x}_{(m_i)}^w, \mathbf{x}_{(m_i)}^g]$   $\in \mathbb{R}^{1 \times N}$ , where  $\mathbf{x}_{(m_i)}^t$  represents the list of all actor,  $\mathbf{x}_{(m_i)}^s$  is the list of all actress,  $\mathbf{x}_{(m_i)}^w$  represents the list of all writer,  $\mathbf{x}_{(m_i)}^d$  represents the list of all director and  $\mathbf{x}_{(m_i)}^g$  represents the list of all genre of a movie  $m_i$ .  $N$  is the sum of length of those lists. We will provide detailed representations in Section 4.1. Besides those factors covered in the above movie configuration definition, various other relevant factors (e.g., movie language, production country, etc.) can also be effectively incorporated with a simple extension to the definition, which will not be studied in this paper.

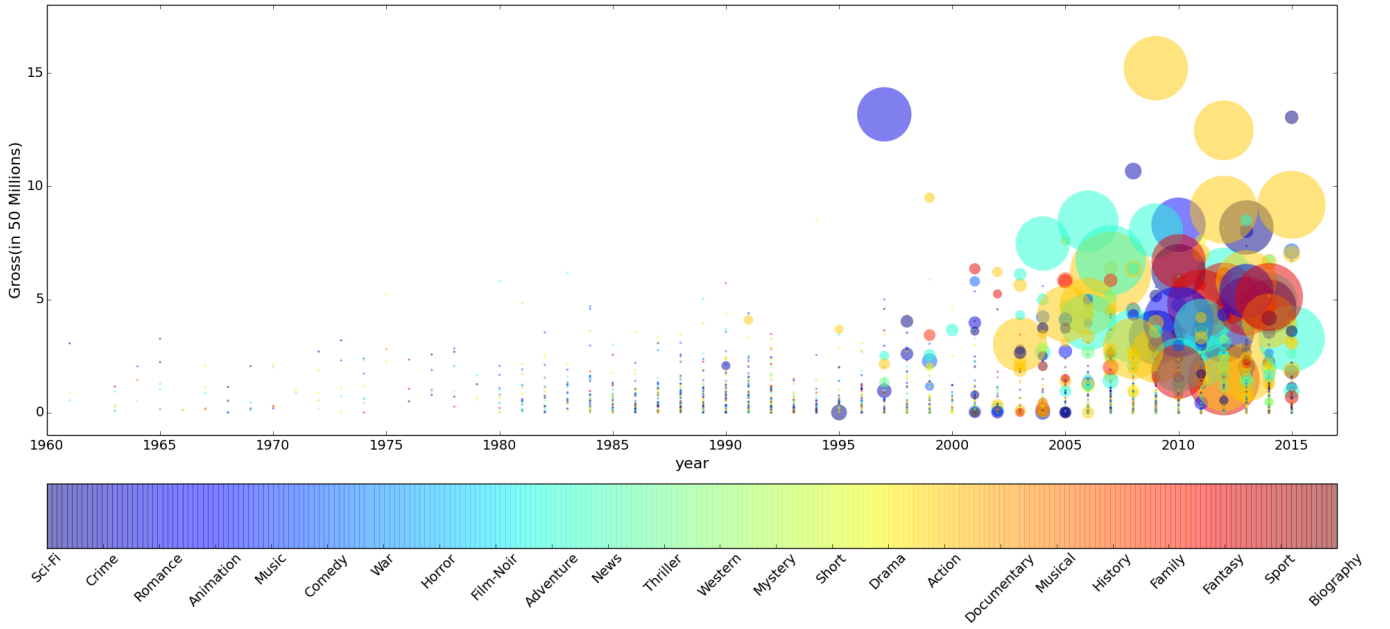


Fig. 1. Movie General information of IMDB Movie

**Definition 3. Movie Configuration Acquaintance:** Given two movie team members  $c_i, c_j \in \mathcal{C}, c_i \neq c_j$  and a movie genre  $g_k \in \mathcal{A}_{(m_i)}^g$ , their acquaintance can be represented as  $Acquaintance(c_i, c_j, g_k)$ , denoting their historical collaboration frequency. For instance, if crews  $c_i$  and  $c_j$  participate  $t$  times in  $g_k$  genre movie,  $Acquaintance(c_i, c_j, g_k) = t$ .

### C. Problem Formulation

**Definition 4. Blockbuster Planning Problem:** Given a fixed budget  $B$ , the objective of BP is to plan a movie configuration  $\mathbf{x}$  that achieves maximum movie gross and maximum movie configuration acquaintance simultaneously, subject to the budget  $B$ .

Let  $Budget(\mathbf{x})$  denotes the cost by using movie configuration  $\mathbf{x}$ ,  $Gross(\mathbf{x})$  estimates the gross earned by using  $\mathbf{x}$  and  $Acquaintance(\mathbf{x})$  measures the acquaintance of movie configuration. Formally, the BP problem aims at inferring the optimal movie configuration  $\mathbf{x}^*$  which can maximize the following objective function

$$\mathbf{x}^* = \arg \max_{\mathbf{x}} \alpha \cdot Gross(\mathbf{x}) + \beta \cdot Acquaintance(\mathbf{x}), \quad (1)$$

*s.t.*  $Budget(\mathbf{x}) \leq B$

In above equation, the concrete representation of function  $Budget(\mathbf{x})$  and  $Gross(\mathbf{x})$  will be provided in Section IV and function  $Acquaintance(\mathbf{x})$  will be provided in Section V-A.  $\alpha$  is the coefficient of  $Gross(\mathbf{x})$  and  $\beta$  is the coefficient of function  $Acquaintance(\mathbf{x})$ . Analysis of parameters  $\alpha$  and  $\beta$  will be provided in the Section V-E.

## III. ONLINE MOVIE KNOWLEDGE LIBRARY STATISTICAL ANALYSIS

Before introducing the method to solve the blockbuster planning problem, in this section, we first study the IMDB<sup>2</sup> datasets to provide some statistical analysis about the factors affecting movie gross. The analysis of the IMDB movies focuses on several important aspects like the gross, budget, genres and production team information (Actor, Actress, Director, Writer), which provides fundamental insights for the blockbuster planning framework. Among the crawled IMDB movies, only 3,156 movies contain the gross and budget information, and they belong to 24 genres and cover 72,786 actors, 38,951 actresses, 4,576 writers and 1,682 directors.

### A. General Movie Information Statistics

In this section, we study general information, like budget and genre regarding the movie gross.

The results are shown in Figure 1. In this figure, we provide the information distribution of IMDB datasets in terms of their production years. In the plot, each circle denotes a movie, whose x axis and y axis denote the movie gross and the production year, respectively. Meanwhile, the circle diameter represents the budget of the movies (larger circle corresponding to movies with bigger budgets). Additionally, we use different colors, shown in the color bar below the figure, to represent the corresponding genre of each movie.

According to Figure 1, we observe that the number of movies produced in recent years are increasing. For instance, according to our dataset, the number of movies produced in years 1980, 1990, 2000, 2010, and 2015 are 12, 58, 77, 137 and 158 respectively. Besides the movie numbers, we also

<sup>2</sup><http://www.imdb.com>

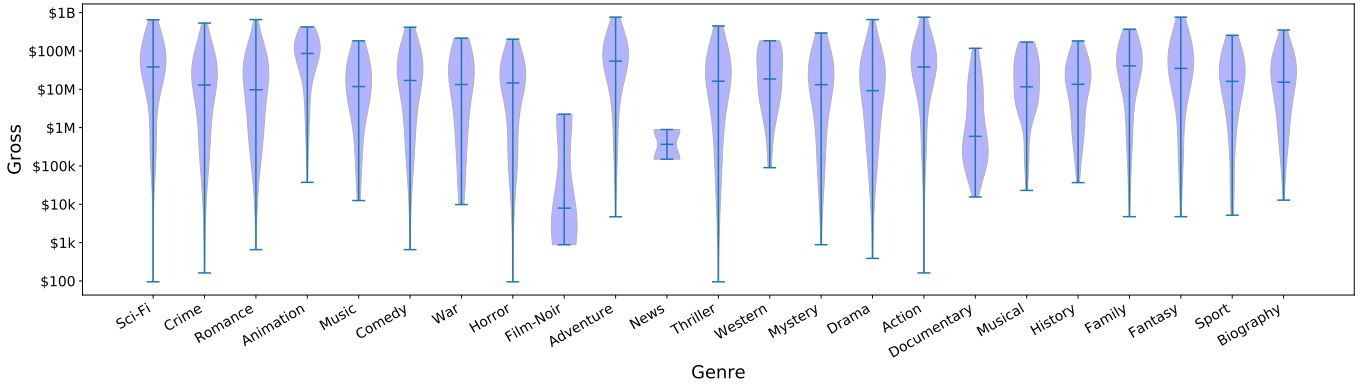


Fig. 2. Movie Gross Distribution of IMDB Movie Genres

discovered several important observations regarding the movie budget, gross and genre information.

**Recent Movies Have Higher Budget And Gross:** According to the movie budget data, a majority of the movies produced before 2000 have budgets under \$200 million while recent movies have relatively higher budgets (i.e., the circles in recent years are much larger). Among the top ten movies receiving the highest budget, six of them were produced within the past five years. Simultaneously, the movie gross of the past ten years is much higher than before (i.e., dot in recent years are much higher). Few of movies produced before 2000 had a gross of more than \$250 million while some movies produced after 2000 reached more than \$750 million on gross, which shows the growth of the movie industry. Among top ten highest gross movies, five of them were produced within the past five years, for example movie “Jurassic World” (\$652 million), “The Avengers” (\$623 million) and “The Hunger Games: Catching Fire” (\$424 million). The movie “Avatar” (produced in 2009) achieves the highest gross in our dataset, which is \$760 million. Additionally, the growth of the movie industry brings a big gross discrepancy from the recent movies, because the gross variance between movies produced before 2000 is smaller than movies produced after 2000.

**Movie Genre Distribution And Performance:** Generally, each movie can belong to more than three movie genres. For all movies, the top three movie genres on most movies include “Drama”, “Adventure” and “Fantasy”. Movies belonging to any of those three genres are more than 91% of the total movies. In order to further analyze the overall genre preference of audiences, we show the violin plot on gross of all movies in Figure 2. In this figure, the horizontal bar in each box denotes the median gross of each genre, vertical bar denotes the range of gross in each genre, and the width of the violin shows the quantity of the movie in the same gross.

By comparing the positions of the horizontal bar of each movie genre, we observe that the median movie gross fluctuates widely on different genres. For instance, the median gross of the “Animation” and “Adventure” genres are \$85 million and \$68 million respectively, but those of the “Film-Noir” and “News” genres only have \$89k and \$95k. Additionally, the box

height of some movie genres, like “News” and “Short”, are relatively short compared to the remaining movie genres. By studying the data, we observe that these movie genres are of a relatively small minority, and less than ten movies in total belong to these genres according to our IMDB dataset.

### B. Movie Production Team Statistics

After analyzing the common movie information, we believe that production team information which influence movie gross are more important than those common movie information. For example, it’s more likely that an audience watch a movie due to his/her favorite actress or actor participation. Therefore, in this section, we will analyze some latent movie information such as the movie production team information, which are actor, actress, writer and director. Moreover, we will discuss the movie configuration acquaintance and why it’s important to consider it when planning the blockbuster.

1) *Production Team and Movie Gross:* We show the stacked bar plot of the top ten movie production team members whose movies have the highest accumulative gross. In each stacked bar, the different color represents different movies. The height of the bar represents gross of the given movie, and the higher the bar is, the higher the gross is.

**Movie Gross vs. Actor:** Frank Welker’s movies have the highest gross according to Figure 3(a); He participated in 66 movies based on our dataset and most of them, he acts as voice actor. His movies earned a total of \$6,579.99 million with an average of \$99.7 million. Among the top ten gross maker actors, Stan Lee is the actor who has the highest average gross of \$217.8 million and he is second highest grossing actor. From Figure 3(a), we can know that most of the those actors act in an average of more than 30 movies, which means that the famous actors are very popular.

**Movie Gross vs. Actress:** Compared to actors, actresses relatively act in less movies as shown in Figure 3(b). Cate Blanchett is the actress who acts in the highest number of movies; She was involved in 35 movies, and she ranks fifth in the top ten actresses. The number of movies she has acted in is much less than Frank Welker. That shows that actors usually act in more movies than actresses. Generally speaking, we

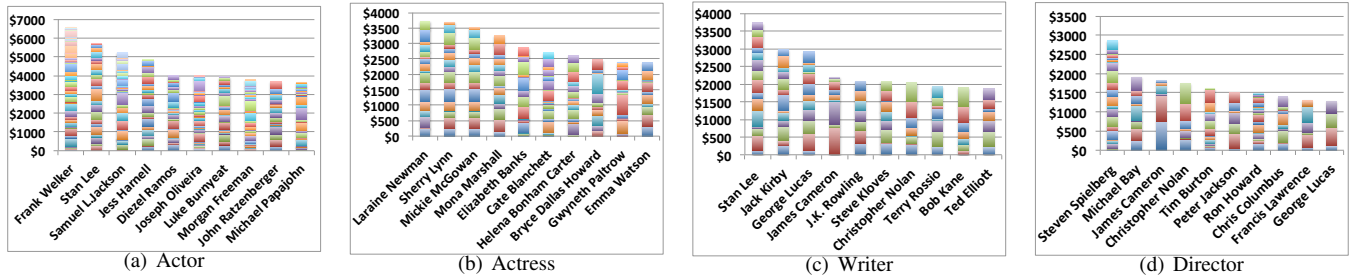


Fig. 3. Accumulative Movie Gross vs. Production Team Member. From the left to the right which are Actor, Actress, Writer and Director

Actor Actress collaboration				Actor Writer collaboration			
Actor	Actress	Times	Genre	Actor	Writer	Times	Genre
Bernard Lee	Lois Maxwell	7	Action Adventure Thriller	Adam Sandler	Tim Herlihy	10	Comedy Drama
Johnny Depp	Helena B. Carter	6	Adventure Fantasy Comedy	Daniel Radcliffe	J.K. Rowling	7	Adventure Family Fantasy
Burt Young	Talia Shire	6	Drama Sport	Bernard Lee	Ian Fleming	7	Action Adventure Thriller
Actor Director collaboration				Actress Writer collaboration			
Actor	Director	Times	Genre	Actress	Writer	Times	Genre
Johnny Depp	Tim Burton	8	Horror Comedy Drama	Lois Maxwell	Ian Fleming	14	Action Adventure Thriller
Adam Sandler	Dennis Dugan	7	Comedy Romance	Lois Maxwell	Richard Maibaum	11	Action Adventure Sci-Fi
Antonio Banderas	Robert Rodriguez	7	Action Adventure Crime	Mia Farrow	Woody Allen	11	Comedy Drama
Actress Director collaboration				Writer Director collaboration			
Actress	Director	Times	Genre	Writer	Director	Times	Genre
Mia Farrow	Woody Allen	11	Comedy Drama	Ethan Coen	Joel Coen	16	Comedy Crime
Giannina Facio	Ridley Scott	8	Comedy Drama	Fran Walsh	Peter Jackson	10	Adventure Fantasy
Helena B. Carter	Tim Burton	7	Fantasy Adventure Horror	Bobby Farrelly	Peter Farrelly	8	Comedy Romance

TABLE I  
THE COLLABORATION AMONG PRODUCTION TEAM

can make a conclusion that actresses act in fewer movies than actors.

**Movie Gross vs. Writer:** From Figure 3(c), we observe that the total gross difference between the last seven writers is not obvious. Their movies' gross only have a few million difference. But compared to the first writer and the last writer, their difference is observable. The total gross of first writer is twice that of the last writer, which shows that famous writers are in great demand.

**Movie Gross vs. Director:** Directors are responsible for the whole production process of the movies, and their crucial roles may determine the movie quality. The Figure 3(d) shows that top ten directors participated in relatively few movies. Steven Spielberg participated in 23 movies which is the highest of all the directors in our dataset. The average gross of each director is around \$165.7 million dollars, which is higher than actor, actress and writer. We can see that best directors relatively act less time of making movie, but each movie they making have a high gross.

We show the different character between actor, actress, writer and director regarding to movie gross. But they all have a strong relation to the movie gross and are the necessary factors for planning the blockbuster.

2) *Movie Configuration Acquaintance:* We show that the production team has a strong connection with the movie gross. Some of them are the guarantee to a high gross movie. To ensure a high gross movie, effective collaboration among production team also needs to be analyzed. We will see that production members have a strong collaboration with each

other. There are six different types of collaborations, which are shown in the Table I.

**Actor and Actress Collaboration:** Bernard Lee and Lois Maxwell participated in seven movies, like "Dr. No" which is the first James Bond film, and the genre of those movies are "Action", "Adventure" and "Thriller". The second frequent collaborating partners are Johnny Depp and Helena B. Carter. They participated in recent well-known movies, like "Sweeney Todd: The Demon Barber of Fleet Street" and "Alice in Wonderland".

**Actor and Writer Collaboration:** Daniel Radcliffe and J.K. Rowling collaborated in the series of "Harry Potter". For the third frequent partners, Bernard Lee and Ian Fleming collaborated in many "Action", "Adventure" or "Thriller" genre movies and actress Lois Maxwell also participated in most of those movies, which shows strong collaboration among the three of them.

**Actor and Director Collaboration:** Johnny Depp and Tim Burton collaborated in eight movies which makes them the most frequent partners. Among those eight movies, Helena B. Carter also participated in five of them, like "Corpse Bride" and "Dark Shadows". Adam Sandler and Dennis Dugan collaborated seven times, and in those movies, Tim Herlihy also participated.

**Actress and Writer Collaboration:** Lois Maxwell was a famous actress during the 1960s and 1970s. She collaborated fourteen times with Ian Fleming and eleven times with Richard Maibaum. The most common genres they participated in are "Action", "Thriller" and "Adventure".



**Actress and Director Collaboration:** Mia Farrow and Woody Allen collaborated in the “Comedy” or “Drama” genre movies eleven times and those movies have “Comedy” or “Drama” genre.

**Director and Writer Collaboration:** Directors and writers collaborate more often than the other relationships. Ethan Coen and Joel Coen even collaborated sixteen times, with most of the movie genres being “Comedy” or “Crime”.

All of these collaborations show a strong relationship between the production team members and movie genre. The movies with high collaborations have a high gross and are well-known by viewers. Besides, the binary relationship cannot fully represent those collaborations. For example, Bernard Lee as actor, Ian Fleming as actress and Ian Fleming as writer, participated in many movies together. Moreover, those movies have the same genre, “Action”, “Adventure” and “Thriller”, which shows that only considering the collaboration between team members is not enough. Instead, considering the collaboration between team members based on movie genre is necessary. In our dataset, there are a lot of same or more complex collaborations like this. Therefore, the movie configuration acquaintance must need be studied when we make the blockbuster planning. The more details of movie configuration acquaintance term will be discussed in the Section V-A.

#### IV. MOVIE CONFIGURATION VERIFICATION

To verify the effectiveness of these factors aforementioned on estimating the movie gross and budget and to learn the weight of movie configuration, in this section, we will build a prediction model to learn their correlations. A set of features (i.e., the configuration) will be extracted for the movies based on each of the factors first. After that, a regression model will be built to project the movie configurations to their budget and gross.

##### A. Feature Extraction and Movie Budget/Gross Estimation

Features like actor, actress, writer and director are a bag-of-words. Moreover, the relationship between a movie and its feature is one-to-many. For example, each movie belongs to more than one genre. And each movie has more than one actor or actress.

We use  $e$  to represent an element in the movie  $m_i$  configuration  $\mathbf{x}_{(m_i)} = [\mathbf{x}_{(m_i)}^t, \mathbf{x}_{(m_i)}^s, \mathbf{x}_{(m_i)}^d, \mathbf{x}_{(m_i)}^w, \mathbf{x}_{(m_i)}^g]$ . We use the binary value to set the element. Namely, for example, if actor  $t_j$  participates in movie  $m_i$ , then  $\mathbf{x}_{(m_i)}^{t_j}$  equals to 1, otherwise, it equals to 0. In the same way, we can get the vector representation of  $\mathbf{x}_{(m_i)}^s, \mathbf{x}_{(m_i)}^d, \mathbf{x}_{(m_i)}^w$  and  $\mathbf{x}_{(m_i)}^g$ .

After extracting all the features of a movie, we can train an approximation model of budget function and gross function. Since the cost and income of each production team member can not be negative, we use Lasso linear regression [9] and force the coefficients to be non-negative. Formally, they can be represented as:

$$\text{Budget}(\mathbf{x}) = \min_{\mathbf{w}_b, b_b} \|\mathbf{B} - (\mathbf{w}_b^T \mathbf{x} + b_b)\|_2^2 + \lambda \|\mathbf{w}_b\|_1 \text{ s.t. } \mathbf{w}_b \geq 0 \quad (2)$$

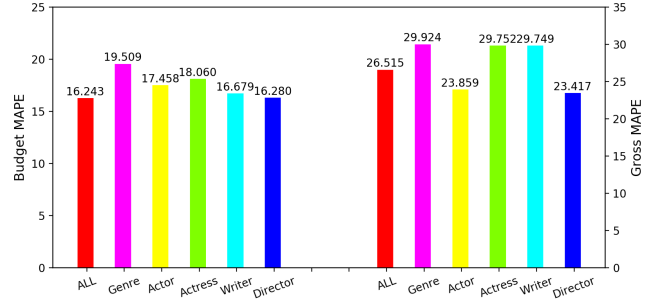


Fig. 4. The MAPE of different approximation models

$$\text{Gross}(\mathbf{x}) = \min_{\mathbf{w}_g, b_g} \|\mathbf{G} - (\mathbf{w}_g^T [\mathbf{B}, \mathbf{x}] + b_g)\|_2^2 + \lambda \|\mathbf{w}_g\|_1 \text{ s.t. } \mathbf{w}_g \geq 0 \quad (3)$$

where  $\mathbf{x}$  is the movie configuration and  $B$  is the budget and  $G$  is the gross of movie in the IMDB knowledge base.  $\mathbf{w}_b$  and  $\mathbf{w}_g$  are the weights and  $b_b$  and  $b_g$  are the intercepts of function  $\text{Budget}(\mathbf{x})$  and function  $\text{Gross}(\mathbf{x})$  function respectively.  $\lambda$  is the coefficient of the  $\ell_1$ -norm regularization, which we set to 0.1 in the experiment.

##### B. Movie Budget/Gross Estimation Experiment Results

In the experiments, 80% of 3156 movies are used as training data and 20% are used as testing data. The 5-fold cross validations are performed on training data. We analyze the effectiveness of those features on function  $\text{Budget}(\mathbf{x})$  and function  $\text{Gross}(\mathbf{x})$  separately. To measure the performance, we use the mean absolute percentage error (MAPE) as the evaluation metrics which represent as follow:

$$\text{MAPE} = \frac{100}{n} * \sum_{t=1}^n \left| \frac{A_t - E_t}{A_t} \right| \quad (4)$$

where  $A_t$  is the actual value and  $E_t$  is the estimated value.

The following is our compared methods:

##### Models using all information

- ALL: Method ALL builds the gross approximation with all features, which are genre, actor, actress, writer and director.

##### Models using partial information

- Genre: Method Genre builds the gross and budget approximation model with the genre information alone.
- Actor: Method Actor just uses the actor feature to build the gross and budget approximation model.
- Actress: Method Actress builds the gross and budget approximation model with the actress feature.
- Writer: Method Writer builds the gross and budget approximation model with writer feature.
- Director: Method Director builds the gross and budget approximation model with only director feature.

The experiment results are available in Figure 4. The left part shows the MAPE on predicting the movie budget and the right part shows the MAPE of movie gross estimated by different methods.

Using all features together gives us the lowest MAPE, 16.243% as shown in the left part of Figure 4. When looking at the production team information, we find that writer and director have relative low MAPE, 16.679% and 16.280%, which implies writer and director positively correlated to the movie budget. While in reality, the salary of writer and director can determine how much the movie producer needs to invest in the movie. Production team information achieves a relatively lower MAPE than genre, due to wide difference of movie budgets in the same genre; therefore genre is not a good factor for the budget.

By comparing all features in the gross approximation, we can observe that Director achieves the lowest MAPE (23.417%). Actor gets the second-lowest MAPE (23.859%). Compared to those two models, the performance of ALL is not good. It's probably because feature actress, writer and genre have a large MAPE, meaning that those features have no (or weak) correlation to the movie gross. Therefore, combining all features together will achieve 26.515% on MAPE. Such result is reasonable because a director with great reputation is more likely to produce a good movie. Moreover, similarly as the conclusion in Section 4, production team information can reach a relatively lower MAPE than the genre.

Even if the performance of ALL is not as good as same only feature such as director or actor, the performance is still good. And our goal is to learn a function which can map the movie configuration with gross and budget.

## V. BLOCKBUSTER PLANNING: BIGMOVIE

Since we have already demonstrated that good collaboration between production team members is the safeguard for the profit of movie. In this section, we first formulate the movie configuration acquaintance. Based on the movie gross, budget estimation and movie configuration acquaintance function, we provide the joint objective function of BigMovie and a cubic programming algorithm to effectively solve the objective.

### A. Movie Configuration Acquaintance

As we discussed previously, it is advantageous to have production team members that have a strong collaboration to the other specific members and furthermore have strong acquaintance to the certain movie genre. If team members have already participated together before, they will have chemistry when they participate in the next movie, which may stimulate the increase of movie gross. Besides, production team members that have joined in a certain movie genre previously can more easily work together when making the same genre type movies. Those two types of acquaintances have a great effect on the movie gross and movie budget, so finding the mathematical representation of the movie configuration acquaintance is important.

**Movie Configuration Acquaintance Function:** We discuss in section III-B2 the binary relationship between two members cannot well represent their collaboration and movie genre need to be considered as well. To solve these problems, we use a three dimensional tensor  $\mathbf{W}_a \in \mathbb{R}^{C \times C \times G}$  to represent movie

configuration acquaintance, where the  $C$  is the dimension for the size of all cast,  $G$  is the dimension for the size of all movie genres. We propose to define the movie configuration acquaintance as follow:

$$Acquaintance(\mathbf{x}) = \sum_{n=0}^{C-1} \sum_{m=0}^{C-1} \sum_{l=0}^{G-1} \mathbf{W}_a[n][m][l] \cdot \mathbf{x}[n] \cdot \mathbf{x}[m] \cdot \mathbf{x}[l], \quad (5)$$

where  $n$  and  $m$  are the production team member  $\in \mathcal{C}$ . And  $l$  is the movie genre  $\in \mathcal{A}_{(m_i)}^g$

### B. Joint Objective Function

Maximize movie gross can be mathematically represented as:  $\max_{\mathbf{x}} \sum_{i=0}^{N-1} \mathbf{w}_g[i+1] \cdot \mathbf{x}[i] + b_g + \mathbf{w}_g[0] \cdot B$ , where  $\mathbf{w}_g$  is the weight and  $b_g$  is the intercept of function  $Gross(\mathbf{x})$  that we learned from section IV-A. The movie budget bound can be mathematically represented as:  $\sum_{i=0}^{N-1} \mathbf{w}_b[i] \cdot \mathbf{x}[i] + b_b \leq B$ , where  $\mathbf{w}_b$  and  $b_b$  are weights and intercepts of function  $Budget(\mathbf{x})$  we learned from section IV-A.

The objective of the BP problem is to find the optimal movie configuration that can maximize the movie gross and movie configuration acquaintance while not exceeding the movie budget bound. So the joint objective function represents as:

$$\begin{aligned} \max_{\mathbf{x}} & \alpha \left( \sum_{i=0}^{N-1} \mathbf{w}_g[i+1] \cdot \mathbf{x}[i] + b_g + \mathbf{w}_g[0] \cdot B \right) \\ & + \beta \sum_{n=0}^{C-1} \sum_{m=0}^{C-1} \sum_{l=0}^{G-1} \mathbf{W}_a[n][m][l] \cdot \mathbf{x}[n] \cdot \mathbf{x}[m] \cdot \mathbf{x}[l] \\ s.t. & \sum_{i=0}^{N-1} \mathbf{w}_b[i] \cdot \mathbf{x}[i] + b_b \leq B, \quad \forall i: \mathbf{x}_i \in \{0, 1\} \end{aligned} \quad (6)$$

where  $\alpha$  and  $\beta$  are parameters to adjust movie gross estimation and movie configuration acquaintance which are studied in Section V-E.

### C. Prove NP-Hardness

In this section, we prove that the Blockbuster Planning with maximize movie configuration acquaintance problem is a NP-hard problem. In Equation 6 of the BP problem, two objectives equations are involved: the gross equation weighted by  $\alpha$ , and the acquaintance equation weighted by  $\beta$ . By assigning  $\alpha = 1$  and  $\beta = 0$ , we will show that the *Knapsack problem* can be reduced to the BP problem, which is a classic NP-hard problem. Meanwhile, by assigning  $\alpha = 0$  and  $\beta = 1$ , we will show that the *Maximal Clique problem* can be reduced to the BP problem.

Given a set of items, each with a weight and a value, the *Knapsack problem* aims at picking the items to be included in a bag so that the total weight is less than a given limit while maximizing the total value. By treating items as features in the movie configuration vector  $\mathbf{x}$  with corresponding values in vector  $\mathbf{w}_g$  and weights in vector  $\mathbf{w}_b$ , *Knapsack problem* can be exactly reduced to the BP problem (with  $\alpha = 1$  and  $\beta = 0$ ), where the bag limit is denoted as the provided budget  $B$ . If we can identify an optimal movie configuration vector

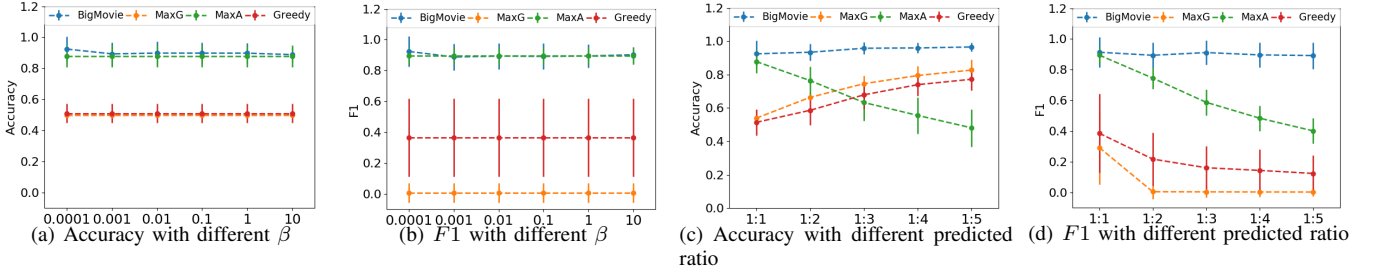


Fig. 5. Planning production team with different  $\beta$  and different predicted ratio measured by Accuracy and F1

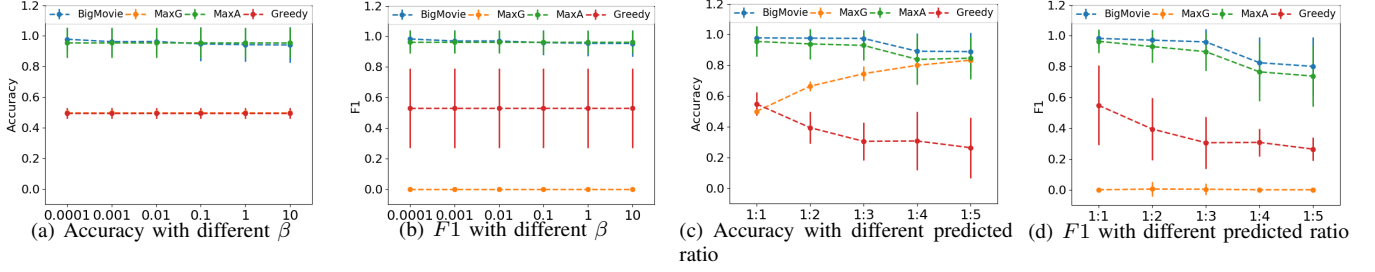


Fig. 6. Planning movie genre with different  $\beta$  and different predicted ratio measured by Accuracy and F1

$\mathbf{x}$ , the items corresponding to the features with value 1 can be selected, which will be the optimal solution to the *Knapsack problem*.

Given an undirected graph formed by a finite set of nodes and a set of undirected edges, the *k-Clique problem* aims at determining whether there exist a clique involving  $k$  nodes in graph or not. Let a tensor  $\mathbf{W}_p$  denote whether nodes can form a triangle in the input graph or not. If nodes  $n_i, n_j, n_k$  can form a triangle, then  $W_p[i][j][k] = 1$ , and 0 if not. By treating each node in the graph as a feature to be determined in the movie configuration vector  $\mathbf{x}$  and assign vectors  $\mathbf{w}_g, \mathbf{w}_b$  to be 1, the problem of identifying a clique of size  $k$  in the input graph (i.e., *k-Clique problem*) can be reduced to the problem of obtaining the optimal value of  $k(k-1)(k-1)$  in the BP problem, where the budget  $B$  takes value  $k$ . If we can identify an optimal value  $k(k-1)(k-1)$ , then the nodes corresponding to features with value 1 will be selected to form a clique of size  $k$  in the input graph.

Therefore, the BP problem containing these two objectives makes itself at least as difficult as the *Knapsack problem* [10] and the *Maximal Clique problem* [11], which renders the BP problem to be NP-hard.

#### D. Solving the Objective

Since this problem is NP-complete [12] and no polynomial-time solutions can solve the problem efficiently, we propose to solve the problem with two steps: (1) integer constraint relaxation, and (2) result post-processing. We relax the integer constraint on variables, and allow them to take real values in the range of  $[0, 1]$  to help address the problem in polynomial time [13]. Based on the obtained real-valued solution  $\mathbf{x}$  denoting the score of the feature-gross links, we post-process

the variable to binary values by pruning with a confidence threshold  $\theta$  in  $[0, 1]$ . For the variables, e.g., if  $x_i > \theta$ , we will map  $x_i$  to value 1; otherwise,  $x_i$  will be mapped to value 0. After post-processing, we can obtain the final optimal movie plan by finding corresponding genre  $g_i$  and team member  $c_i$  for the movie  $m_i$  having values  $x_i = 1$ .

#### E. Experiments

In this section, we give the experimental analysis of BigMovie, and evaluate its performance for designing new movies. We seek to answer two main questions:

- Q1.** How well does BigMovie quantitative performance?
- Q2.** Is the movie planned by BigMovie reasonable?

**1) AI: Quantitative Evaluation:** In order to quantitatively measure the performance, we use our method to design the movie production team setting or movie genre to see whether our model can match with the ground truth, the movie setting in the database. The BP problem is a new problem, and no existing methods can be applied to address it directly. To show the advantages of the framework BigMovie, we compare some other methods with BigMovie measured by accuracy and F1 score on production team and genre. The comparison methods used in the experiments are listed as follows:

- **BigMovie:** framework in the paper that achieves the maximum gross and maximizes the movie configuration acquaintance. By setting  $\alpha = 1$ , then the performance of different  $\beta$  will be studied in the experiments.
- **MaxG:** aims to maximize the movie gross, which sets the objective function Eq. (6) with  $\alpha = 1$  and  $\beta = 0$ .
- **MaxA:** only considers the movie configuration acquaintance by setting the objective function Eq. (6) with  $\alpha = 0$  and  $\beta = 1$ .



- **Greedy**[14]: iteratively picks the greedy choice to maximize movie gross, by always choosing the maximum ratio of  $\frac{w_g}{w_b}$ .

We verify the effectiveness of production team and genre planning on 3,156 IMDB movies separately. When studying the production team feature, we set the production team as unknown while the movie genre is given, and vice versa.

**Observation 1.** As shown in Figures 5 and 6, BigMovie outperforms the competing baselines **MaxG**, **MaxA** and **Greedy** and obtains higher accuracy and F1 score.

From the experiment results, we can see that BigMovie can get more than 90% accuracy on both genre and production team, which is consistently better than other methods on different  $\beta$ . When  $\beta = 0.0001$ , the best performance is achieved on production team and genre, as shown in Figures 5(a), 5(b) and 6(a), 6(b). With  $\beta = 0.0001$ , we study different ratios of positive and negative samples that are randomly selected, as shown in Figure 5(c), 5(d), 6(c) and 6(d).

**Observation 2.** Maximizing both the movie gross and movie configuration acquaintance simultaneously can achieve best performance.

For methods only depending on maximizing the gross, MaxG and Greedy, because they do not depend on  $\beta$ , their performances do not change with different  $\beta$ . Since we use Lasso with non-negative constraint to learn the weight of budget and gross, most of weights we learned are zeros. Therefore, most candidates are not selected in the production team study, when using MaxG. Such planning is less ideal because not choosing any candidate is not the goal of an optimal planning. The Greedy method achieves higher accuracy than MaxG, but still has the same problem as MaxG. In the genre study, MaxG select all the candidate genre, so when the planning scale gets larger, its performance gets worse.

For the MaxA method which only uses movie configuration acquaintance part of objective function, in both production team and genre planning, MaxA doesn't have the problem of not selecting any samples. This shows the importance of considering movie configuration acquaintance. But when the negative ratio gets bigger, MaxA performs worse. Since BigMovie outperforms MaxA, the results show the importance of considering simultaneously maximizing the movie gross and movie configuration acquaintance.

## 2) A2: Case Study:

We show a case study to demonstrate the reasonable and effectiveness of the proposed method. We choose "The Avengers" to plan, which has the fourth-highest movie gross on our dataset, \$623.27 million. For fairness, we delete the other sequel movies of "Avengers". We provide movie genre "Action", "Adventure" and "Sci-Fi", movie budget and 250 random candidates to build an about 20 casts production team.

The detail of planned movie configuration and actual movie configuration is shown in Table II.

**Observation 3.** BigMovie is rational and interpretive on planning the blockbuster movie.

In the planned movie configuration, we get a \$654.52 million on gross which is higher than the gross in original movie, \$623.27 million. All the members in the original movie are selected by BigMovie. Besides the actual members, we plan one more actor, actress and writer and two more directors which are shown with underscores. For actors, we selected Sebastian Stan. We find that the reason why Sebastian Stan was selected is that he has a strong collaboration with Chris Evans as they collaborated twice and he participated in Marvel's "Captain America" series of movies, which share the same genre with "The Avengers". For actress, Elizabeth Olsen was selected. She has a strong collaboration with others, like Scarlett Johansson, Cobie Smulders and Chris Evans, because they participated in the movie "Captain America: The Winter Soldier". Additionally, she has participated in many the "Action", "Adventure" and "Sci-Fi" type movies, as she participated in the same genre movie "Godzilla". For writer, Joe Simon was selected. He is the writer in "Captain America" series of movies. And he collaborated with Chris Evans twice as well. For director, Jon Favreau was planned, as he is the director for the "Iron Man" series of movie. He collaborated with Robert Downey Jr. three times. Michael Bay, who is the director for the "Transformers" movie series, has collaborated with Scarlett Johansson in the movie "The Island". Besides, "Iron Man" series and "Transformers" series have the same genre as the "The Avengers" series. This planned configuration shows the effectiveness of our model, because it has a higher gross than the actual configuration and even if some casts was mistaken by us, all the planned casts are reasonable.

Planned movie Configuration	
<b>Gross</b>	\$654.52
<b>Actor</b>	Robert Downey Jr., <u>Sebastian Stan</u> , Tom Hiddleston Chris Hemsworth, Clark Gregg, Mark Ruffalo Chris Evans, Jeremy Renner
<b>Actress</b>	<u>Elizabeth Olsen</u> , Scarlett Johansson, Cobie Smulders Gwyneth Paltrow, Tina Benko, M'laah Kaur Singh
<b>Writer</b>	Joss Whedon, <u>Joe Simon</u> , Zak Penn
<b>Director</b>	<u>Jon Favreau</u> , Joss Whedon, <u>Michael Bay</u>

Actual movie Configuration	
<b>Gross</b>	\$623.27
<b>Actor</b>	Robert Downey Jr., Tom Hiddleston Chris Hemsworth, Clark Gregg, Mark Ruffalo Chris Evans, Jeremy Renner
<b>Actress</b>	Scarlett Johansson, Cobie Smulders Gwyneth Paltrow, Tina Benko, M'laah Kaur Singh
<b>Writer</b>	Joss Whedon, Zak Penn
<b>Director</b>	Joss Whedon

TABLE II  
THE PLANNED AND ACTUAL MOVIE CONFIGURATION OF MOVIE "THE AVENGERS"

## VI. RELATED WORK

We have clearly illustrated the significant differences of the BP problem from the existing works in Section I. In this section, we provide a brief review of recent developments on related works.

**Movie Gross Prediction** People use different resource of information to predict the movie gross. Mestyán and Yasserli [2] used the knowledge base, Wikipedia, to predict the movie box office. Joshi et al. [15] use the sentiment analysis on movie reviews to predict the movie gross. The recent analysis of the movie gross was done through social media, like Twitter and YouTube [16][17].

**Viral Marketing** This problem focuses on finding a small set of seed nodes in a social network that maximizes the spread of influence. Kempe et al. [4], [18] first proposed two basic diffusion models, namely independent cascade model(IC) and linear threshold model(LT). These two models set the foundation of almost all existing algorithms finding seed in social networks [19]. The major drawback of their algorithm is that its inefficiency and ineffectiveness to the large networks. Later, Chen [20] proposed a greedy algorithm to approximate the influence regions of nodes. However, when the scales beyond million-sized graphs, greedy algorithm becomes unfeasible. Chen et al. proposed to use local directed acyclic graphs to explore a large-scale influence maximization algorithm [21].

**Team Formation** Lappas et al. first proposed this problem [5]. They described an approach that defined the total communication cost among the social relationships to select a subset of experts to form a qualified team for certain projects. Recently, Nikolaev et al. proposed the EngTFP problem to find the subset of users that was the most important for keeping the whole user base together [22]. Different from those two works that find a subset of users to form a qualified team for certain projects, several recent works focus on training the team members [23] [24]. Their motivation is to build a team so that teammates can benefit from interaction to improve their skills.

## VII. CONCLUSION

In this work, we studied the Blockbuster Planning (BP) problem where professional movie planning are made by exploring the accumulated knowledge in the online movie knowledge library. A novel movie planning framework named BigMovie is introduced, where we first build the gross estimation function by analyzing and investigating the real-world online movie library dataset. The weights of the movie factors learned by the gross estimation are easily interpretable, and can be directly applied to the objective function for blockbuster planning. The BigMovie framework is optimized to maximize the movie gross as well as the production team preference simultaneously. In addition, the limited budget is used as a hard constraint for the objective function to guarantee the plan achievement. Extensive experiments have been done on the real-world dataset to demonstrate the effective and advantages of the proposed framework in addressing the BP problem.

## VIII. ACKNOWLEDGEMENTS

This work is supported in part by NSF through grants IIS-1526499, IIS-1763325, and CNS-1626432, and NSFC 61672313. This work is also partially supported by NSF through grant IIS-1763365 and by FSU through the startup package and FYAP award.

## REFERENCES

- [1] J. W. Fowler, P. Wirojanagud, and E. S. Gel, "Heuristics for workforce planning with worker differences," *European Journal of Operational Research*, vol. 190, no. 3, pp. 724–740, 2008.
- [2] M. Mestyán, T. Yasserli, and J. Kertész, "Early prediction of movie box office success based on wikipedia activity big data," *PloS one*, vol. 8, no. 8, p. e71226, 2013.
- [3] M. Richardson and P. Domingos, "Mining knowledge-sharing sites for viral marketing," in *KDD*. ACM, 2002, pp. 61–70.
- [4] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *KDD*. ACM, 2003, pp. 137–146.
- [5] T. Lappas, K. Liu, and E. Terzi, "Finding a team of experts in social networks," in *KDD*. ACM, 2009, pp. 467–476.
- [6] A. Anagnostopoulos, L. Becchetti, C. Castillo, A. Gionis, and S. Leonardi, "Online team formation in social networks," in *WWW*. ACM, 2012, pp. 839–848.
- [7] R. Sharda and D. Delen, "Predicting box-office success of motion pictures with neural networks," *Expert Systems with Applications*, vol. 30, no. 2, pp. 243–254, 2006.
- [8] D. Koutra, A. Dighe, S. Bhagat, U. Weinsberg, S. Ioannidis, C. Faloutsos, and J. Bolot, "Pnp: Fast path ensemble method for movie design," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2017, pp. 1527–1536.
- [9] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.
- [10] S. Sahni, "Approximate algorithms for the 0/1 knapsack problem," *Journal of the ACM (JACM)*, vol. 22, no. 1, pp. 115–124, 1975.
- [11] I. M. Bomze, M. Budinich, P. M. Pardalos, and M. Pelillo, "The maximum clique problem," in *Handbook of combinatorial optimization*. Springer, 1999, pp. 1–74.
- [12] C. H. Papadimitriou, "On the complexity of integer programming," *Journal of the ACM (JACM)*, vol. 28, no. 4, pp. 765–768, 1981.
- [13] D. S. Hochbaum, N. Megiddo, J. Naor, and A. Tamir, "Tight bounds and 2-approximation algorithms for integer programs with two variables per inequality," *Mathematical programming*, vol. 62, no. 1, pp. 69–83, 1993.
- [14] G. B. Dantzig, "Discrete-variable extremum problems," *Operations research*, vol. 5, no. 2, pp. 266–288, 1957.
- [15] M. Joshi, D. Das, K. Gimpel, and N. A. Smith, "Movie reviews and revenues: An experiment in text regression," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2010, pp. 293–296.
- [16] K. R. Apala, M. Jose, S. Motnam, C.-C. Chan, K. J. Liszka, and F. de Gregorio, "Prediction of movies box office performance using social media," in *ASONAM*. IEEE, 2013, pp. 1209–1214.
- [17] L. Doshi, J. Krauss, S. Nann, and P. Gloor, "Predicting movie prices through dynamic social network analysis," *Procedia-Social and Behavioral Sciences*, vol. 2, no. 4, pp. 6423–6433, 2010.
- [18] D. Kempe, J. Kleinberg, and É. Tardos, "Influential nodes in a diffusion model for social networks," in *ICALP*. Springer, 2005, pp. 1127–1138.
- [19] W. Chen, Y. Wang, and S. Yang, "Efficient influence maximization in social networks," in *KDD*. ACM, 2009, pp. 199–208.
- [20] N. Chen, "On the approximability of influence in social networks," *SIAM Journal on Discrete Mathematics*, vol. 23, no. 3, pp. 1400–1415, 2009.
- [21] W. Chen, C. Wang, and Y. Wang, "Scalable influence maximization for prevalent viral marketing in large-scale social networks," in *KDD*. ACM, 2010, pp. 1029–1038.
- [22] A. Nikolaev, S. Gore, and V. Govindaraju, "Engagement capacity and engaging team formation for reach maximization of online social media platforms," in *KDD*, vol. 16, 2016, pp. 225–234.

- [23] J. Zhang, P. S. Yu, and Y. Lv, “Enterprise employee training via project team formation,” in *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. ACM, 2017, pp. 3–12.
- [24] S. Bahargam, D. Erdos, A. Bestavros, and E. Terzi, “Team formation for scheduling educational material in massive online classes,” *arXiv preprint arXiv:1703.08762*, 2017.