PICK THE RIGHT TEAM AND MAKE A BLOCKBUSTER

A SOCIAL ANALYSIS THROUGH MOVIE HISTORY
THE CHALLENGE:
TO UNDERSTAND HOW TEAMS CAN WORK BETTER

SOCIAL NETWORK + MACHINE LEARNING TO THE RESCUE
Previous research: team success

- Teamwork selection as an optimisation problem
  Anagnostopoulos et al. [2012], Tseng et al. [2004], …

- Studied team success without social parameters
  Kim et al. [2013], Elberse [2007], …

- Did not study the team as a whole
  Nemoto et al. [2011], Singh et al. [2011], …
Previous research: social features

- Studied social parameter of individuals
  Papagelis et al. [2011], Li et al. [2013], …

- Studied single social features
  Chen and Guan [2010], Schilling and Phelps [2007], …

- Performed on small datasets
  Ghiassi et al. [2015], Oghina et al. [2012], …

- No predictive analysis
  Uzzi and Spiro [2005], [Burt, 2009], …
IN PREDICTIVE ANALYSIS OF TEAM SUCCESS, DOES USING MANY TOPOLOGICAL FEATURES FROM TEAMS HELP?
Methodology

- Start with large set of collaboration data (IMDB)
- Form a social network
- Filter irrelevant data
- Extract social features from team
- Characterize this never-before-seen data
- Apply Machine Learning Techniques
  - Assess how social features help predict team success
<table>
<thead>
<tr>
<th>DATASET</th>
<th>IMDB [INTERNET MOVIE DATABASE]</th>
</tr>
</thead>
<tbody>
<tr>
<td>WORLD’S LARGEST MOVIE DATASET</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DATE</th>
<th>SIZE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1808–2014</td>
<td>12,250 MOVIES 31,698 PRODUCERS</td>
</tr>
</tbody>
</table>
MOVIE’S TYPICAL PRODUCING TEAM

PRODUCERS THAT WORK TOGETHER ARE LINKED IN A SOCIAL NETWORK
Forming a Social Network

Movies

Producers

Producer’s Social Network
Removing inactive nodes
Filtering: 238K $\rightarrow$ 32K Movies

- Filtering out movies that are
  - Not connected to giant component
  - Not from cinema
  - Just one producer
  - Released before 1930 (used for bootstrapping)
  - Not feature length (< 30 min.)
  - Not relevant (< 1,000 votes)
MOVIE’S SUCCESS PARAMETERS

NUMBER OF RATINGS (POPULARITY), AVERAGE RATING (ACCEPTANCE), GROSS (FINANCIAL SUCCESS)
Characterization: Movie Success

- Distribution of movie success
- Historical evolution of success distribution
- Correlation between different success metrics
HISTOGRAM OF MOVIE’S SUCCESS PARAMETERS

G1: TOP 10% MOVIES, G2: TOP 10—50% MOVIES, G3: ALL OTHER MOVIES
EVOLUTION OF MOVIE’S SUCCESS PARAMETERS DISTRIBUTION

ON TOP, HISTOGRAM OF MOVIE PRODUCTIONS COLOR CODED BY SUCCESS GROUP

EXPLOSION IN MOVIE PRODUCTION
MORE MOVIES WITH LOWER GROSS NOW
OLD MOVIES RECEIVE LESS VOTES
BIASED HIGHER RATINGS FOR OLD MOVIES
HEXAGONAL SCATTER PLOT BETWEEN SUCCESS PARAMETERS

DARKER BLUE SHADES REPRESENT HIGHER CONCENTRATION OF MOVIES

POPULAR MOVIES HAVE HIGHER GROSS

POPULAR MOVIES HAVE HIGHER RATINGS

MOVIES WITH HIGHER GROSS DEVIATE MORE FROM THE AVERAGE RATING
RESEARCH QUESTION:
(IN THE CONTEXT OF MOVIE PRODUCING TEAMS)

GIVEN DIFFERENT TEAMS THAT COULD PRODUCE A MOVIE, WHICH IS MORE LIKELY TO ACHIEVE SUCCESS?
Movie Characteristics

GENRES (21)

RUNTIME

PRODUCTION BUDGET (NORM.)

CONTINENTS (6)
Movie team Parameters:
Ego

# OF PAST EXPERIENCES
LEVEL OF PREVIOUS SUCCESS
IN-DEGREE
CLOSENESS
CLUSTERING COEFFICIENT
BETWEENNESS
NETWORK CONSTRAINT
SQUARE CLUSTERING COEFFICIENT
Movie team Parameters: Pairwise

- **Shared Friends**
- **Neighbour Overlap**
- **Shared Experience**
Movie team Parameters: Global

GLOBAL CLUSTERING COEFFICIENT

AVERAGE SHORTEST PATH

SMALL-WORLD-COEFFICIENT
Problem: many numbers from a single parameter

ARITHMETIC MEAN
HARMONIC MEAN
MEDIAN
STANDARD DEVIATION
MINIMUM VALUE
MAXIMUM VALUE

NODE CONTRACTION
NUMBER OF FEATURES FOR EACH MOVIE

70 EGO FEATURES 10 PARAMS. X 7 AGG. WAYS
27 MOVIE FEATURES 21 GENRES + 6 CONTINENTS
3 MOVIE PARAMS. RUNTIME, TEAM SIZE, BUDGET
3 GLOBAL METRICS Q, CLUSTERING, AVG. PATH LENGTH

121 TOTAL DISTINCT FEATURES
Characterization: Movie Teams’ Parameters

- Distribution of parameters
- Historical evolution of parameters
- Relation between success metrics and parameters
- Distribution of movies in pairs of characteristics
EVOLUTION OF MOVIE’S FEATURES DISTRIBUTION

PREV. VOTES, PREV. RATINGS, MOVIE RUNTIME

40’S–70’S: MOVIES WERE LONGER

EFFECT OF CHRONOLOGICALLY BIASED RATINGS

RECENT MOVIES: TEAMS THAT PREVIOUSLY PRODUCED POPULAR MOVIES

40’S–70’S: MOVIES WERE LONGER

EFFECT OF CHRONOLOGICALLY BIASED RATINGS

RECENT MOVIES: TEAMS THAT PREVIOUSLY PRODUCED POPULAR MOVIES

EVOLUTION OF MOVIE’S FEATURES DISTRIBUTION

PREV. VOTES, PREV. RATINGS, MOVIE RUNTIME
Mean: Gross (Billion USD)

Team metrics: Previous experience

EVOLUTION OF MOVIE’S FEATURES DISTRIBUTION
MEAN OF PREVIOUS GROSS, TEAM’S PREVIOUS EXPERIENCE
**Team metrics: Degree**

- 2000’s: Explosion of team’s degree

**Team metrics: Team size**

- 2000’s: Much higher # of producers per team

**Median: Closeness**

- Closeness: An evolving characteristic

**Harmonic mean: Clustering**

- Clustering: Fairly stable distribution

**DISTRIBUTION MOVIE’S FEATURES**
INTERACTION BETWEEN PAIRS OF FEATURES

DARK CLUSTERS SHOW CONCENTRATION OF BLOCKBUSTERS
A SUCCESSFUL, FEATURE LENGTH MOVIE CAN’T BE TOO SHORT

TEAMS WITH MODERATE PREVIOUS RATINGS PERFORM BETTER (!)

TEAMS THAT HAVE PRODUCED MORE POPULAR MOVIES BEFORE PERFORM BETTER

HISTOGRAM OF MOVIE’S PARAMS, PER SUCCESS GROUP
(a) Team metrics: Degree

(b) Team metrics: Network Constraint

(c) Team metrics: Team size

(d) Median: Closeness

(e) Harmonic mean: Clustering

Teams that have produced more money before perform better.

Teams with summed low experience perform badly.

Teams with low degree perform badly.

Socially unconstrained teams perform better.

Histogram of movie's params, per success group.
HISTOGRAM OF MOVIE’S PARAMS, PER SUCCESS GROUP

BEST PERFORMING TEAMS ARE NEITHER SMALL NOR BIG

TEAMS WITH LOWER CLOSENESS PERFORM WORSE

TEAMS WITH LOWER CLUSTERING (BY HARMONIC MEAN) PERFORM BETTER

(a) Team metrics: Degree

(b) Team metrics: Network Constraint

(c) Team metrics: Team size

(d) Median: Closeness

(e) Harmonic mean: Clustering

Team metrics: Team size

Median: Closeness

Harmonic mean: Clustering
Movie Success Forecast

- Movie Producing teams characteristics as features
- Movie success parameters as target variables
- Regressor: Bayesian Ridge (better to handle noise)
- Feature selection: eliminate features with less significance until model starts loosing accuracy
Feature selection

- Out of 121 features, 23 features were selected
- **19 Non-topological:** Genres (9), Continent (3), Runtime (1), Budget (1), Previous success (4), Previous Experience (1)
- **4 Topological:** Degree (1), Team Size (1), Closeness (1), Clustering (1)
Test $R^2$: 0.694
Baseline $R^2$: 0.399

IMPROVEMENTS IN PREDICTION ACCURACY WITH SOCIAL FEATURES
RED BARS REPRESENT ACCURACY GAINS IN THIS SAMPLE, RED BARS, LOSSES

× True value  – Baseline  + Test
Table 5.4: Coefficient of determination ($R^2$) and confidence interval for 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.

<table>
<thead>
<tr>
<th>Target</th>
<th>Years</th>
<th>Non Topol.</th>
<th>Topologic</th>
<th>All</th>
<th>Gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Votes</td>
<td>2008–2013</td>
<td>.529, ±.0008</td>
<td>.310, ±.0006</td>
<td>.556, ±.0008</td>
<td>5.10%</td>
</tr>
<tr>
<td></td>
<td>2000–2013</td>
<td>.484, ±.0004</td>
<td>.294, ±.0005</td>
<td>.517, ±.0004</td>
<td>6.82%</td>
</tr>
<tr>
<td></td>
<td>1990–2013</td>
<td>.437, ±.0003</td>
<td>.246, ±.0004</td>
<td>.464, ±.0003</td>
<td>6.18%</td>
</tr>
<tr>
<td>Gross</td>
<td>2008–2013</td>
<td>.431, ±.0008</td>
<td>.170, ±.0013</td>
<td>.448, ±.0009</td>
<td>3.94%</td>
</tr>
<tr>
<td></td>
<td>2000–2013</td>
<td>.419, ±.0004</td>
<td>.175, ±.0005</td>
<td>.447, ±.0004</td>
<td>6.68%</td>
</tr>
<tr>
<td></td>
<td>1990–2013</td>
<td>.392, ±.0004</td>
<td>.174, ±.0004</td>
<td>.435, ±.0003</td>
<td>10.97%</td>
</tr>
<tr>
<td>Rating</td>
<td>2008–2013</td>
<td>.271, ±.0011</td>
<td>.033, ±.0009</td>
<td>.281, ±.0012</td>
<td>3.69%</td>
</tr>
<tr>
<td></td>
<td>2000–2013</td>
<td>.267, ±.0006</td>
<td>.038, ±.0003</td>
<td>.273, ±.0006</td>
<td>3.37%</td>
</tr>
<tr>
<td></td>
<td>1990–2013</td>
<td>.258, ±.0004</td>
<td>.031, ±.0003</td>
<td>.262, ±.0005</td>
<td>1.55%</td>
</tr>
</tbody>
</table>
Contributions

- Improvement to the state-of-the-art in movie success forecasting
- In-depth characterization of social aspects of a large collaborative network
- Presented a new approach for extensive aggregation of social metrics from agents in teams
THIS IS ONLY A FIRST LOOK IN HOW NETWORK TOPOLOGY ANALYSIS CAN HELP EXPLAIN COMPLEX HUMAN BEHAVIOR.