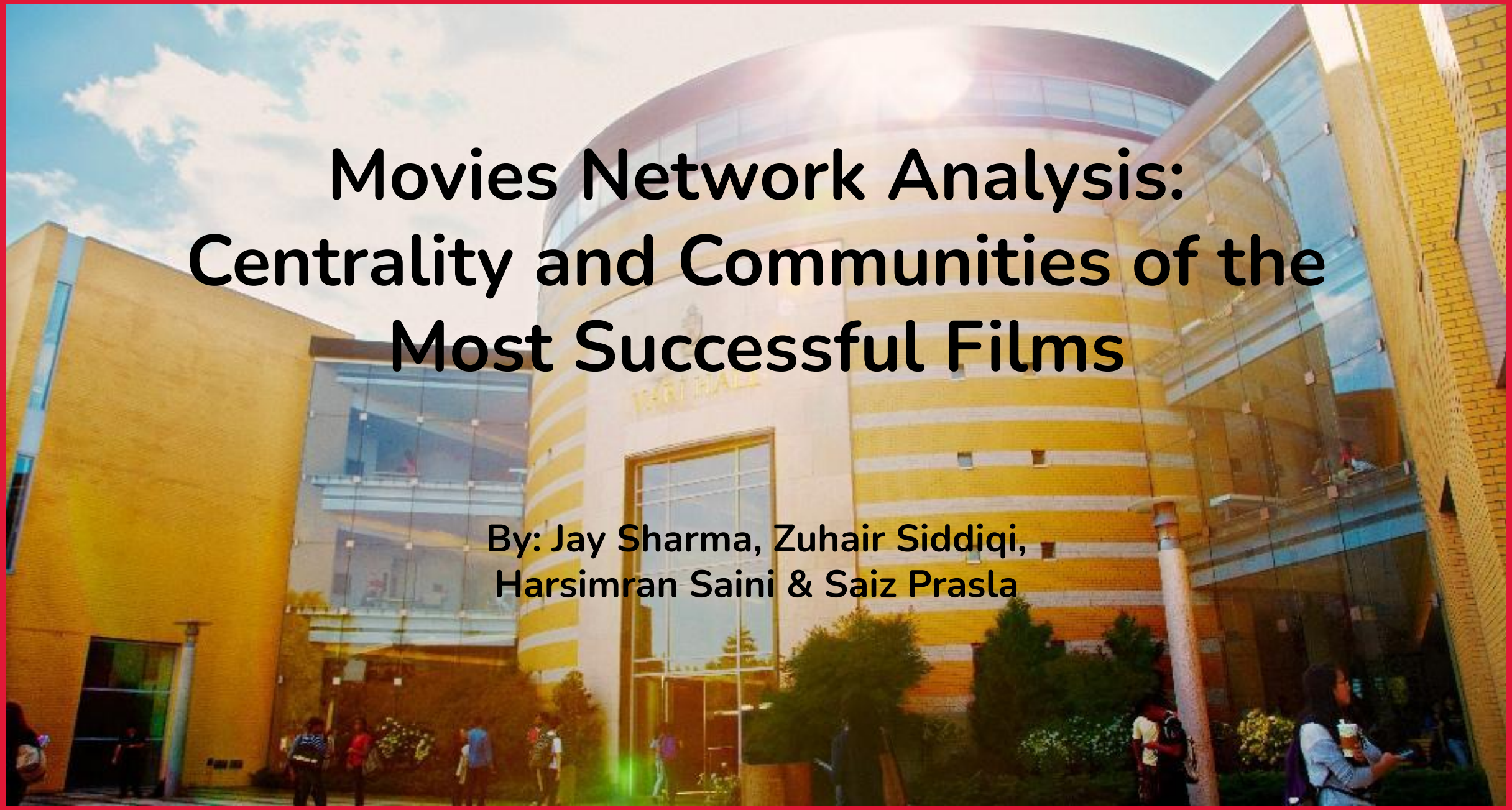


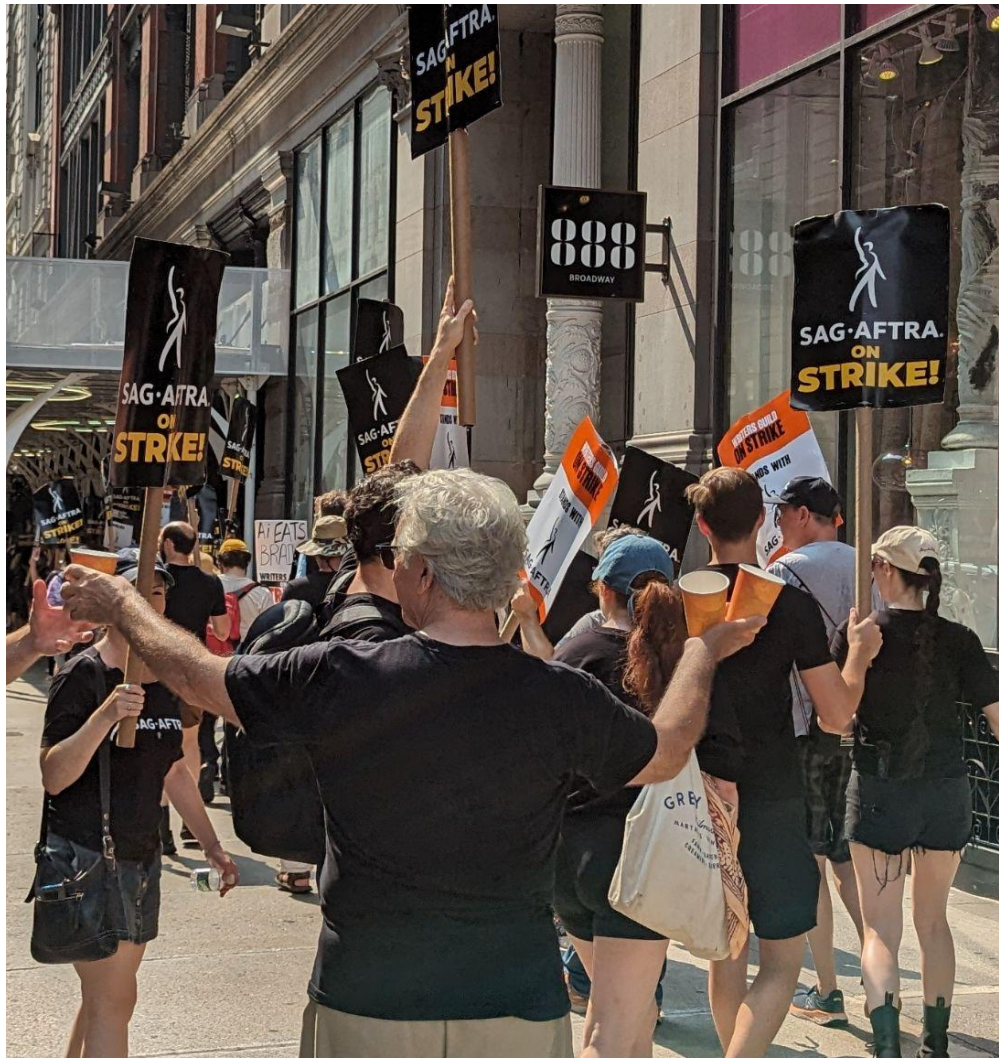
Movies Network Analysis: Centrality and Communities of the Most Successful Films

By: Jay Sharma, Zuhair Siddiqi, Harsimran Saini & Saiz Prasla

Outline

Introduction

- Our focus for this project lies within finding out what determines and results in a successful film
- With COVID-19, the rise of streaming services and the SAG-AFTRA strikes, the film industry has seen a decline in terms of theatre performance (seen in aspects such as box office revenue, cancellations/delays)
- Looking into methods of analyzing successful films and how to repeat past successes



YORK U

Problem Definition

Problem 1

Given a dataset of films, what similarities and structure can we find to connect the films?

Problem 2

What factors can we use to measure each film and rank them?

Problem 3

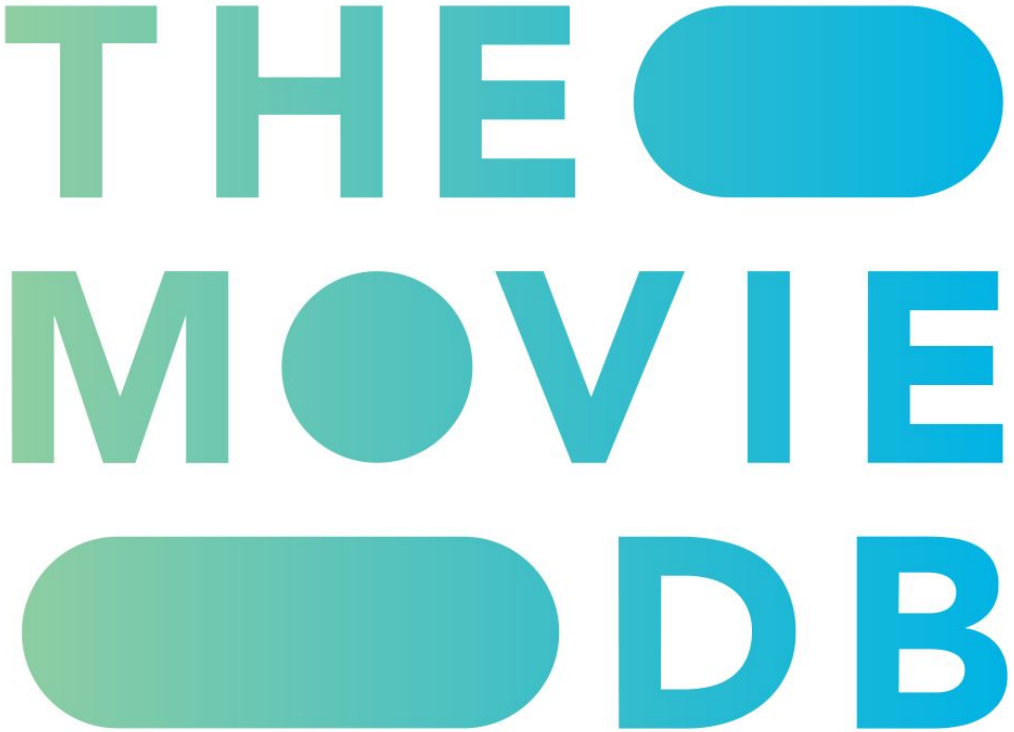
What analysis and evaluations can we do to measure success and see what films to replicate?

Datasets

- Used the TMDb API to get the dataset of the top 10000 top rated films of all time

- Got details of the films including:
  - Budget - How much the film costed to produce
  - Genres - What genres are attributed to the film
  - Actors - Who acted in the film
  - Production Companies - What studios produced the film
  - Viewer Rating - The rating of the film
  - Revenue - How much the film made at the Box Office
  - Series - Whether or not the film is a part of a series
  - Release Year - The year the film released

- Removed films without necessary data including films with no revenue or budget



Graph Creation

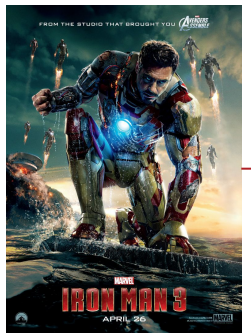
Node Creation:

- Introduced **success score** for each node, combining profit and rating
- Ratio favored low-budget films, leading to skewed scores, thus **normalized profit** and rating between 0 and 1 to address bias
- Used MinMaxScaler() for automated normalization
- **Equally weighted profit and rating** for balanced assessment

Edge Creation:

- Determine edge weights based on **shared attributes**
- Takes two nodes & outputs weighted edges
- Identifies common attributes (actors and production companies); computes edge weight if criteria met
- Assigns **weights** based on attribute **importance**
- Iteratively refines to optimize graph relations

5.47 (score)



7 (weight)

5.83 (score)

YORK U

PageRank

- Used NetworkX’s pagerank(G, personalization = “success”) function,
  ➜ G is movie graph
  ➜ personalization is the success score of each node

- PageRank works by ranking nodes based on the structure of links that are incoming

  ➜it goes under the assumption that important nodes are linked by other important nodes.

- Personalized PageRank is a variation of PageRank that allows for a personalization vector that adds bias to nodes with higher personalization scores

  ➜ in our case, "success" scores

Node Degree Centrality

- Used NetworkX's degree_centrality(G) function,
  ➜ G is movie graph

- Node degree centrality was used alongside PageRank to measure the nodes without bias of success score

- Measures nodes based on number of connections

- Films with a lot of high profile actors who are in many big films would score higher, since there would be many connections

Louvain

- Used NetworkX's built in function louvain_communities(G, weight)
  ➜ detect communities within our graph while taking edge weights into account

- Incorporating these weights ensured stronger connections which lead to more cohesive and meaningful clusters
  ➜ actors who frequently collaborated were given more significance in the detection process

- Louvain was selected due to its scalability compared to Girvan Newman
  ➜ Louvain time complexity = O(n log n)
  ➜ Girvan Newman time complexity = $O(n^3)$

- We ran initially Girvan Newman, however it led to a very large community with over 4000 nodes, and then many small communities
  ➜ Louvain gave us an even spread of communities with similar sizes

Experiments

- The network consisted of **5,755 nodes** and **13,998 edges**, representing a sizable subset of the initial dataset
- Evaluation involved various algorithms including PageRank, node degree centrality, and Louvain community detection
- Evaluation also included analysis of **communities** with the highest average success scores, reflecting different attributes
- Similarities in top 5 PageRank and Degree centrality shows graphically dominant films g

Top 5 Films PageRank

Films	PageRank Score
Harry Potter and the Order of the Phoenix	0.0013234276868831663
Avengers: Infinity War	0.0012905504009495543
Avengers: Endgame	0.0012624523143546574
Harry Potter and the Deathly Hallows: Part 1	0.0011429170249350508
Avengers: Age of Ultron	0.001110811313672943

Top 5 Films Degree Centrality

Films	PageRank Score
Harry Potter and the Order of the Phoenix	0.007994438651372959
The Lego Movie	0.007820646506777894
Sausage Party	0.0076468543621828295
Avengers: Endgame	0.0074730622175877654
Avengers: Infinity War	0.0074730622175877654

Lord of the Rings Community

Top 5 Films in community by success score:

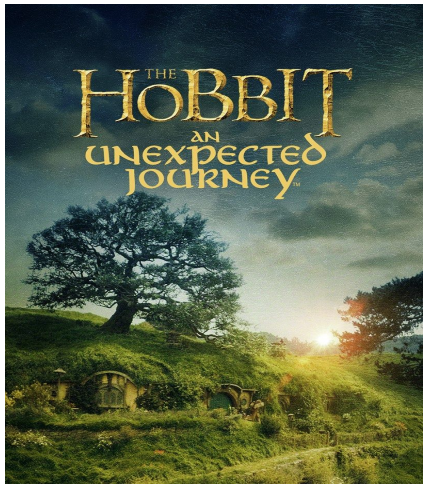
6.93

6.53

6.43

5.53

5.44



Most Common actors:
David Wenham: 12 appearances
Hugo Weaving: 11 appearances
Brad Dourif: 10 appearances

Most Common Genres:
Drama: 50 films
Action: 38 films
Thriller: 35 films

Size of Community: 436 films
Percentage of series: 33%
Average release year: 2006
Average success score: 3.560

YORK U

Classic Sci-Fi/Inspirational Community

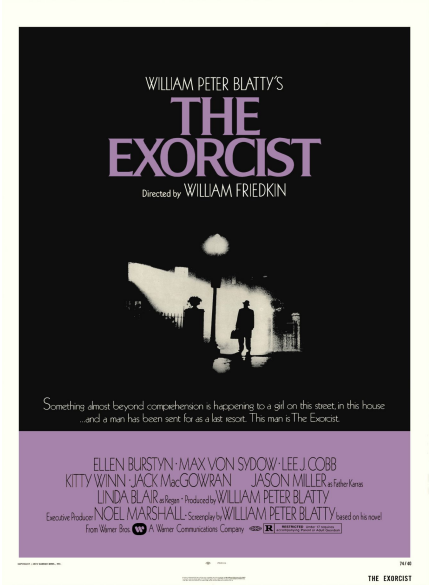
Top 5 Films in community by success score:

5.62

5.51

5.20

5.19



5.18

Most common actors:
Clint Eastwood: 21 appearances
Mel Gibson: 12 appearances
Arnold Schwarzenegger: 10 appearances

Most Common Genres:
Drama: 55
Thriller: 55
Action: 49

Size of Community: 149 films
Percentage of Series: 34%
Average Release Year: 1980
Average Success Score: 3.710

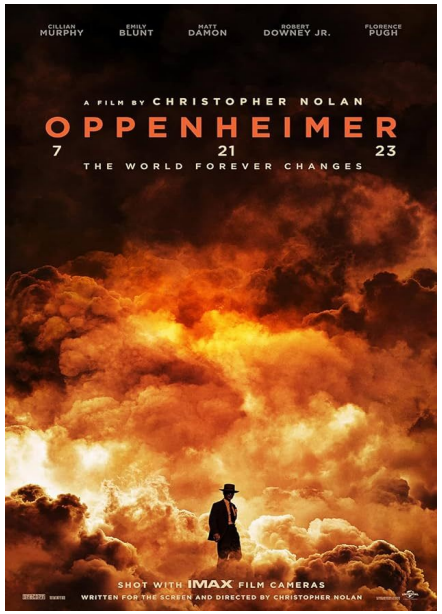
YORK U

Nolan/Modern Classics Community

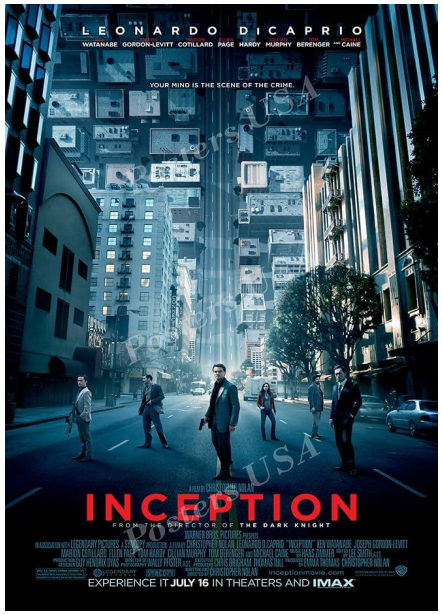
Top 5 films in community by success score:

6.5916

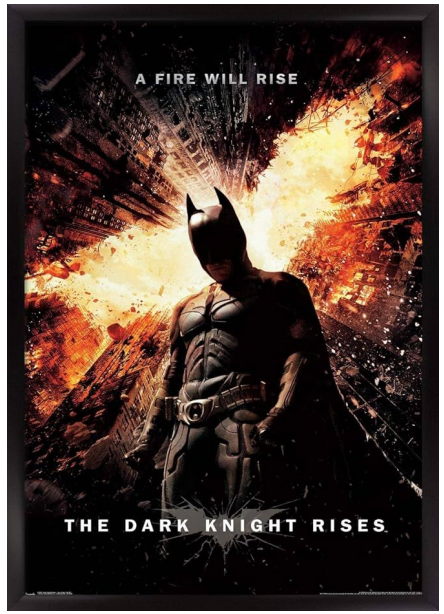
6.2587



6.1886



5.9345



5.4153

Most Common Actors:
Morgan Freeman: 14 appearances
Penelope Cruz: 12 appearances
Michael Caine: 11 appearances

Most Common Genres:
Drama: 96
Thriller: 82
Action: 53

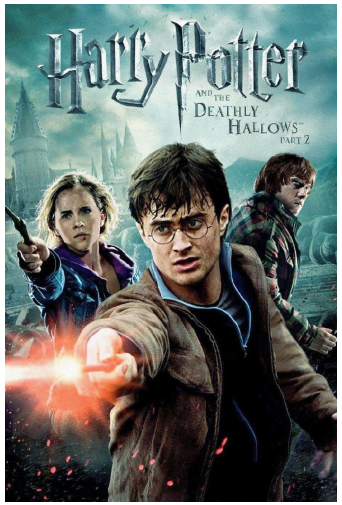
Size of Community: 173 films
Percentage of Series: 20%
Average Release Year: 2011
Average Success Score: 3.734

YORK U

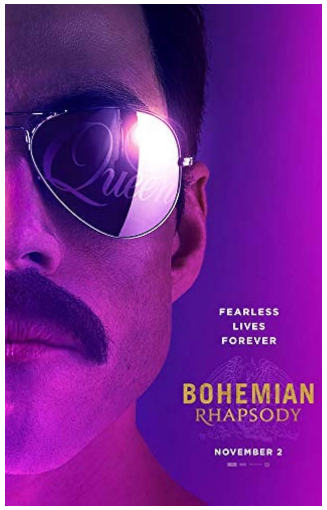
Harry Potter Community

Top 5 Films in community by success score:

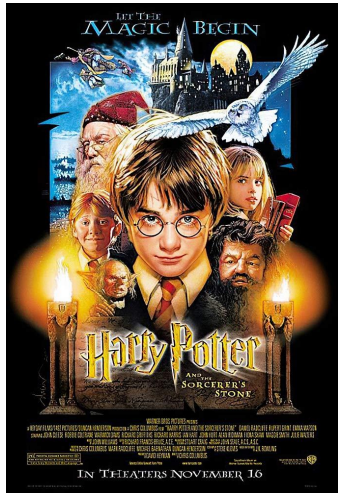
6.89



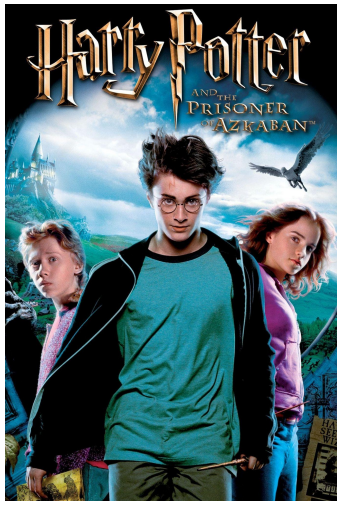
6.19



6.09



5.86



5.82

Most common actors:
Johnny Depp: 25 appearances
Colin Firth: 23 appearances
Jim Broadbent: 22 appearances

Most common genres:
Drama: 246 films
Action: 122 films
Thriller: 114 films

Size of Community: 436 films
Percentage of series: 20%
Average release year: 2004
Average success score: 3.751

YORK U

Marvel Cinematic Universe Community

Top 5 Films in community by success score:

9.17

7.92

7.68

5.86

5.83

Most common actors:
Samuel L. Jackson: 16 appearances
Anthony Mackie: 14 appearances
Idris Elba: 13 appearances

Most common genres:
Action: 70 films
Drama: 66 films
Adventure: 61 films

Size of Community: 171 films
Percentage of series: 34%
Average release year: 2009
Average success score: 3.762

YORK U

Visualizations

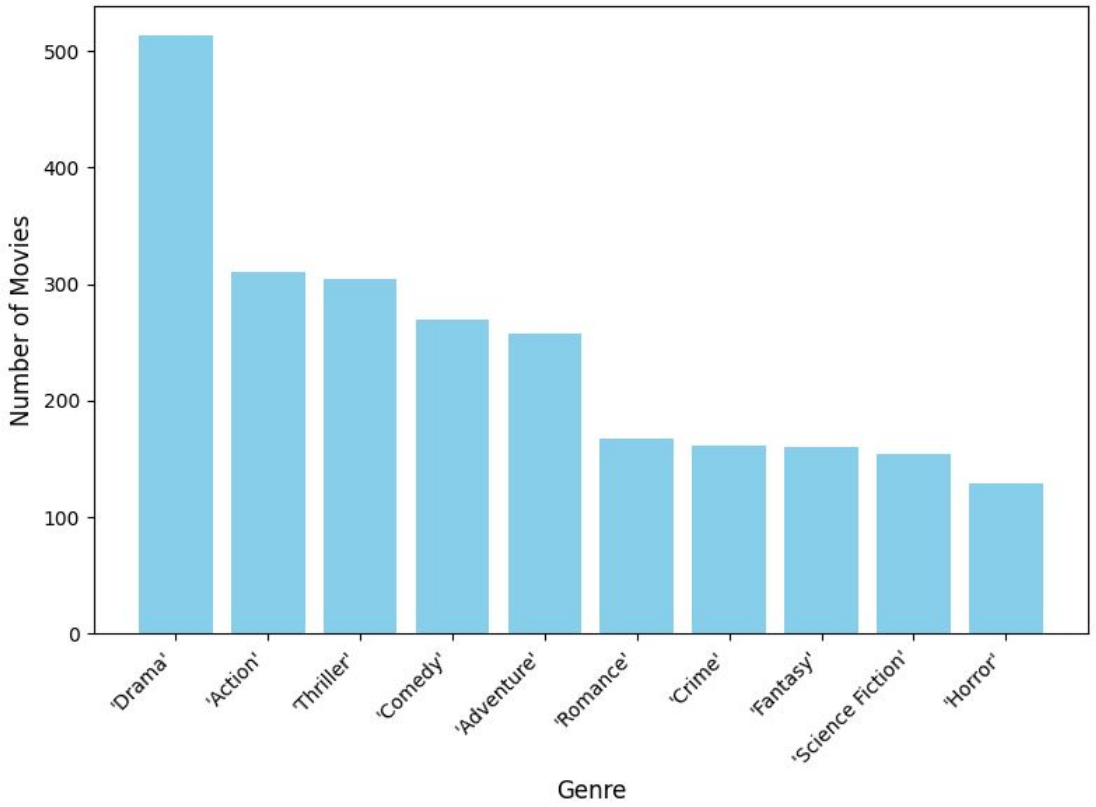


Figure 1: Top 10 genres in the top 5 communities by number of movies

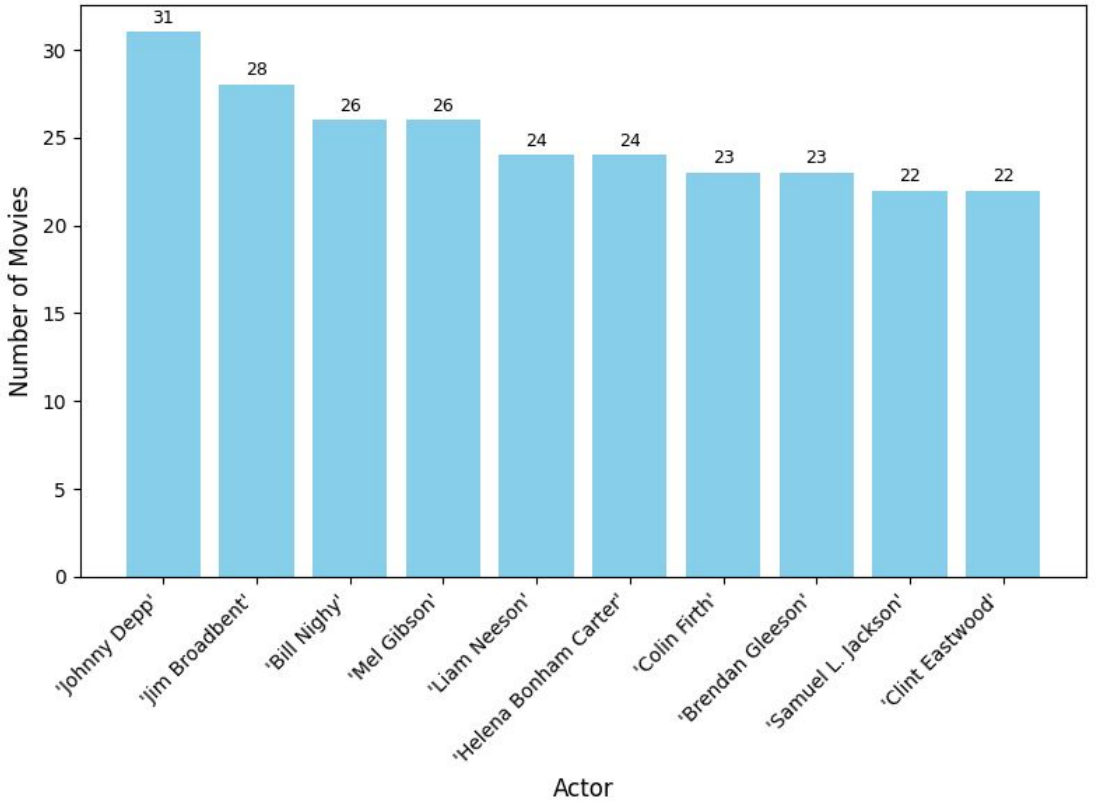


Figure 2: Top 10 actors in the top 5 communities by number of movies

YORK U

Visualizations

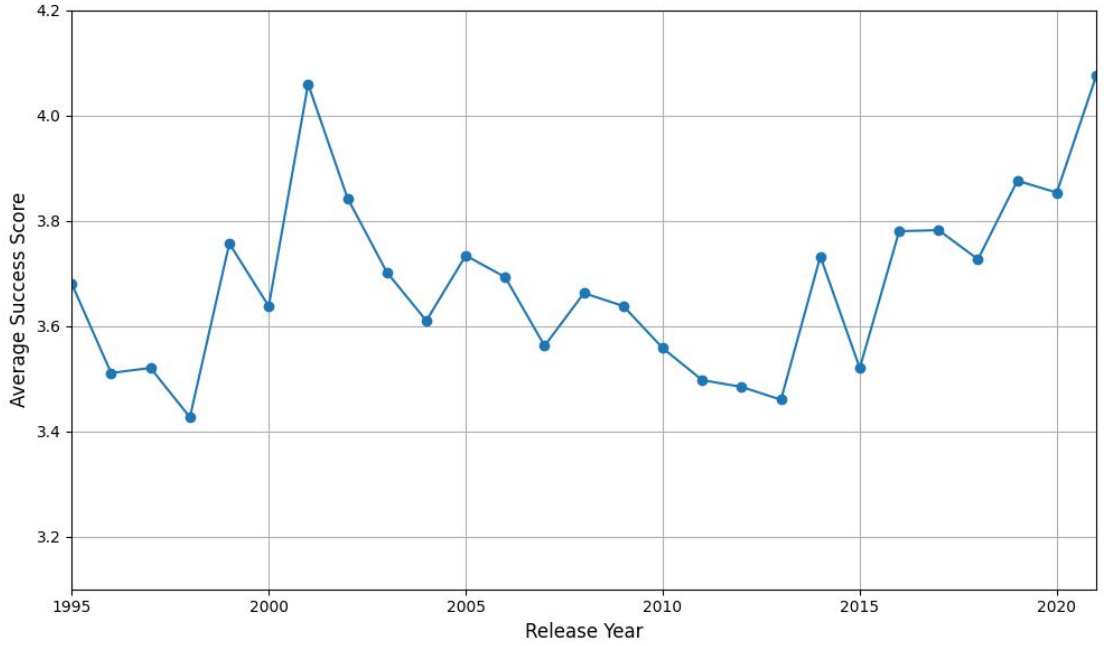


Figure 3: Average success score for movies in the top 5 communities over time (1995-2022)

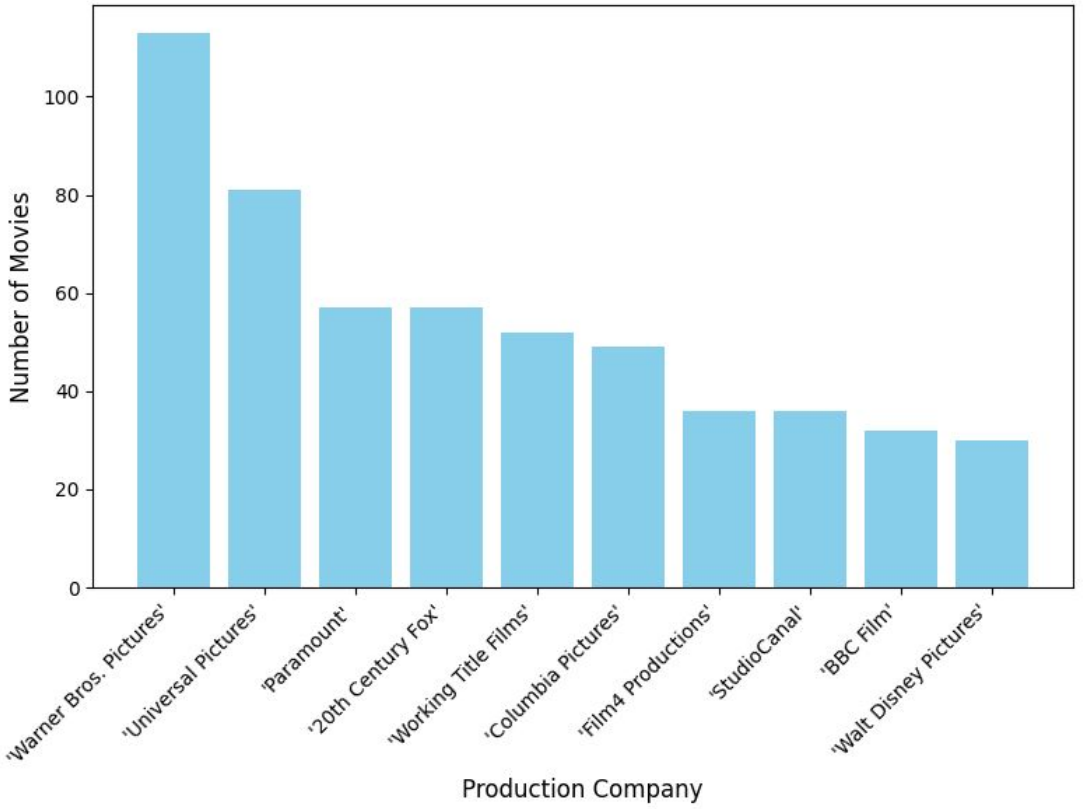


Figure 4: Top 10 production companies in the top 5 communities by number of movies

YORK U

Analysis

1	Top films with **PageRank** are **successful**, notably Avengers titles and MCU community stands out
2	Communities like Lord of the Rings and Harry Potter have drama and action genres
3	**Community analysis** reveals commonality in genres and actors
4	Statistical analysis identifies investment opportunities, highlighting **casting impact**
5	Metrics show varying success, with Marvel Cinematic Universe leading
6	Major studios like Warner Bros. and Universal are prominent
7	**Temporal bias** in release years indicates dominance of newer movies
8	Disney lags due to animated films in user ratings and box office performance
9	**Recurring themes** and **actors** increase film success

Conclusions

- We have constructed our network of films with each film being a movie and each edge being weighted with common actors.
- We developed a scoring system for the success of each film that uses a normalized profit and rating to attribute each node with a success score
- We used different algorithms such as PageRank, Node degree centrality, and the Louvain community detection algorithm on our graph to see some commonalities and correlations with our various scoring mechanisms and common films.
- Possible future works could be based on the fact there is the aforementioned temporal bias, with newer films scoring better on average in most metrics relative to older films
  ➜ This discrepancy could be attributed to changes in currency inflation or advancements in film technology and to address this bias is to incorporate currency inflation calculations
- Sorting this inconvenience out would help see a pattern as to what continues to be successful films across decades regardless of the vast differences
  ➜ film-making techniques
  ➜ technological advancements
  ➜ growth of film culture.

Image Sources

- Kristis Luhaers. (n.d.). *Person watching movie* [Photograph]. Unsplash. https://unsplash.com/photos/AtPWnYNDJnM
- By Eden, Janine and Jim from New York City - SAG-AFTRA Picket I, CC BY 2.0, https://commons.wikimedia.org/w/index.php?curid=134582496
- https://www.themoviedb.org/about/logos-attribution
- https://www.themoviedb.org/

Thank you!

Questions?