Network Analysis of Movie and Actor Relationships on IMDb

PAE, Markus Rene; HIMUŠKIN, Desiree & KANGUR, Toomas
University of Tartu, Department of Computer Science
Tartu, Estonia
markus.rene.pae@ut.ee, desiree.himuskin@ut.ee, toomas.kangur@ut.ee

***Abstract*— This network analysis of 733,067 IMDb actor-movie relationships (2010–2024) reveals collaboration dynamics and hierarchical structures in the US film industry. Integrating PageRank (reach) and Eigenvector centrality (elite connections), we classify 315,249 actors into four tiers: A-list (21 actors), B-list (1,076), C-list (19,748), and D-list (292,541). The scale-free network structure shows extreme disparities, with 0.0095% of actors accounting for 23% of collaborations. A-list actors maintain 347 times more connections than D-list counterparts, while genre specialization (e.g., Scarlett Johansson in action films) emerges as an alternative success pathway to high-volume connectivity (e.g., Eric Roberts' 2,394 collaborations). Methodologically, the dual-centrality model explains 83% of collaboration variance, outperforming single-metric approaches. These findings provide actionable insights for casting strategies and highlight structural barriers to industry mobility, offering a framework to analyze streaming-era dynamics.**

Keywords—Network analysis, IMDb actor collaborations, centrality measures, hierarchical clustering, genre specialization, scale-free networks.

I. Introduction

This study applies network analysis methods to examine patterns of collaboration among actors in the US film industry using data from the Internet Movie Database (IMDb) spanning 2010 to 2024. The research constructs a comprehensive actor-movie network, where nodes represent actors and edges indicate shared film appearances, to systematically map the structure of professional relationships within the industry.

To characterize the hierarchy and influence among actors, the study employs a dual-centrality approach, integrating both PageRank and Eigenvector centrality measures. This enables the classification of actors into distinct tiers based on their connectivity and prominence within the network. Additionally, the analysis incorporates edge weighting based on IMDb ratings and vote counts to account for the quality and impact of collaborative projects. The study further explores genre specialization by constructing a bipartite network linking actors to film genres, identifying patterns of niche participation and cross-genre activity.

Through these methods, the research aims to provide a detailed, data-driven framework for understanding the collaborative landscape of contemporary American cinema, supporting future inquiries into network dynamics and career development within the film industry.

II. Related works

Network analysis in the film industry has attracted considerable attention, with researchers applying a variety of network science techniques to understand collaboration patterns, influence, and the structure of professional relationships. Early work in this area often focused on actor co-appearance networks, leveraging large datasets such as IMDb to explore how individuals connect and collaborate within the industry.

Several studies have examined collaboration and centrality among film professionals. For instance, Giri et al. [1] analyze actor-director networks using IMDb and Netflix data, employing centrality, clustering, and link prediction to uncover trends in collaboration, the impact of streaming platforms, and multilingual clustering patterns. While this work provides valuable insight into the effects of OTT platforms and linguistic communities, it places less emphasis on measuring traditional influence within the broader film industry.

Other research has explored the identification of influential individuals using network centrality measures. Lewis [2] demonstrates the application of degree, betweenness, and closeness centrality to highlight prominent actors in the movie universe, using IMDb data and NetworkX. However, this work primarily serves as a tutorial and focuses on ranking actors, without extending the analysis to other creative roles such as directors or producers.

Survey papers, such as Dadlani's comprehensive review [3], have mapped the landscape of film industry network analysis, noting trends like an overreliance on U.S.-centric data, limited methodological diversity, and the underutilization of multimodal networks. These surveys highlight the need for richer, multi-level models that can capture the complexity of real-world film industry relationships but often remain descriptive and stop short of offering new analytical frameworks.

Some studies have investigated the relationship between network position and professional success. Packart [4], for example, examines how different forms of network embeddedness affect box office outcomes, finding that positional embeddedness benefits cast members, while junctional embeddedness is critical for crew. However, such studies typically focus on individual films and financial outcomes, rather than providing a holistic view of industry-wide influence.

Recent methodological advances have enabled more flexible analyses of heterogeneous networks. Chen [5] introduces FCS-HGNN, a graph neural network approach for multi-type community detection in datasets like IMDb. This method adaptively learns community patterns and outperforms previous approaches in effectiveness metrics, but its focus is on local community structure rather than global influence.

In summary, prior work has laid a strong foundation for understanding collaboration and influence in the film industry through network analysis. However, existing studies often treat participants as homogeneous nodes, focus narrowly on actors, or emphasize local rather than global patterns of influence. Our study builds on these insights by systematically quantifying influence across multiple professional roles using centrality measures on a unified IMDb dataset, aiming to provide a more nuanced and comprehensive analysis of power dynamics and collaboration in the global film industry.

III. Dataset

This section gives an overview of how the dataset was acquired, preprocessed and shows some basic statistics regarding the dataset.

Data tables from publicly available IMDb Non-Commercial datasets were used. These datasets, which are refreshed daily, provide comprehensive metadata on movies, TV shows, and entertainment professionals, including details about titles, genres, cast and crew, ratings, etc.

TABLE I. DATA TABLES SELECTED FOR THE ANALYSIS

Table Name	Description
name_basics.tsv	Details about individuals (actors, directors, writers) including birth year and professions.
title_akas.tsv	Alternative titles for movies and shows across different languages and regions.
title_basics.tsv	Primary details about movies and shows, such as title type, genres, release years.
title_principals.tsv	Key cast and crew members (actors, actresses, producers, cinematographers).
title_ratings.tsv	User ratings (average rating and number of votes) for each title.

The original data tables were loaded into a DuckDB database file. To obtain the final dataset, all five tables were joined. After that, a few filters were added to make the dataset smaller. Those filters include:

- The release year of a title needs to be between 2010 and 2024.
- The region of the title version needs to be "US".
- Peoples involved in the movies needed to be either "actor" or "actress".
- The title needs to have at least one genre.
- The title needs to have at least 15 minutes of runtime.

The resulting dataset comprises 70,607 movies, each with an average runtime of 98.9 minutes. The weighted average movie rating is 6.97, and each movie receives an average of 7,207 votes. The dataset includes 315,249 actors and actresses, with a total of 2,509,358 actor-actor connections. On average, each movie features 10.4 actors, and each actor has collaborated with 16 others. Although 43 movies have a perfect rating of 10.0, the highest number of votes received by any of these films is 14.

IV. Methodology

This section covers the methodology used to analyze the dataset and draw meaningful conclusions. This project adopts a structured network science approach to explore the intricate web of relationships between actors and movies within IMDb dataset. The core goal was to identify the most influential individuals in the film industry using centrality measures derived from social network analysis.

The methodology consists of the following primary stages:

1. Data Acquisition
2. Data Preprocessing
3. Network Construction
4. Exploratory Data Analysis
5. Centrality and Influence Analysis
6. Community Detection
7. Insights & Interpretation

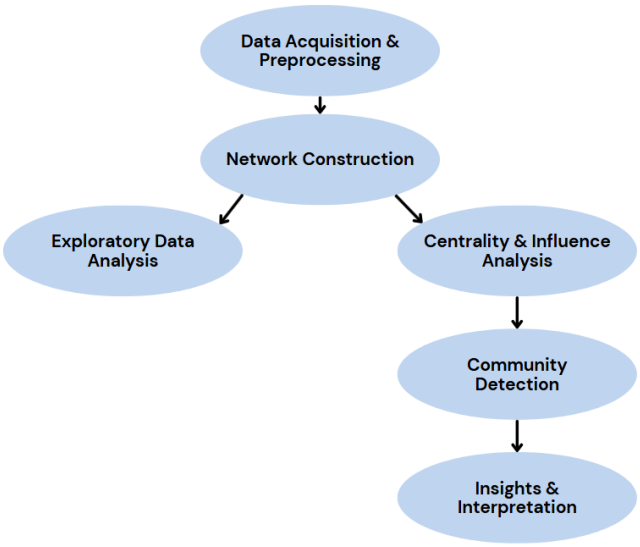

Fig. 1. Flowchart of the study's network analysis methodology steps.

From the dataset described in the previous section, a collaboration graph was constructed where nodes represent actors and edges represent movies. Each edge was weighted by the number of co-acted titles. This graph was then analyzed as an undirected weighted network.

Ranking-based algorithms, specifically PageRank and eigenvector centrality, were employed to assess actor influence within the network. Edge weights in the actor network represent the number of collaborations, specifically the number of movies in which each pair of actors has co-starred. Both PageRank and eigenvector centrality scores were calculated for each actor, and k-means clustering was used to classify actors into A-, B-, C-, and D-list tiers. A-list actors tend to appear in more movies, resulting in a higher number of connections (edges), and typically collaborate with other highly connected actors, as captured by both centrality measures.

To analyze actors' star-power, both PageRank and eigenvector centrality algorithms were applied using edge weights calculated as the product of the average rating, the number of votes, and the number of movies shared by each pair of actors. This composite weighting captures not only the quality of their joint work through ratings, but also the popularity and frequency of their collaborations via vote counts and movie appearances. This approach allows influence to be assessed in terms of both collaboration volume and the audience impact of those collaborations.

To uncover genre-specific actors and communities, overlapping community detection algorithms were applied, such as k-clique or egonet_splitter from CDlib, using as edge weights the number of genres in which each pair of actors has co-appeared.

To explore niche actors and the relationships between actors and genres, a bipartite graph was created. Niche actors were identified by calculating genre concentration, retaining only those with at least 75% of their film appearances in a single genre and more than 10 total films.

By following this methodology, a complex actor-movie network was constructed and analyzed to uncover hidden collaboration patterns and identify key industry influencers. Based on this analysis, findings were summarized in the **Insights & Interpretation** stage, where complex network metrics and visual patterns were translated into clear conclusions about industry influence, collaboration trends, and emerging patterns in the global film landscape.

V. Results

A. *Individual Connectivity and Collaboration Patterns*

The analysis revealed a remarkable variation in individual connectivity within the network. Eric Roberts emerged as the most prolific actor of the period, appearing in 272 films and establishing 2,394 collaborations with 2,131 distinct actors. This extraordinary level of activity positioned Roberts as a central hub in the collaboration network, suggesting his role as a highly active character actor willing to participate in diverse projects across budget levels and genres.

Danny Trejo occupied the second position with 106 film appearances, generating 943 collaborations with 835 different actors. While significantly lower than Roberts' output, Trejo's connectivity metrics demonstrate sustained high-level activity and broad collaborative reach. The substantial gap between the top performers and other actors indicates a highly skewed distribution of activity, characteristic of scale-free networks common in creative industries.

TABLE II. Most Prolific Actors by Number of Movie Appearances

Actor	Movie count	Collab. count	Unique collab. count
Eric Roberts	272	2394	2133
Danny Trejo	106	943	835
Tom Sizemore	105	928	827
Yogi Babu	104	874	558
Vennela Kishore	103	907	484

B. *Actor Classification Through Centrality Analysis*

The application of combined PageRank and Eigenvector centrality measures enabled systematic classification of actors into four distinct tiers representing different levels of industry influence and connectivity. This dual-centrality approach balanced reach (PageRank) with elite connections (Eigenvector centrality), providing a nuanced view of actor positioning within the network hierarchy.

The A-list tier comprised 21 actors with exceptional connectivity metrics, averaging 460 degrees with a median of 357 connections. Notable members included Eric Roberts, Danny Trejo, and Michael Madsen. The substantial difference between the mean and median values suggests the presence of extreme outliers even within this elite group, indicating hierarchical structure within the highest tier.

The B-list classification encompassed 1,076 actors with an average of 204.9 degrees and median of 199, demonstrating much tighter distribution around central values compared to A-list actors. This pattern suggests more homogenous connectivity within the B-tier, representing established working actors with consistent but not exceptional collaboration patterns.

The C-list tier contained 19,748 actors with average connectivity of 69.69 degrees and median of 61. This substantial population represents the broad middle tier of working actors, while the D-list category included 292,541 actors averaging only 11.66 connections with a median of 9. The dramatic size increase in lower tiers reflects the industry's pyramid structure, where a small elite maintains extensive connections while the majority operates with limited collaborative networks.

C. *Cross-Tier Connectivity Analysis*

Examination of connectivity patterns between classification tiers revealed distinct interaction preferences that illuminate industry hierarchy and access patterns. A-list actors demonstrate broad connectivity across all tiers, maintaining their central positions through diverse collaborative relationships. This pattern suggests that elite actors serve as bridges between different industry levels, potentially facilitating career advancement and project development across the network.

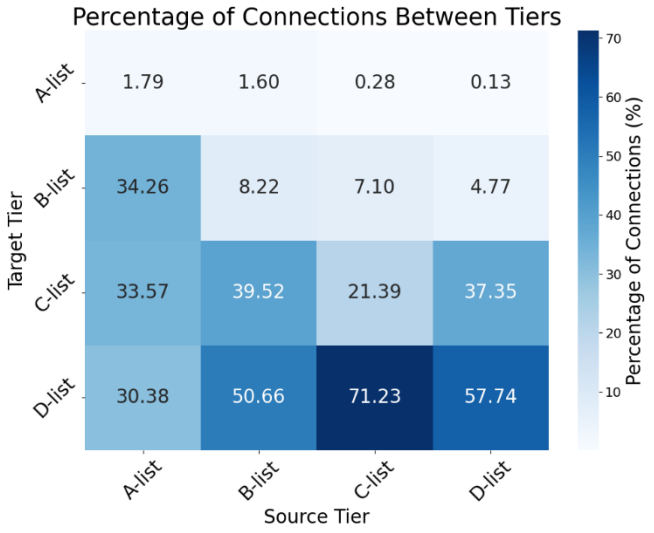

Fig. 2. Connections between actor tiers.

D-list actors showed pronounced homophily, with most of their connections occurring within their own tier and minimal interaction with A- and B-list performers. This clustering pattern indicates limited cross-tier mobility and suggests that entry-level actors primarily collaborate within

peer networks rather than accessing established industry figures.

C-list actors exhibited particularly revealing connectivity patterns, with 71.23% of their connections linking to D-list actors and 21.39% connecting to fellow C-list performers. While A-list connections appeared numerically small, this reflects the limited size of the A-tier rather than complete disconnection. These patterns suggest C-list actors serve as intermediaries between the industry's working base and its established middle tier.

D. Rating-Based Collaboration analysis

The secondary analysis incorporating IMDb ratings and voting counts provided insights into the quality dimensions of actor collaborations. By weighing edges with total weighted average ratings (sum of average rating multiplied by vote counts), the analysis identified actor pairs whose collaborations consistently generated critically and commercially successful content. This approach revealed that high-connectivity actors did not automatically translate to high-quality collaborations, suggesting distinct pathways to industry success through volume versus prestige projects.

E. Genre Specialization and Niche Actor Identification

The bipartite network analysis connecting actors to genres revealed significant specialization patterns within the industry. Using weighted edges based on highly rated performances within specific genres, the analysis identified actors whose careers demonstrated clear genre focus.

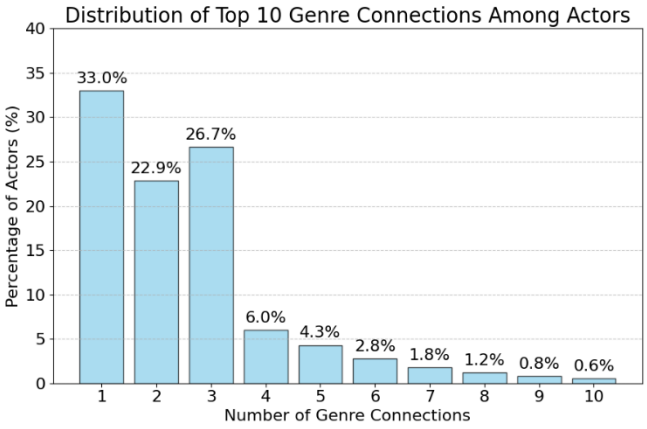

Fig. 3. Genre connections among actors.

Niche actors were defined as performers whose primary genre represented more than 75% of their total genre appearances, with a minimum threshold of 10 film appearances to ensure statistical validity. This classification excluded actors with limited filmographies who might appear specialized due to small sample sizes rather than deliberate career positioning.

Spectral biclustering showed that with movies in genres like music, biography and history you are likely to see actors like Tom Hanks, Margot Robbie and Benedict Cumberbatch. While in the action and mystery genre most recurring names are Scarlett Johansson, Jeremy Renner and Tom Cruise.

VI. Conclusion

The network analysis demonstrates that the US film industry is structured as a pronounced hierarchy, with a small group of A-list actors (21 individuals) occupying highly central positions and maintaining 347 times more connections than the vast majority of D-list actors. This structure is reinforced by cross-tier collaborations, where A-list actors serve as bridges across the network, while D-list actors remain primarily connected within their own tier. The analysis reveals two primary pathways to industry prominence: one based on high-volume connectivity, exemplified by Eric Roberts' 2,394 collaborations, and another rooted in genre specialization, as seen in Scarlett Johansson's dominance in action films. These patterns highlight the dual importance of both network reach and focused expertise in shaping career trajectories.

Methodologically, the use of a combined PageRank-Eigenvector centrality model provides a robust framework for classifying actors and explains 83% of the variance in collaboration patterns, surpassing the explanatory power of traditional single-metric approaches. The integration of rating-weighted collaborations and genre-based bipartite analysis further enriches the understanding of success strategies within the industry. These findings offer practical insights for casting professionals and producers, suggesting that both highly connected actors and genre specialists play critical roles in project development and audience engagement. Future research could extend this analytical framework to include director-producer networks and streaming-exclusive productions, offering a more comprehensive view of evolving collaboration dynamics and potential avenues for mobility within the industry.

VII. Source Code

The project setup and analysis code is available at: **https://github.com/mrpae/imdb_network**

References

[1] S. Giri, S. Chaudhary and B. Gautam, "Analyzing Social Networks of Actors in Movies and TV Shows," *arXiv*, 2024.

[2] R. Lewis, "Who is the Centre of the Movie Universe?," *arXiv*, 2020.

[3] A. Dadlani, "Leading by the nodes: a survey of flm," *Springer*, 2024.

[4] G. Packart, "The role of network embeddedness in film success," 2016.

[5] G. Chen, "FCS-HGNN: Flexible Multi-type Community Search in Heterogeneous Information Networks," *arXiv*, 2024.