Network Analysis within the Movie Industry

Yuri Noviello, Artificial Intelligence, ID number (0001038692)
Francesco Olivo, Artificial Intelligence, ID number (0001036865)

1 Introduction

The film industry is a multifaceted domain that has continually captivated global audiences through its storytelling ability. Over time, the film industry has undergone significant transformations, not only in creative aspects but also in terms of its business dynamics, distribution methodologies, and audience engagement strategies. With the rise of digital platforms and the internet, the film industry has entered a new era characterized by unprecedented connectivity and information exchange. In this context, Social Network Analysis (SNA) emerges as a potent analytical tool for unraveling the intricate web of relationships, influences, and trends within the industry.

1.1 The Movie Industry: A Dynamic Ecosystem

The movie industry represents a dynamic ecosystem, incorporating a multitude of stakeholders, including filmmakers, actors, producers, distributors, critics, and, most significantly, the audience. It is an industry propelled by creativity, innovation, and collaboration, where success often hinges on the capacity to forge connections with both industry peers and the broader public. The filmmaking process entails a complex network of relationships, partnerships, and collaborations that span from scriptwriters to post-production teams, and from casting directors to marketing professionals.

1.2 The Role of Social Network Analysis

Social Network Analysis (SNA) is a research methodology that has garnered significant attention in recent years due to its aptitude for unveiling the intricate connections between individuals, entities, and communities in diverse domains. SNA entails the systematic visualization and analysis of networks to elucidate concealed patterns, pinpoint pivotal influencers, and comprehend the flow of information and resources. When applied to the movie industry, SNA can offer valuable insights into the intricate social dynamics that underpin the creation, distribution, and reception of cinematic productions. It can help with finding collaborations that happen frequently, not only between actors but also between actors and producers, who are often overlooked from the public.

2 Problem and Motivation

The core objective of this research is to meticulously map out and scrutinize the network of professional relationships that defines the contemporary movie industry. By analyzing a comprehensive dataset of recent cinematic productions, we will highlight the influential figures —

actors, directors, writers, and producers — who are central to this creative domain. Our analysis will go beyond mere recognition of star-studded names, aiming to uncover the less visible, yet crucial, contributors who anchor the industry's collaborative network.

In pursuit of these insights, we will employ social network analysis methods to reveal the prominence and frequency of interactions among individuals within the industry. We are particularly interested in identifying those who serve as vital links or hubs in the network. These are the individuals through whom the most paths of collaboration pass, indicating a level of influence that may not always correlate with public fame.

Furthermore, our research will extend to the examination of the industry's global interconnectedness. By incorporating data from international film productions, we will assess the network's geographic diversity and inclusivity. Our analysis will seek to discern whether the network operates as a singular global entity or is fractioned into smaller, perhaps more regionally focused, clusters. This aspect of the study will shed light on the degree to which national film industries are integrated into the global scene or remain self-contained.

Ultimately, this project aims to provide a detailed portrait of the social architecture of moviemaking and to contribute to a deeper understanding of how creative collaborations shape the films that capture the world's imagination. Through this exploration, we aim to articulate the subtle intricacies of the movie industry's social network, offering a valuable perspective for industry stakeholders and film scholars alike.

3 Datasets

3.1 IMDB Movie Dataset

The IMDB movie dataset [1] is a rich and extensive database that encompasses information about movies, television shows, and the people involved in the entertainment industry. It is maintained and updated by IMDB, a popular online resource for movie and television enthusiasts. The dataset covers a wide range of movie-related details, making it a valuable resource for film researchers, data analysts, and movie aficionados.

3.2 Key Features of the IMDB Movie Dataset

The IMDB movie dataset is a comprehensive and versatile database that encompasses a wide range of information related to movies, television shows, and the individuals involved in the entertainment industry. It serves as a valuable resource for film researchers, data analysts, and movie enthusiasts. The dataset includes the following key features:

1. **Movie Information:** Details about movies such as titles, release dates, genres, languages, and runtime are available.

2. **Cast and Crew:** Information about actors, directors, writers, producers, and other personnel engaged in movie production is provided.

3. **Ratings and Reviews:** User and critic ratings, reviews, and votes for movies allow for the analysis of movie popularity and sentiment.

4. **Box Office Data:** Some versions of the dataset include box office earnings and financial information for movies.

5. **Awards and Nominations:** Information regarding awards won and nominations received by movies and individuals in the film industry is included.

6. **User-generated Content:** User-generated content, such as user reviews, user ratings, and user-generated lists of movies, is part of the dataset.

7. **Connections and Collaborations:** The dataset tracks collaborations between actors, directors, and other industry professionals, facilitating the study of network relationships within the industry.

8. **Plot Summaries and Keywords:** It contains plot summaries and keywords associated with movies, aiding in content analysis and categorization.

9. **Release Details:** Information about release dates, countries, and distribution of movies is provided.

These features collectively make the IMDB movie dataset a versatile resource that supports a wide range of applications, including movie recommendation systems, sentiment analysis, social network analysis, academic research, market research, and content categorization.

3.3 Data reduction

The sheer size of the IMDB movie database presents a significant challenge when attempting to develop applications and algorithms that utilize the entire dataset, particularly without access to exceptionally resourceful computing infrastructure. Consequently, we made the deliberate decision to focus our analysis on a subset of the database. Specifically, we included movies and TV series that have garnered a substantial presence, each with a minimum of 50000 user reviews, within the last 20 years (from 2003 and on).

The 50000 threshold may appear very high, but it was necessary: some empirical tests showed that the computational resources only allowed to work with networks having less than 15000 nodes. Thus, as Table 1 shows, the 50000 threshold was selected.

threshold	**nodes**	**edges**	**movies**	**% of kept movies**
1000	138721	1320090	30784	13.5%
2000	95954	905533	20912	9.2%
10000	37303	356797	8113	3.6%
50000	13948	134263	3031	1.3%
10000	8324	77928	1754	0.8%

Table 1: Initial estimate for the network dimension according to the number of reviews

This strategic filtering approach serves several purposes. First, it enables us to concentrate our analysis on content that has demonstrated a degree of prominence and relevance within the IMDb community, eliminating less-noteworthy entries. Additionally, it allows us to target content with broader appeal rather than niche selections.

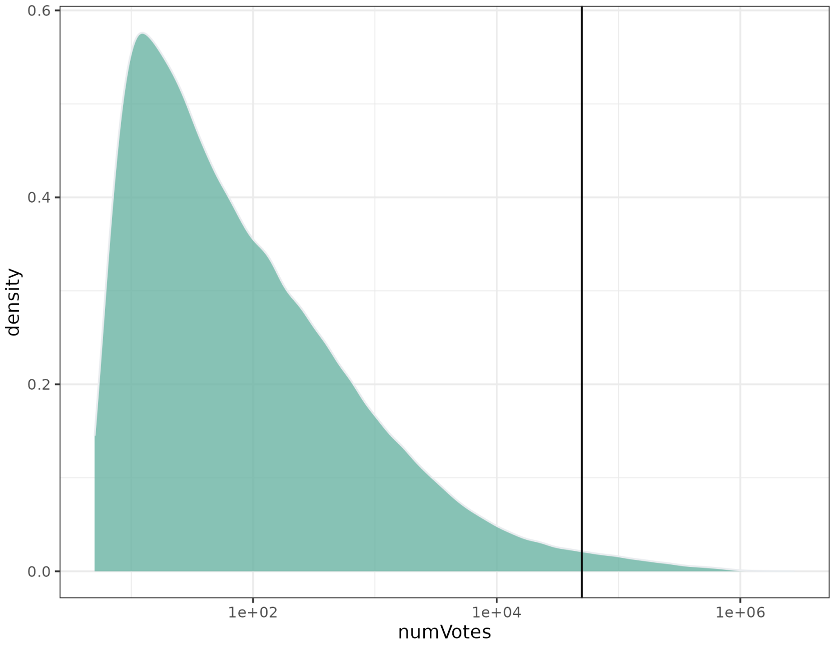


Figure 1: Distribution of votes and 50000 threshold

By applying these criteria, we effectively reduced our dataset size from its original scale of 10,255,958 movies/episodes to a more manageable and focused collection of 3031 instances. We believe that this volume of movie products is more than suited for the goal of our project, since it allows us to inspect the most relevant section of the movie industry. In fact, as figure 1 shows, a relevant amount of movie products has less than 100 reviews, making it less representative and relevant in the field.

This refinement facilitates a more efficient and insightful analysis, enabling us to derive meaningful patterns, trends, and insights from a representative subset of IMDb's extensive repository.

3.4 Network Structure

3.4.1 Nodes

The nodes of the network represent people in the movie industry. Each node has a feature *role* representing the role of that person, such as actor, actress, director or producer.

3.4.2 Edges

The edge between two nodes represents a collaboration within two people in a project. The weight of the edge represents the number of collaborations. Due to the mutual nature of collaborations, edges are not oriented.
The edge with the highest weight is the one between Eric Fellner and Tim Bevan: both of them are producers, and they worked together in 66 projects among the ones selected in our dataset.

4 Validity and Reliability

4.1 Validity

The validity of our dataset model plays a pivotal role in ensuring that our analysis closely represents the realities of the movie industry. Given the comprehensive nature of the data provided by IMDB, our dataset offers a robust foundation for conducting meaningful research and analysis. IMDB provides extensive information about movies, actors, ratings, and industry-related details, allowing us to capture a wide range of aspects related to the film industry. This level of detail enhances the validity of our dataset, enabling us to explore and draw insights that closely align with the multifaceted nature of the movie industry.

However, the process of data reduction, necessitated by our limited computational resources, does have implications for the validity of our dataset. By filtering out movies with fewer user reviews, we potentially exclude a segment of the industry, particularly those actors and movies that may not be mainstream but still contribute significantly to the industry's diversity and richness.

This filtration led to the selection of only 13,948 professionals from a total of 859,384 listed in our dataset for movies released after 2003, representing a mere 1.6% of the total database. Despite this, we believe that this subsample effectively represents the most influential and relevant sectors of the global movie industry. It prioritizes prominent productions and mainstream projects, thereby capturing the core dynamics of the industry while unavoidably sidelining niche and international productions.

We acknowledge that this reduction approach does affect the overall validity of our dataset to a certain extent. However, we assert that this impact is relatively marginal concerning the overarching goals and scope of our project, which is focused on analyzing the dominant trends and patterns within the mainstream movie industry.

4.2 Reliability

Our dataset's reliability is significantly high, largely attributable to the robustness of IMDb as a source. IMDb's reputation for comprehensive and accurate data collection in the film industry lends a high degree of credibility to our dataset.

Aside from the data filtering process described in Section 3.3, our methodology involves no non-deterministic steps. This deterministic approach enhances the reproducibility of our dataset. The consistency in data collection and processing ensures that our dataset can be replicated for similar studies or analyses, offering a reliable resource for researchers and analysts interested in exploring trends and patterns within the movie industry.

5 Measures and Results

There are many different applicable measures that we could apply to our dataset. We will adopt the following ones:

- **Size:** The total number of nodes or entities in the network. Semantically, in our dataset, this represents the total count of professionals in the movie industry, such as actors, directors, and producers, indicating the industry's breadth.

- **Connectivity:** Connectivity pertains to how these nodes or entities are interlinked within the network. It involves understanding whether there are any isolated nodes that have no connections and whether the network comprises multiple connected components or forms a single cohesive unit. In the movie industry, this shows how professionals are linked, highlighting whether the industry is a cohesive unit or if there are isolated individuals.

- **Diameter:** The diameter of a network represents the longest distance, in terms of edges or steps, between any pair of nodes within that network. It offers valuable insights into the farthest possible connections that can exist within the network. In our context, it reflects the maximum distance between professionals in terms of collaborations, indicating how spread out the industry's connections are.

- **Average Path Length:** This metric calculates the average distance between pairs of nodes in the network. It provides a more typical measure of how far apart nodes are on average, illustrating how information or influence might propagate across the network. For the movie industry, this measures the typical separation between professionals, revealing how quickly information or influence can traverse the network.

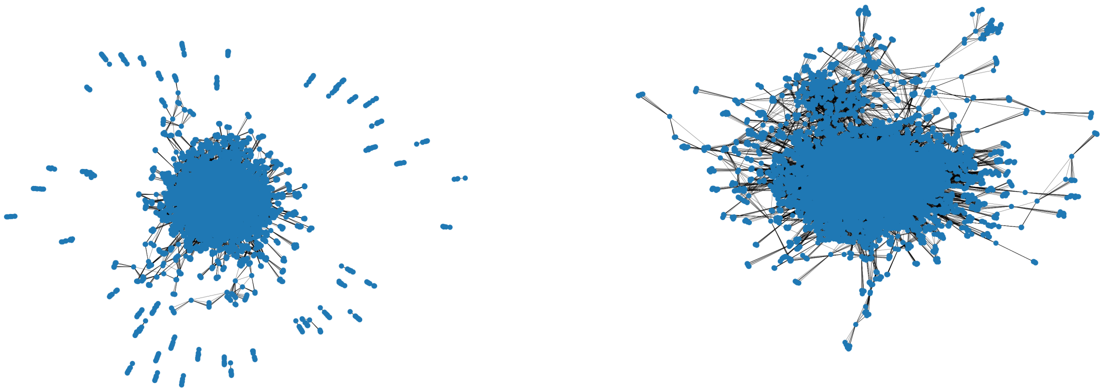
- **Sparsity:** Sparsity quantifies the density of connections or edges in the network. It compares the actual number of edges present to the total possible edges within the network. A lower sparsity value indicates that the network has fewer connections relative to its maximum potential, while a higher sparsity value suggests a denser network with more connections. In our network, it indicates how densely the industry's professionals are connected, with higher sparsity pointing to a more interconnected industry.

- **Degree Centrality:** Degree centrality measures the number of direct connections (edges) that each node in the network possesses. It helps identify nodes that have a large number of direct interactions, indicating their prominence and involvement in collaborations or relationships within the network. Semantically, it signifies the level of a professional's active involvement and prominence in the industry, based on their number of collaborations.

- **Betweenness Centrality:** This metric identifies nodes that serve as critical intermediaries in the network. It quantifies the extent to which a node lies on the shortest paths connecting pairs of other nodes. Nodes with high betweenness centrality act as essential bridges, facilitating communication and information flow between different parts of the network. In our dataset, it identifies those who are key connectors or 'bridges' in the industry, critical for the flow of information and resources.

- **Eigenvector Centrality:** Eigenvector centrality measures a node's influence based on its connections to other influential nodes within the network. Nodes with high eigenvector centrality are not only well-connected but are also connected to other nodes that hold significant relevance in the network. This indicates not just who is well-connected in the movie industry, but who is connected to other highly influential professionals.

- **Cliques:** A clique is a subgroup of nodes where each node is directly connected to every other node within that subset. Cliques represent tightly-knit and highly interconnected



(a) The original network

(b) The largest connected subgraph of the network

groups or clusters of nodes, often indicating strong collaboration or affiliation among their members. In the movie industry context, this represents close-knit collaboration groups, such as recurring casts or production teams.

- **Clustering Coefficient:** The clustering coefficient measures the likelihood that two randomly selected neighbors of a node are also connected to each other. It quantifies the tendency for nodes to form local clusters or tightly-connected neighborhoods. This reflects the likelihood of professionals in the movie industry to form tight-knit collaboration circles.
- **Triadic Closure:** Triadic closure reflects the propensity of nodes to form triangular relationships or closed loops in the network. It suggests the potential for additional connections within existing groups, contributing to the network's overall structure and cohesion. In our network, this suggests the potential for creating new connections within existing collaborative circles in the industry.
- **Small-Worldness:** Small-worldness is a measure that assesses whether a network exhibits the small-world effect. This effect is characterized by high local clustering, akin to regular networks, and short average path lengths, similar to random networks. It gauges the network's efficiency in transmitting information or influence. In the movie industry, high small-worldness implies an efficient network where professionals maintain close collaborations and can easily connect with distant network members.
- **Power Law:** A power law distribution describes a mathematical relationship in which a small number of nodes, known as hubs, possess significantly higher connectivity compared to the majority of nodes. This distribution pattern is often used to characterize degree distributions in complex networks, revealing the presence of highly influential nodes. This would indicate in our network the presence of a few highly influential professionals who dominate the industry's collaborative landscape.

6 Network Characteristics

The network obtained from the IMDB dataset was manipulated using NetworkX [2], and exhibits the following key characteristics:

6.1 Size and Connectivity

The network comprises 13,924 nodes and 127,154 edges, making it a substantial and interconnected representation of the movie industry. Despite its size, the network is not fully connected; instead, it consists of 46 distinct fully connected components. The largest component, which contains 13,455 nodes, accounts for 96.6% of the entire model. The remaining components vary in size, ranging from a maximum of 24 nodes to a minimum of 4 nodes, often corresponding to foreign films. For the sake of a more manageable analysis, we focus solely on the largest connected component, as shown in image 2b.

6.2 Diameter and Average Path Length

The network's diameter is 15. This implies that the longest shortest path between any two nodes within the network requires 15 steps. Additionally, the average path length in the network is 3.9, indicating that, on average, nodes are relatively close to each other in terms of their connectivity.

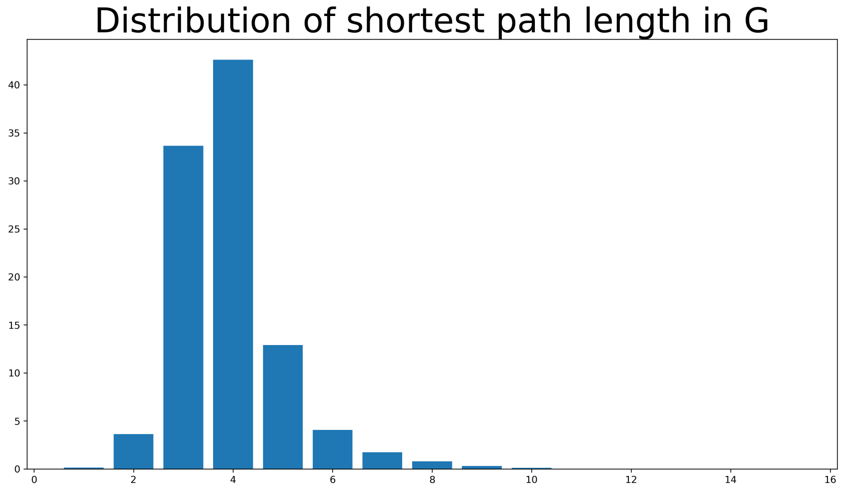


Figure 3: The original network

6.3 Sparsity

Analysis of the shortest paths in the network reveals that the vast majority of them fall within the range of 3 to 5 edges. Only a few nodes exhibit a shortest path length exceeding 9 edges, as can be seen in image 3. This observation underscores the network's sparsity, indicating that there are limited direct connections between nodes. This sparsity is further confirmed by the network's low density, which measures at only 0.0014. This low value suggests that there could possibly be many more connections between the nodes of the movie industry which have not been explored yet.

6.4 Degree centrality

Degree centrality provides valuable insights into the network structure: it represents the number of edges that each node has, in our case, the number of different professionals that a subject worked with. Examining the top actors and producers by degree centrality reveals noteworthy

Table 2: Top 10 Actors and Their Centrality

Actor	Category	Centrality	Neighbours
Mark Wahlberg	Producer	0.0199	269
Tim Bevan	Producer	0.0193	259
Samuel L. Jackson	Actor	0.0188	253
Ryan Reynolds	Actor	0.0183	246
Jason Blum	Producer	0.0183	246
Dwayne Johnson	Actor	0.0180	242
Scott Rudin	Producer	0.0178	240
Colin Farrell	Actor	0.0178	239
Matt Damon	Producer	0.0175	236
Scarlett Johansson	Actress	0.0175	235

patterns:

Visual examination of the network image reinforces the degree centrality analysis, as can be seen in image 4. The majority of individuals in the dataset exhibit degree centrality values well below 0.005, underscoring the network's inherent sparsity. This observation becomes even more intriguing when considering that our analysis is based on a limited subsample of the entire dataset, highlighting the potential for even greater network complexity beyond our current scope.

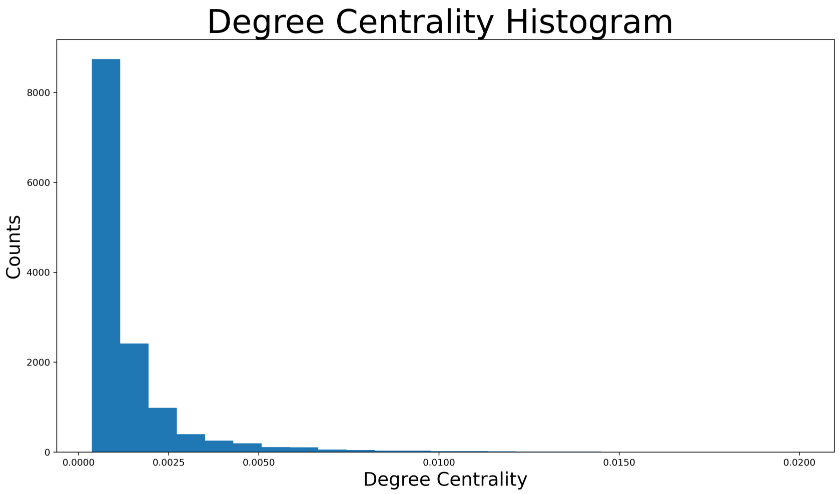


Figure 4: Degree centrality

6.5 Betweenness Centrality

Betweenness centrality provides valuable insights into the network by identifying individuals who play critical bridging roles in connecting various parts of the industry.

It's important to note that the vast majority of betweenness centralities in the network are below 0.005, as can bee seen in image 5. This observation aligns with the network's inherent sparsity, where most nodes do not act as bridges in shortest paths. However, this sparsity results in some nodes, such as Mamoru Miyano and Samuel L. Jackson, having higher betweenness

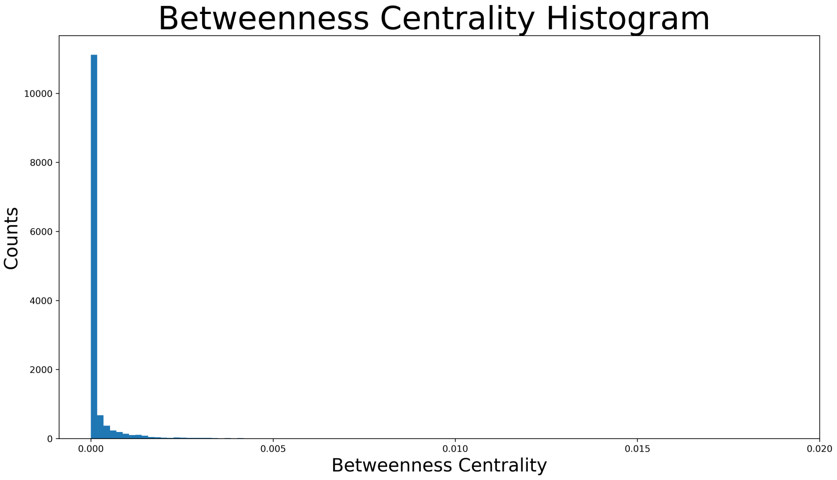


Figure 5: Betweenness centrality

Table 3: Top 10 Actors and Their Betweenness Centrality

Actor	**Betweenness Centrality**
Mamoru Miyano	0.0175
Dwayne Johnson	0.0156
Samuel L. Jackson	0.0154
A.R. Rahman	0.0145
Irrfan Khan	0.0141
Jason Blum	0.0131
Mark Wahlberg	0.0120
Priyanka Chopra Jonas	0.0119
Jason Statham	0.0116
Anupam Kher	0.0115

centralities, highlighting their unique roles as key intermediaries in connecting various parts of the industry network.

6.6 Eigenvector Centrality Insights

Eigenvector centrality assesses an individual's influence in the network, considering not only their connections but also the quality of those connections. In the movie industry network, high eigenvector centrality signifies nodes that hold substantial influence and are connected to other influential nodes.

The results reveal several individuals with noteworthy eigenvector centrality values. These individuals wield significant influence over the movie industry network, often through well-placed connections to other influential figures. Their roles in shaping industry dynamics and decisions are indicative of the network's structure and the key players within it.

6.7 Cliques

The search for cliques is trivial in this case: a clique is a is a set of nodes such that every member of the set is connected by an edge to every other. Given the design of the network,

Table 4: Top 10 Actors and Their Eigenvector Centrality

Actor	Eigenvector Centrality
Mark Wahlberg	0.0935
Matt Damon	0.0890
Scott Rudin	0.0881
Tim Bevan	0.0854
Cate Blanchett	0.0845
Colin Farrell	0.0809
Brad Pitt	0.0780
Christian Bale	0.0769
Ryan Reynolds	0.0756
Eric Fellner	0.0752

each movie creates a clique, since all of the actors/directors/producers of a movie are connected to each other.

6.8 Clustering Coefficient and Triadic Closure

- **Clustering Coefficient (0.778):** The clustering coefficient of the network measures the likelihood that two randomly selected friends of a node are also friends with each other. A value close to 1 suggests a highly cohesive and interconnected graph, indicating strong triadic closure.

- **Nodes with Coefficient of One:** More than 8,000 out of 13,455 nodes in the network have a clustering coefficient of one. This signifies that many nodes form tightly-knit groups where virtually all friends of a node are interconnected. These high clustering coefficients indicate strong local relationships within subgroups of the network.

- **Unique Triangles (383,525):** The presence of a large number of unique triangles in the network signifies that sets of three nodes are frequently interconnected. This reflects the prevalence of triadic closure, where nodes often connect in groups of three.

- **Average Node Participation in Triangles (85.5, Median 36):** On average, a node is part of approximately 85.5 triangles, while the median node participation is 36 triangles. This indicates that while some nodes are highly central to numerous triangles, the majority of nodes participate in a substantial number of triadic relationships.

6.9 Scale-Free

Scale-free networks are a prominent class of complex networks commonly observed in various real-world systems, including social networks, the World Wide Web, and citation networks. They exhibit a specific structural property characterized by a power-law distribution, which means that the probability of a node having a certain degree decreases exponentially as the degree increase. In a power-law distribution, the probability of observing a value x is proportional to $x^{-\alpha}$, where α is the power-law exponent.

The power-law exponent (α) is a critical parameter that defines the shape of the distribution. It quantifies the rate at which the probability of observing larger values decreases with x. The formula for alpha is:

$$\alpha = 1 + n\left(\sum_i \ln \frac{d_i}{d_{min} - \frac{1}{2}}\right)^{-1}$$

The power-law exponent α provides insights into the distribution's behavior:

- When $\alpha < 2$, the distribution has a very heavy tail, indicating extreme values.
- When $2 \leq \alpha \leq 3$, the distribution is heavy-tailed with scale-free characteristics.
- When $\alpha > 3$, the distribution has a less heavy tail and is more well-behaved.

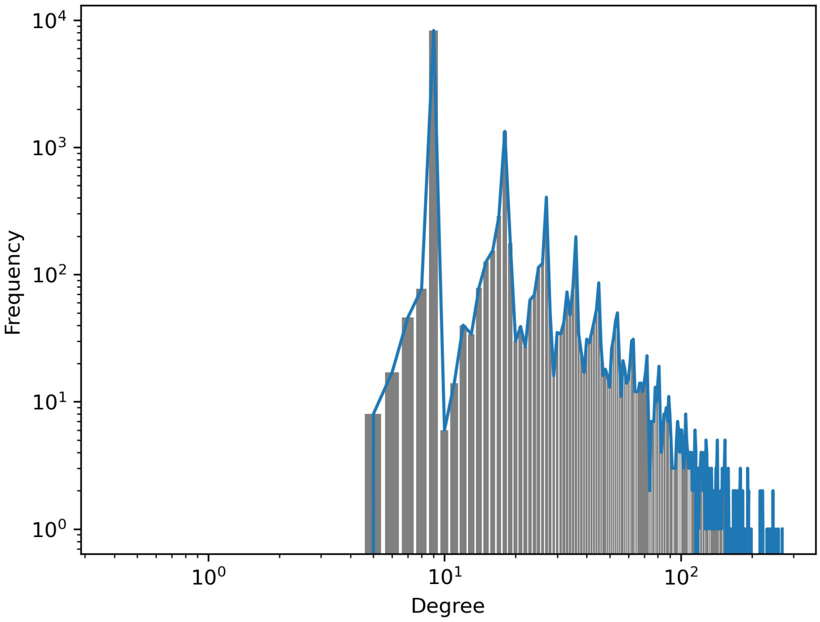


Figure 6: Power-law distribution

In our analysis of the movie network, we aimed to assess whether certain network properties follow a power-law distribution and determine the power-law exponent (α). Our findings reveal that the distribution of these network properties aligns with a power-law distribution, with an estimated α value of approximately 3.273.

The obtained α value does not fall within the range of 2 to 3, which is means that our network does not exhibit scale-free behavior. This suggests the following interpretations in the context of the movie network:

- **Moderate Inequality in Connectivity:** The higher α value implies a network with more evenly distributed connections, reducing the dominance of highly connected hubs.
- **Resilience of the Network:** Networks with an α value greater than 3 are generally more resilient to random node failures, suggesting stability in the movie industry network.
- **Diverse Collaborative Opportunities:** This structure indicates potential for more democratic collaboration opportunities across a wider range of professionals.

- **Less Extreme "Winner-takes-all" Dynamics:** The network exhibits less pronounced "winner-takes-all" phenomena, favoring a more inclusive environment.
- **Strategies for Networking:** The findings suggest that building a network in the industry might not require connections only with top-tier individuals, as influence is more evenly spread.
- **Potential for New Influential Nodes:** The current structure allows for the emergence of new influential nodes over time.

In conclusion, the analysis of the movie industry network with a computed power-law exponent α of approximately 3.273 indicates a system where opportunities for collaboration and influence are not limited to a few highly connected individuals. Instead, there is potential for a wider range of professionals to contribute significantly to the industry. This structure could lead to a more resilient and adaptable industry, capable of withstanding changes and incorporating new talents and ideas. It highlights the evolving nature of the movie industry and suggests a landscape where diversity and inclusivity might play a crucial role in its future development.

6.10 Small-Worldness Analysis

A small-world network is a network which typically exhibits a high degree of local clustering and short average path lengths, making it efficient for information or influence to traverse the network.

To assess the presence of small-worldness, we calculated two key metrics: sigma (σ) and omega (ω). These metrics provide insights into the network's clustering coefficient and average path length compared to those of random networks.

- **Sigma (σ):** The sigma value quantifies how the clustering coefficient of our network compares to that of a random network with similar characteristics. A higher sigma value indicates that our network exhibits a higher clustering coefficient than expected in a random network. In our analysis, we obtained a sigma value of 522, which strongly suggests a significantly higher clustering coefficient in our network.

  $$\sigma = \frac{\frac{C}{C_r}}{\frac{L}{L_r}}$$

  Nonetheless, the sigma value alone is prone to bias and error, since it is susceptible to the network size and density, which makes it necessary to evaluate omega before drawing any conclusion.

- **Omega (ω):** Omega measures how the average path length of our network compares to that of a random network. A negative omega value indicates that our network has a shorter average path length than expected in a random network. In our analysis, we found an omega value of -0.34, implying that our network's average path lengths are shorter than those of random networks.

  $$\omega = \frac{L_r}{L} - \frac{C}{C_r}$$

The high sigma value (522) indicates a significantly higher clustering coefficient than expected in random networks. This points to the presence of local clustering or tightly connected groups of nodes within our network On the other hand, the negative omega value (-0.34) suggests that our network's average path lengths are shorter than those in random networks. This indicates that despite the local clustering, our network maintains efficient global connectivity, allowing for relatively short paths between any two nodes.

In conclusion, our small-worldness analysis reveals compelling evidence that our network exhibits the small-world property. The combination of a high clustering coefficient (sigma) and shorter average path lengths (negative omega) indicates an efficient and structurally advantageous network.

The presence of small-worldness in our network can have implications for information propagation, robustness, and the overall dynamics of our network. It signifies that while nodes tend to cluster locally, the network maintains global connectivity and efficient pathways for interactions or influence to spread.

7 Conclusion

7.1 Connectivity

As previously mentioned, the original network is not fully connected. By inspecting the smaller non connected components, we can state that they are either international movies, such as the case of the Italian movie "Perfetti Sconosciuti", or strongly political and controversial products which Hollywood is not willing to link itself with, such as "Citizienfour", the documentary on Edward Snowden.

7.2 Centrality

Mark Wahlberg, a producer, holds the highest degree centrality with 0.0199, indicating his extensive collaborations within the industry, as he has worked with nearly 2% of the entire network. Similar high values are observed for other prominent figures, including producers like Tim Bevan and notable actors such as Samuel L. Jackson, Ryan Reynolds, and Dwayne Johnson (also known as "The Rock"). Notably, Scarlett Johansson is the sole female actress to appear in the top 10 list, reflecting her substantial industry connections.

Despite not ranking as prominently in degree centrality, Mamoru Miyano stands out with a high betweenness centrality. This suggests that while he may have fewer direct collaborations, he plays a pivotal role in connecting disparate components of the industry network. His ability to act as a bridge between different individuals is a noteworthy feature.

Dwayne Johnson and Samuel L. Jackson, known for their extensive collaborations, also exhibit significant betweenness centrality values. This reaffirms their roles as key bridges, connecting multiple nodes within the network.

Jason Blum, recognized for his numerous collaborations, holds substantial betweenness centrality, indicating his importance in facilitating connections among industry professionals.

7.3 Clustering

The high clustering coefficient and the abundance of nodes with a coefficient of one highlight a network structure characterized by tightly-connected groups. This observation aligns with the unique nature of the movie industry network, where numerous subgroups exist, and individuals collaborate extensively. Unlike data gathered from typical social networks, where relationships are often one-to-one, the movie industry's context naturally leads to the formation of significant clusters and groups. This phenomenon is primarily attributed to the collaborative nature of movie production, where the cast of a movie forms a small network within the larger industry framework.

On the other hand, is clear how in the movie industry, clusters are strongly linked to the movie genre: this kind of analysis is a possible future development of this project.

7.4 Scale-free and Small-worldness

The analysis of the network suggests that the movie industry does not conform to a scale-free structure. This finding challenges the notion of a rigidly hierarchical organization within the industry and suggests a more egalitarian distribution of connections and influence.

Instead of being dominated by a few highly influential figures, the network appears to be characterized by a broader distribution of influence. This means that while there are still key figures, such as well-known actors, respected directors, and successful producers, their role in the network is not as disproportionately central as it would be in a scale-free network. Their connections, while significant, are part of a more extensive web of interactions that includes a wide range of professionals.

In this context, the industry's network is more democratized, with opportunities for influence and collaboration more evenly spread across different participants. This structure may facilitate a more diverse and dynamic creative process, as it allows for a wider range of voices and talents to participate and be recognized.

The absence of a scale-free structure implies that the industry is potentially more resilient to disruptions affecting individual nodes. In a network where influence and connectivity are not overly concentrated in a few nodes, the loss or diminished activity of any single individual is less likely to cause major disruptions.

For stakeholders within the industry, this means that building a broad network of connections could be as valuable as connecting with a few key figures. For emerging talents, this structure might offer a more accessible entry point into the industry, as the field is not as dominated by a small elite. It highlights the importance of building diverse connections and collaborations to establish oneself in the industry.

In summary, the absence of scale-free characteristics in the movie industry network points to a more evenly distributed field of influence and collaboration. This could foster a more inclusive and dynamic environment, encouraging a wide range of participants to contribute to and shape the cinematic landscape.

8 Critique

The analyses we performed are a solid and comprehensive analysis, which allowed us to gather and better understand the underlying dynamics of the movie industry. The centrality of the actors is coherent with their popularity and relevance within the movie business.

On the other hand, it is fair to assess that our work has been strongly influenced by the computational limits that such a huge task posed, given the immensity of the addressed dataset. If we had more computation power, using a broader sample would be a natural extension of our work, and would allow to delve also into niche sectors of the movie industry, which we had to overlook in this case.

The project is also clearly prone to further extensions in the way the network is build, which allows and encourages further analyses on it. A clear development could be building a network including the genre of the movie product on which the collaboration is based. This would provide a great basis for a deeper and more accurate cluster analysis, allowing us to inspect not only topological features of the network but also more business-oriented features.

References

[1] International movie database. https://imdb.com.

[2] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using networkx. In Gaël Varoquaux, Travis Vaught, and Jarrod Millman, editors, *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA, 2008.