Social Network Analysis of Netflix Movies and TV Shows Actors

Mohsen Hosseini 30234961 (672), Samiha Hoque 30130973 (572), Bardia Alidoust 30232124 (672)

April 4, 2024

Abstract

Our study delves into the network of actors within Netflix's extensive library, which boasts over 8,000 titles. By constructing a network where nodes represent actors and edges signify shared projects, our analysis aims to uncover the latent structures and dynamics that govern actor collaborations.

Our dataset has been downloaded from Kaggle that has attributes like cast, country, duration and more. We picked cast as our nodes featuring 36,439 actors and 289,205 connections to explore community structures, centrality measures, and collaboration patterns. The collaborations was found mainly by the country attribute and our preliminary results reveal a rich tapestry of actor communities and geographical influences.

Furthermore, our centrality analysis spotlighted actors who not only have a broad range of collaborations but also serve as crucial links between otherwise disparate communities, highlighting their role in the cohesive fabric of the entertainment industry. The parameters such as degree centrality and betweenness centrality were pivotal in identifying these key actors. Degree centrality highlighted actors with the highest number of connections, reflecting their popularity and influence. Betweenness centrality, on the other hand, underscored actors who frequently appear on the shortest paths between others, indicating their role as connectors or bridges within the network.

These insights offer a nuanced understanding of the collaborative landscape on Netflix, providing valuable perspectives for producers, marketers, and strategists aiming to navigate the complex interplay of preferences, trends, and network dynamics. By leveraging our findings, stakeholders in the entertainment industry can make informed decisions about content development, marketing strategies, and casting choices, ultimately enhancing the alignment with viewer preferences and fostering greater success in the competitive marketplace.

1 Research Questions

- **Community Detection:** Can the network be divided into distinct communities or groups based on collaboration patterns? What characteristics define these communities, and are they related to specific genres, countries, or directors, release year?
- **Centrality Analysis:** Which actors have the highest degree centrality in the network, indicating that they have worked with the most diverse set of colleagues? Are there any actors who act as bridges or connectors between different clusters of actors in the network?
- **Collaboration Patterns:** What patterns emerge in terms of actors collaborating with the same set of individuals repeatedly? Are there certain pairs or groups of actors who frequently collaborate with each other?

2 Introduction

In the rapidly evolving landscape of digital entertainment, the interplay between actors' collaborations and content popularity has garnered significant scholarly attention. With streaming platforms like Netflix revolutionizing how content is consumed globally, understanding the network of actor collaborations offers invaluable insights into content strategy and audience engagement. This project situates itself at the intersection of network science and entertainment analytics, aiming to unravel the complex web of relationships within Netflix's vast catalog of movies and TV shows.

Recent studies have increasingly focused on applying network analysis to various domains, including social networks, biological systems, and, pertinent to our study, the entertainment industry. However, the unique nature of streaming platforms, characterized by their extensive and diverse content libraries, presents new challenges and opportunities for network analysis. Our project draws inspiration from pioneering works such as Newman's (2001) [1] study on the structure of scientific collaboration networks and Borgatti and Halgin's (2011) [2] exploration of social media networks. These studies lay the foundation for our approach, emphasizing the importance of network structure and actor centrality in understanding complex systems. Furthermore, the concept of community detection in networks, as explored by Fortunato (2010) [3], provides a crucial analytical framework for our study. By identifying distinct communities within the Netflix actor network, we aim to uncover patterns of collaboration that transcend individual projects, hinting at broader industry trends and preferences.

Our methodology builds upon the robust analytical toolkit provided by network science, employing centrality measures to identify key actors and community detection algorithms to explore the network's subdivision into meaningful clusters. Specifically, we leverage the Netflix dataset, comprising over 8,000 titles and 36,439 actors, to construct a network where nodes represent actors, and edges signify their collaborations on Netflix titles. Through this lens, we aim to answer pressing questions about community structures, centrality, and collaboration patterns within the entertainment giant's offerings.

This project not only contributes to the academic discourse on network analysis in the entertainment industry but also offers practical insights for content creators, marketers, and strategists looking to navigate the complex dynamics of actor collaborations. By illuminating the hidden structures and key influencers within the Netflix network, we provide a roadmap for leveraging actor networks in content strategy and audience engagement efforts.

In sum, our study stands at the confluence of network science and entertainment analytics, offering a novel perspective on actor collaborations within Netflix. Through meticulous analysis and the application of advanced network theory, we aspire to shed light on the patterns and implications of these collaborations, enriching both academic understanding and industry practice.

3 Dataset Description

Raw Data Source and Characteristics : The dataset underpinning our analysis is sourced from Kaggle, an online community of data scientists and machine learning practitioners that provides access to datasets and computational resources. The "Netflix Movies and TV Shows" dataset is a comprehensive collection of titles available on Netflix as of mid-2021, reflecting the platform's diverse offerings to its global subscriber base exceeding 200 million. Comprising over 8,000 entries, the dataset is structured as a CSV file, ensuring ease of access and manipulation for our purposes.

Each entry in the dataset represents a unique title (movie or TV show) and includes a variety of attributes such as the cast (actors), directors, ratings (e.g., TV-MA, PG-13), release year, duration (for movies, in minutes; for TV shows, in seasons), and more. This rich array of information provides a multifaceted view of Netflix's content, allowing for an in-depth analysis of actor networks and collaborations.

Data Cleaning and Wrangling : Given the analytical focus on actor collaborations, our initial data cleaning efforts concentrated on ensuring the integrity of the 'cast' and 'country' columns, which are pivotal for constructing our network. We removed entries lacking cast information, as these do not contribute to our network analysis. Additionally, for entries listing multiple countries, we extracted and retained only the first country listed. This simplification was deemed necessary to maintain focus on the primary production location, facilitating clearer analysis of geographic patterns in actor collaborations.

Network Construction : Transforming the dataset into a network format required identifying actors as nodes and collaborations (shared titles) as edges. A Python script was developed to parse the 'cast' field of each title, generating a list of actors and mapping their collaborations. This process resulted in a network comprising 36,439 nodes (actors) and 289,205 edges (collaborations), with edge weights reflecting the number of shared titles between actor pairs. This network serves as the backbone of our analysis, enabling the examination of community structures, centrality measures, and collaboration patterns.

Figure 1: **Screenshot of the raw dataset**

Metadata Definition : In our network, nodes (actors) are annotated with metadata including their name and country. Edge weights represent the number of projects shared by actor pairs, offering a quantitative measure of collaboration strength. This metadata is crucial for nuanced analysis, allowing us to explore not just the quantity of connections but also their quality and context.

4 Basic Statistics

This table [1] and the statistics it contains provide a comprehensive overview of the network's structure and connectivity, offering insights into the complexity and density of the relationships modeled within the network.

Basic Statistics	Whole Network	Largest Connected Component (LCC)
Number of Connected Components	453	-
Number of nodes	36,039	30,937
Number of edges	289,207	260,645
Average degree	16.0497	16.85
Minimum degree	1	1
Maximum degree	273	273
Average Clustering Coefficient	Use LCC for subsequent analysis	0.818
Average Shortest Path	Use LCC for subsequent analysis	5.482

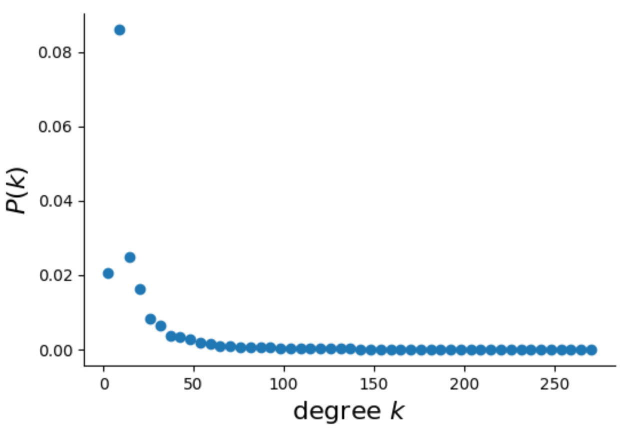
Table 1: **Basic Statistics of the Network and its Largest Connected Component (LCC)**

The basic statistics table provides a detailed look into the network of actors based on the Netflix dataset, highlighting both the overall network's structure and the properties of its Largest Connected Component (LCC). With 453 distinct connected components, the network displays a degree of fragmentation, though a significant majority of actors belong to the LCC, indicating a dense core of collaboration. The network comprises 36,039 actors, with the LCC accounting for 30,937, underscoring the interconnected nature of most actors within this subset.

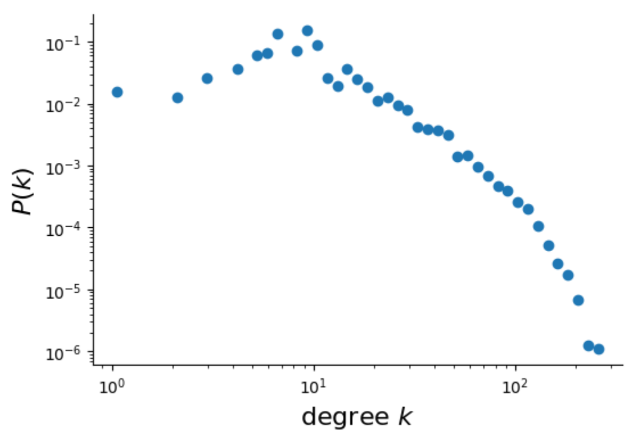
A total of 289,207 collaborations are noted within the entire network, with the LCC holding 260,645 of these connections, showcasing a tight-knit community where actors frequently work together. The average degree of collaboration is about 16 for both the entire network and the LCC, pointing to a high level of engagement among actors. The minimum and maximum degrees of 1 and 273, respectively, reflect a wide range of actor involvement, from those with singular collaborations to highly central figures within the network.

The LCC's average clustering coefficient of 0.818 indicates a strong propensity for actors to form closely-knit clusters, suggesting that collaboration tends to occur within well-defined groups. The average shortest path length of 5.482 within the LCC highlights the small-world nature of the entertainment industry, where any two actors can be connected through a relatively short chain of intermediaries.

In the Figures below, the log-log scale plot provides a clearer representation of the degree distribution of the network, than the linear plot. It is evident that the majority of nodes have a degree around 10, and as the degree increases, the probability of encountering such nodes decreases. This pattern is expected, as there are typically fewer influential actors who appear in numerous titles.



(a) **Linear Scale**



(b) **Log-Log Scale**

Figure 2: **Degree Distribution Plot of LCC**

5 Network Visualization

This main visualization [3] shows a graphical representation of the network with nodes (actors) and edges (collaborations). Nodes are likely color-coded based on the country clusters. The visualization provides a visual map of how tightly actors are grouped by country and who the key influencers are within this network. It gives an intuitive understanding of the global collaboration patterns on Netflix, highlighting the most interconnected regions and individuals in the entertainment industry. The color-coded legend positioned in the upper-left corner corresponds to the respective countries depicted within the network. Upon examination, it is evident that the predominant clusters are those representing the United States, the United Kingdom, and India. Additionally, peripheral clusters, delineated in grey and situated at a distance from the central nexus of the visualization, comprise approximately 3 percent of the total nodes within the dataset.

To create the visualization shown below 3 in Gephi, the first step is to export the model in GML format using NetworkX. Once in Gephi, the ForceAtlas2 layout with default parameters is applied to the graph. This layout helps to separate the nodes and create the overall structure of the graph. The reason ForceAtlas2 was chosen over ForceAtlas is due to the size of the graph; ForceAtlas2 is better suited for larger graphs.

Next, the node sizes are determined based on the degree of each node, ranging from a size of 5 to 50. The node colors are determined based on the "country" field of each node. Gephi automatically assigns colors based on this field.

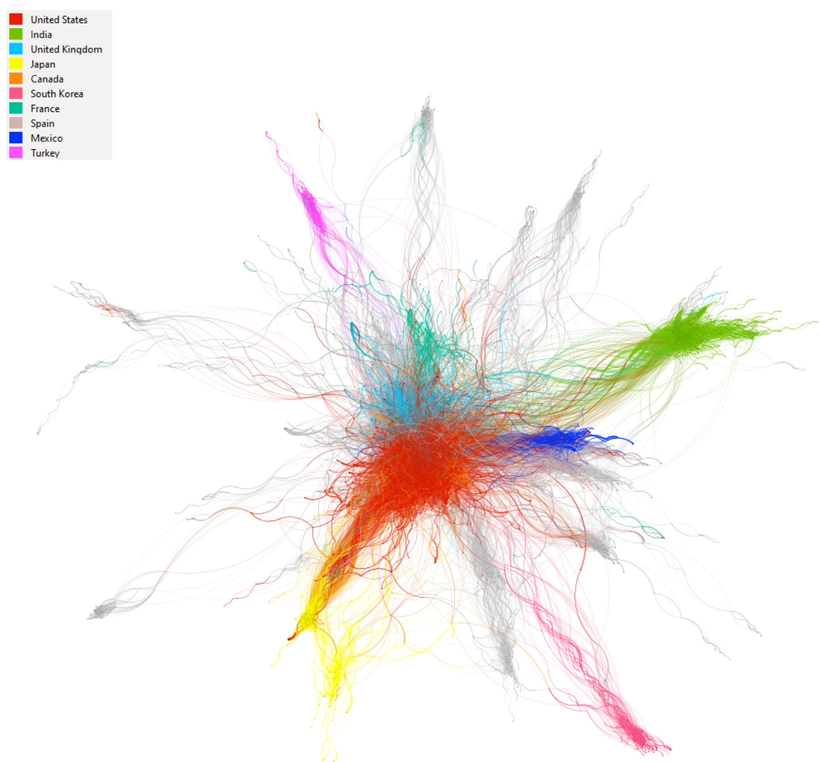


Figure 3: **Visualization of the Whole Network**

The third visualization 6 illustrates the betweenness centrality of nodes across the network. To visualize the top 10 nodes with high betweenness centrality, a new field called "bridge" is added to all nodes. This field is set to "true" for the top 10 nodes with high betweenness centrality, and "false" for the remaining nodes. The visualization properties such as color, node size, label size, and label color are then set based on the "bridge" field of the nodes.

In Gephi, the ForceAtlas2 layout algorithm is once again used to layout the graph based on these updated properties. This layout helps to visually emphasize the nodes with high betweenness centrality, as they are now distinguished by their "bridge" field.

By following these steps, the graph will display the top 10 nodes with high betweenness centrality with specific visual properties assigned to them based on the "bridge" field.

6 Results

The results stem from a comprehensive analysis of actors as nodes, and edges signify the frequency of their collaborations. As depicted in the second table [2], three community detection algorithms were employed to categorize actors into

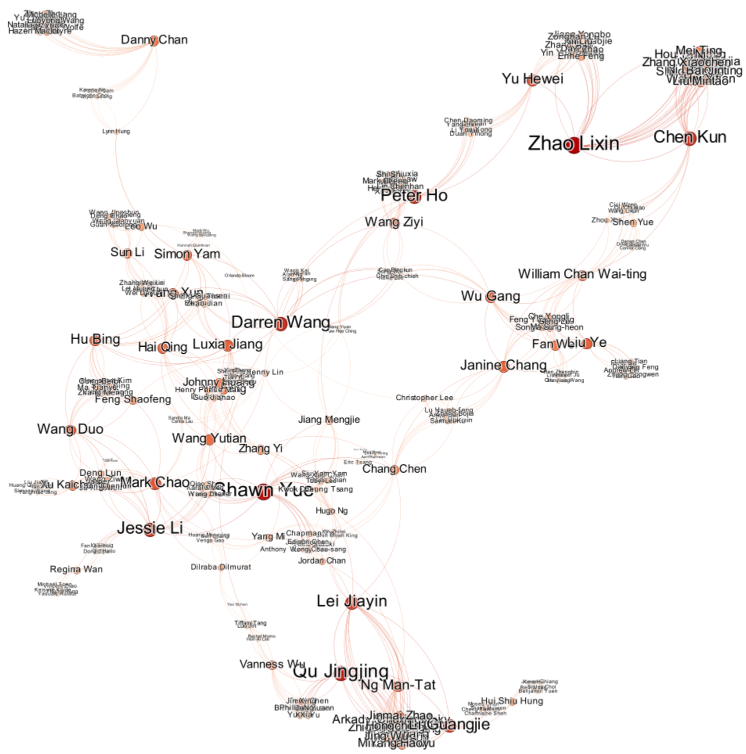


Figure 4: **Collaborative Network among Chinese Cctors**

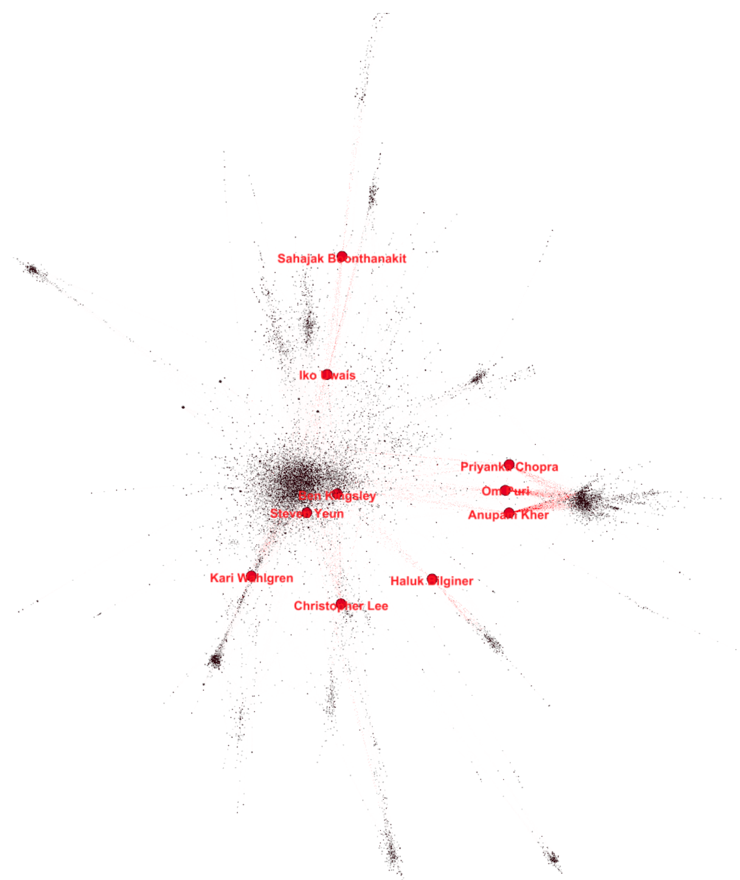


Figure 5: **Betweenness Centrality of Nodes Across the Network**

communities based on nodes, edges and the edge weights. Presented herein are the comprehensive statistics derived from those three community detection algorithms analyzed network:

Community Detection Results :

- **Greedy Modularity Communities:** This algorithm aims to maximize a network's modularity score, measuring the density of links within communities compared to links between communities. It operates on a greedy optimization principle, initially treating each node as its own community and then iteratively merging communities to achieve the highest possible modularity increase. This method is efficient for detecting high-modularity partitions in large networks and is particularly useful for identifying tightly-knit groups that are more densely connected internally than with the rest of the network. This algorithm identified the United States as the most common country within the actor communities. With a similarity ratio of 0.614 and a weighted average of 0.623, it suggests moderate clustering based on the country attribute. This indicates some tendency for American actors to collaborate more frequently with each other than with actors from other countries.
- **Asynchronous Label Propagation Communities (asyn-lpa):** The Asynchronous Label Propagation Algorithm is based on the concept of spreading labels throughout the network and allowing nodes to adopt the most frequent label among their neighbors. Each node updates its label asynchronously, independent of the update order, leading to rapid convergence on a consensus labeling that delineates community boundaries. This algorithm is straightforward and highly scalable, making it suitable for large networks. It excels in uncovering communities based on the natural flow of information within the network but may produce varying results across different runs due to its stochastic nature. This method resulted in a higher community similarity, with a ratio of 0.749 and a notable weighted average of 0.900. This high score implies a strong propensity for American actors to form clusters, indicating that they often work within the same community of actors, possibly reflecting industry or cultural preferences in casting.
- **Louvain Communities:** The Louvain Method is a hierarchical clustering algorithm that also seeks to optimize the network's modularity. It operates in two phases that are repeated iteratively: first, it assigns each node to its own community, and then it aggregates nodes of the same community and builds a new network whose nodes are the communities. In the second phase, it computes the modularity and compares it to the previous level, choosing the partition that yields the highest modularity. This method is known for its efficiency and ability to detect community structures at different scales, from small to very large networks. The Louvain method showed a ratio of 0.771 for the United States, with a weighted average of 0.742. While the ratio is higher compared to Greedy Modularity, indicating more consistent clustering by country, the weighted average is lower than that of the asyn-lpa method, suggesting that while the clusters are strongly defined, they may encompass a more diverse set of countries.

Algorithm	Most Common Country	Weighted Average
Greedy Modularity	United States	0.623
asyn_lpa	United States	0.900
Louvain	United States	0.742

Table 2: **Results of Community Detection Algorithms**

The bar chart [6] resulting from this analysis would provide a visual representation of how different community detection algorithms cluster the actor network based on the country attribute. The visualization supports the interpretation of the data by highlighting the effectiveness of each algorithm in creating country-centric communities within the network. The chart features a metric developed specifically for this project; the rationale behind its creation is detailed in methods section [9].

Degree Centrality Results : The results indicate a few actors with an exceptionally high degree of connections, such as **Anupam Kher and Samuel L. Jackson**, placing them as central figures in the network. Their high connectivity implies a diverse range of collaborations and possibly a significant influence on the flow of information and trends within the industry.

Betweenness Centrality Results : Actors with high betweenness centrality scores serve as bridges in the network. These bridge actors are crucial as they likely bring together disparate clusters of actors, facilitating the spread of new collaborative opportunities and creative influences across the network.

Collaboration Patterns Results : The analysis of edges with the highest weights reveals patterns of frequent collaborations between certain actor pairs. For instance, **John Paul Tremblay** and **John Dunsworth**, with the highest number of shared projects, indicate a strong collaborative bond, which could be reflective of their participation in a particular series or genre of content.

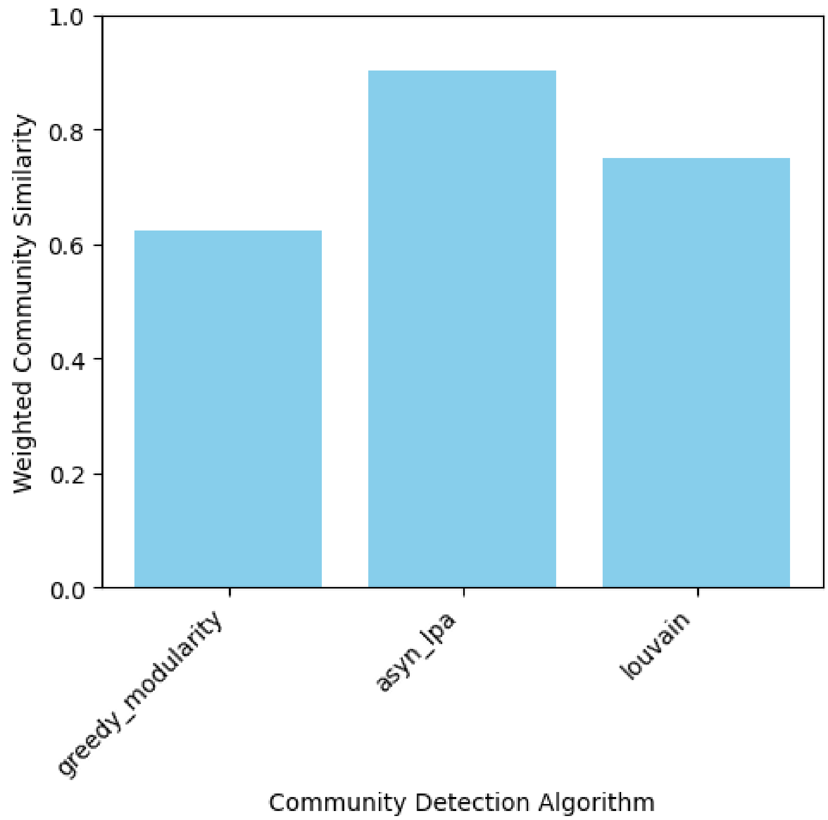


Figure 6: **Community Detection Algorithms with Metric**

Node	**Degree**
Anupam Kher	273
Samuel L. Jackson	230
Yuichi Nakamura	216
Yuki Kaji	209
Shah Rukh Khan	209
Fred Tatasciore	207
Fred Armisen	199
Akshay Kumar	193
Takahiro Sakurai	192
Om Puri	187

Table 3: Top 10 Nodes with Highest Degree

Edge	**Weight**
John Paul Tremblay – John Dunsworth	10
Robb Wells – John Dunsworth	10
Junko Takeuchi – Chie Nakamura	10
Ashleigh Ball – Tabitha St. Germain	9
Junko Takeuchi – Kazuhiko Inoue	8
Chie Nakamura – Kazuhiko Inoue	8
Kappei Yamaguchi – Kumiko Watanabe	8
Kappei Yamaguchi – Satsuki Yukino	8
Kappei Yamaguchi – Koji Tsujitani	8
Kappei Yamaguchi – Houko Kuwashima	8

Table 4: Top 10 Edges with Highest Weight

Actor	**Betweenness Centrality**
Anupam Kher	0.0587
Om Puri	0.0304
Iko Uwais	0.0256
Ben Kingsley	0.0241
Sahajak Boonthanakit	0.0237
Steven Yeun	0.0220
Haluk Bilginer	0.0203
Christopher Lee	0.0196
Priyanka Chopra	0.0181
Kari Wahlgren	0.0173

Table 5: **Betweenness Centrality Measures for Selected Actors**

7 Comparison to Suitable Null Model

Given the constraints imposed by limited hardware resources and the considerable size of our dataset, we were constrained to generating only five ensembles of random networks for our analysis. This limitation necessitated a focused approach to evaluating the structural properties of these networks as a comparative baseline against our observed network.

The random networks were constructed using the Configuration Model, a choice motivated by the need to preserve the degree distribution of the original network. This method ensures that each random network mirrors the degree sequence of our observed network, providing a controlled basis for comparison. The Configuration Model is particularly suitable for this purpose because it allows for the generation of random graphs that maintain specific characteristics of the original graph, thus enabling a more meaningful comparison of network properties.

For each generated random network, self-loops and parallel edges were systematically removed to ensure a more accurate reflection of real-world network structures. Following this, we calculated the average clustering coefficient and average shortest path length for the largest connected component within each random network. These metrics are crucial for understanding the cohesiveness and navigability of the network structure.

The results were summarized by computing the mean and standard deviation for both the clustering coefficients and shortest path lengths across the five random networks. These statistical measures provide insight into the expected variability of these metrics within randomly generated networks that share the original network's degree distribution. The Configuration Model is particularly suitable for our purposes for several reasons:

1. **Degree Preservation:** It retains the original network's degree for each node, making it a stringent test for network properties like clustering and path lengths.
2. **Flexibility:** It can handle any degree sequence, accommodating the diverse connectivity patterns observed in our network.
3. **Runtime Efficiency:** Although generating one ensemble took approximately 11 minutes, this is considered acceptable given the complexity and size of our network.

Comparison with Null Model Results:

1. **The Average Clustering Coefficient :**
   - Random Network: The average clustering coefficient is significantly low (0.0022 ± 5.47e-05), indicating that there are fewer triangles than expected in a random configuration.
   - Implication: Compared to our network, this suggests that the observed network has a higher level of local clustering, indicative of more tightly knit groups or communities within the network, which is a characteristic not captured by the random model.
2. **The Average Shortest Path Length :**
   - Random Network: The average shortest path length is relatively short (3.662 ± 0.00047), similar to what is observed in 'small-world' networks.
   - Implication: This indicates that despite the randomness, the network maintains efficiency in information or influence spread. If our observed network has a similar or shorter path length, it too exhibits small-world characteristics, facilitating rapid communication or connectivity among nodes.
3. **Degree Distribution :**
   - Random Network: The minimum and maximum degrees (1 and 273, respectively) and the average degree (16.85) match our observed network due to the degree-preserving nature of the Configuration Model.
   - Implication: This similarity in degree metrics highlights that the deviations observed in other network properties (like clustering coefficient and path length) are not due to differences in degree distribution but likely due to the network's structural organization.

In summary, our comparative analysis between the observed network and randomized models generated by the Configuration Model illuminates the unique structural characteristics inherent to our network. These attributes include notably enhanced clustering and an efficient dissemination of information, features distinctly absent in the randomized counterparts. The inability of the random model to replicate these specific aspects underscores the existence of non-random patterns and potential organizing principles governing the observed network's structure. Moreover, despite the constraints posed by hardware limitations and the substantial size of our dataset, which necessitated a cap on the generation of random network ensembles, our methodological approach facilitated a concentrated examination. This analysis effectively accentuates the distinctive features of the observed network, distinguishing it from those generated through random chance, and underscoring the insightful nuances brought to light through our study.

Given the intrinsic characteristics of our dataset, particularly the utilization of the data field of nodes for the evaluation of diverse community detection algorithms, an exhaustive comparison of our project's findings with those derived from null models was rendered unfeasible. The Configuration Model, a degree-preserving method selected for its relevance to our study, inherently replicates the degree distribution of our observed network. Consequently, applying this model for centrality analysis and collaboration pattern investigations would invariably yield identical outcomes. As such, embarking on comparative analyses against null models for these specific facets of our research would not engender meaningful differentiation or insights. This limitation underscores the need for alternative approaches or models that might allow for a more nuanced examination of the network's structural properties beyond mere degree distribution.

8 Discussion

- **Framing Results in Existing Literature :** Our investigation into the network of actors within the Netflix catalog revealed distinct community structures, centrality measures, and collaboration patterns. Similar studies in the realm of social networks and organizational theory have long emphasized the importance of socio-cultural factors in shaping collaboration networks [2]. By identifying communities predominantly defined by countries, our findings align with the notion that the entertainment industry, much like other social systems, is influenced by socio-cultural dynamics, echoing the segmentation observed in global markets [1]. The identification of key actors through centrality analysis further supports the concept of "influencers" within networks, whose roles in enhancing cohesion and facilitating information flow have been extensively documented [4].

- **Insights from the Network :** The community detection analysis illuminated the entertainment industry's segmented nature, indicating that collaborations are not merely driven by professional dynamics but are significantly influenced by socio-cultural factors. Centrality analysis uncovered that certain actors, such as **Anupam Kher** and **Samuel L. Jackson**, hold central positions in the network, serving as critical connectors within and possibly across communities. The recognition of actors with high betweenness centrality, notably **Anupam Kher** and **Om Puri**, underscores the existence of bridge figures who facilitate connections across diverse clusters, enriching the network's interconnectedness.

- **Answering Research Questions :** Our analysis successfully addressed the outlined research questions by:
  - Identifying distinct communities within the actor network, revealing the influence of socio-cultural dynamics on collaboration patterns.
  - Highlighting actors with significant centrality, indicating the pivotal roles they play in maintaining network cohesion and connectivity.
  - Unveiling patterns of frequent collaborations, which hint at the industry's preference for stable working relationships.

- **Limitations and Future Directions :** While our study offers valuable insights, it is not without limitations. The analysis was constrained by the dataset's scope, focusing on Netflix titles and may not fully capture the global entertainment industry's complexity. Additionally, the degree to which our findings can be generalized to other platforms or media forms remains an open question. Future work could extend this analysis to include a broader range of data sources, potentially incorporating social media interactions or audience reception data to paint a more comprehensive picture of the actors' influence beyond mere collaborations. Exploring dynamic network analysis to track changes over time could also offer deeper insights into the evolving nature of collaboration patterns within the industry. Also, investigating the impact of network position on actors' career trajectories and content success rates could yield actionable intelligence for industry stakeholders.

9 Methods

For the analysis of our graph, we utilized various methods from NetworkX to gather key metrics. These included the number of connected components, nodes, edges, average degree, minimum and maximum degree, average clustering coefficient, and average shortest path. Additionally, we calculated degree centrality and betweenness centrality, as well as examined collaboration patterns within the graph. NetworkX provides built-in functions to easily obtain these metrics, followed by further calculations as needed.

Regarding community detection, we initially selected 5 algorithms from the NetworkX community package to apply to our graph. However, after a significant runtime of approximately 23 hours, we decided to exclude Girvan Newman and Naïve Greedy Modularity algorithms due to their time-consuming nature. The remaining 3 algorithms were executed, taking nearly 15 minutes on our local machine. For each algorithm, a new field was added to the nodes of the network, which would later help evaluate the algorithm's performance.

To evaluate the results of the community detection algorithms, we needed a suitable metric. Initially, we attempted several clustering metrics using the nodes' community field as predictions and the country field as labels. The country names were label-encoded to integers for this comparison. However, an issue arose as the number of communities generated did not match the number of distinct countries in the network. Consequently, most metrics, except for two, could not be applied. The remaining metrics also encountered issues as they assumed a one-to-one mapping between community numbers and country numbers, which was not the case in our data.

As a solution, we developed our own metric termed 'Weighted Community Similarity'. This metric aims to measure how similar nodes are within each community. For each community, we identified the most frequently occurring country and calculated its ratio within the community. We then aggregated these ratios for all communities, using the number of nodes in each community as a weight. This resulted in a weighted average, giving consideration to the sizes of the communities. This metric indicated that Asynchronous Label Propagation was better suited to our network, as it yielded a higher weighted community similarity score. This suggests that nodes were more effectively clustered with this algorithm.

10 Code

The preprocessing, network construction, and network analysis components of our project have been meticulously developed using Python within a Jupyter Notebook environment. Each segment of the code has been comprehensively commented to elucidate its functionality and rationale, thereby enhancing readability and facilitating a deeper understanding for future researchers and collaborators. To ensure accessibility and foster collaborative review, the complete source code, along with its detailed comments, has been duly archived and is available for public access on Github. This repository encompasses all scripts and notebooks essential for replicating the analyses and results presented in our study.

References

[1] Mark EJ Newman. The structure of scientific collaboration networks. *Proceedings of the national academy of sciences*, 98(2):404–409, 2001.

[2] Stephen P Borgatti and Daniel S Halgin. On network theory. *Organization science*, 22(5):1168–1181, 2011.

[3] Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3-5):75–174, 2010.

[4] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440–442, 1998.