Netflix Movie Network Analysis

Final Report

CPSC 572: Fundamentals of Network Analysis and Data Mining

Authors - Jose Perales, Ishrat Naba, Simrat S. Benipal

Course Instructor - Dr. Emma K. Towlson

04 April 2024

University of Calgary

Winter 2024

UNIVERSITY OF CALGARY

Team

Group # 11

Name	Course	UCID	E-Mail
Jose Perales	CPSC 572	30143354	jose.peralesrivera@ucalgary.ca
Ishrat Naba	CPSC 572	30123121	ishrat.naba@ucalgary.ca
Simrat Benipal	CPSC 572	30129328	simrat.benipal@ucalgary.ca

UNIVERSITY OF CALGARY

Acknowledgements

We would like to express our sincere gratitude to **Dr. Emma K. Towlson, Abbas Omidi, and Dev Patel** for their invaluable guidance, encouragement, and assistance. Their mentorship was instrumental in shaping the project.

Project Summary

The digital era has seen Netflix rise as a major player in the entertainment industry[1], redefining the dynamics of content production and distribution. This project analyzes the underlying network of collaborations that connects films, actors, and directors. In the realm of network science, such an examination unveils the structural nuances and interaction patterns within the world's leading streaming service.

The network in question, comprising titles and production personnel from the Netflix catalog up to 2021, reflects a tripartite structure with nodes representing movies, directors, and actors. The study primarily focuses on movies, acknowledging the wider expanse of content types on the platform as a direction for future research.

Our analysis revealed a notable clustering coefficient within the network, indicative of tight-knit collaborative groups that deviate significantly from the sparseness of a corresponding null model[2]. The average path length suggested a network marked by both efficiency and compartmentalization, characteristic of the film industry's collaborative practices.

Addressing our central research questions, distinct communities within the network were identified, largely delineated by geographical factors, affirming the localized nature of film industry collaborations. Moreover, key individuals were distinguished as conduits between these communities.

These findings, while insightful, come with the caveat of data limitations and the absence of recent trends. Future investigations could expand upon this foundation by incorporating newer data and broadening the analysis to other content forms and platforms. This would enable a more dynamic understanding of the evolving network patterns and the strategic implications for content providers in the competitive landscape of digital streaming.

The implications of our work resonate with the shift towards data-driven approaches in understanding the complexities of creative industries. By deciphering the structure and intricacies of the Netflix movie network, we contribute to the broader discourse on how such platforms shape cultural consumption and production.

[1] https://explodingtopics.com/blog/video-streaming-stats
[2] https://en.wikipedia.org/wiki/Null_model

Research questions

There are a lot of different things to explore and potential areas of interest. However, to maintain clear goals for this project, the following were identified as the key questions for the project:

1. Can distinct communities be identified within the Netflix movie network, and if so, what characteristics (e.g. country of origin, genre, release year), define these communities?

    a. To answer this question, a focus will be taken on the community structure of the Netflix movie network. To detect distinct communities, the Louvian method[3] will be used. This method maximizes the modularity value in order to find community structures and is well suited for large networks.

2. Who are the key individuals within the Netflix movie network, and how do they facilitate connections between different communities?

    a. For this question, the betweenness centrality[4] will be taken as a measure of how often a node is used to transfer information between different communities. This question helps us answer which individuals "act as a bridge" between different communities.

[3] https://en.wikipedia.org/wiki/Louvain_method
[4] https://neo4j.com/docs/graph-data-science/current/algorithms/betweenness-centrality/

Introduction

The advent of digital streaming platforms has catalyzed a transformation in the entertainment industry, with Netflix emerging as a trailblazer in the space of online content delivery. Since its inception as a DVD-by-mail service in 1997 by Reed Hastings and Marc Randolph[5], Netflix[6] has grown exponentially into a giant in the space with over 260 million active subscribers[7], offering a wide array of cinematic works that cross genres, languages, and international borders. This presents an opportunity for examining the interconnectivity and the collaborative network that forms the backbone of this streaming giant.

In the scholarly realm, the analysis of such networks is not unprecedented. Studies in the field of network analysis and data mining have long been concerned with the architecture of various social and professional networks, with recent literature extending this interest to the domain of creative industries[8]. The network dynamics of collaboration and influence among actors, directors, and films, are indicative of broader patterns of human interaction and socio-professional engagement.

Our study delves into this type of networks, focusing on the Netflix movie network—a network composed of cinematic productions and the individuals behind them. The dataset, which contains movies up to the year 2021, is a repository of titles streamed on Netflix alongside the individuals that played a role in their production[9]. However, it is important to note the limitations of our dataset; the stop put by Netflix to data collection through APIs puts a temporal bracket on our findings and implies that our analysis does not capture the platform's most recent evolution.

Framing our study within the context of these limitations, we look to answer two key questions: the discernment of distinct communities[10] within the Netflix movie network and the identification of pivotal individuals who facilitate connections within this network. These questions resonate with the studies that examine modularity and betweenness centrality as a measure of connectivity and influence in networks. Our approach, utilizing methods such as the Louvain algorithm[11] for community detection, and appropriate comparisons with null models[12], allows us to probe the underlying structure of the Netflix network.

While our project establishes a foundational understanding of the Netflix movie network, future research can propel this forward. Extending the dataset to include more recent years and diverse forms of content could provide a more holistic view of the network's evolution. Additionally, a comparative analysis with other streaming

[5] https://en.wikipedia.org/wiki/Netflix
[6] https://www.netflix.com/ca/
[7] https://www.statista.com/statistics/250934/quarterly-number-of-netflix-streaming-subscribers-worldwide
[8] https://www.mdpi.com/2504-3900/85/1/23
[9] https://www.kaggle.com/datasets/shivamb/netflix-shows/data
[10] https://en.wikipedia.org/wiki/Community_structure
[11] https://en.wikipedia.org/wiki/Louvain_method
[12] https://en.wikipedia.org/wiki/Null_model

platforms[13] could offer insights into the competitive landscape of digital entertainment.

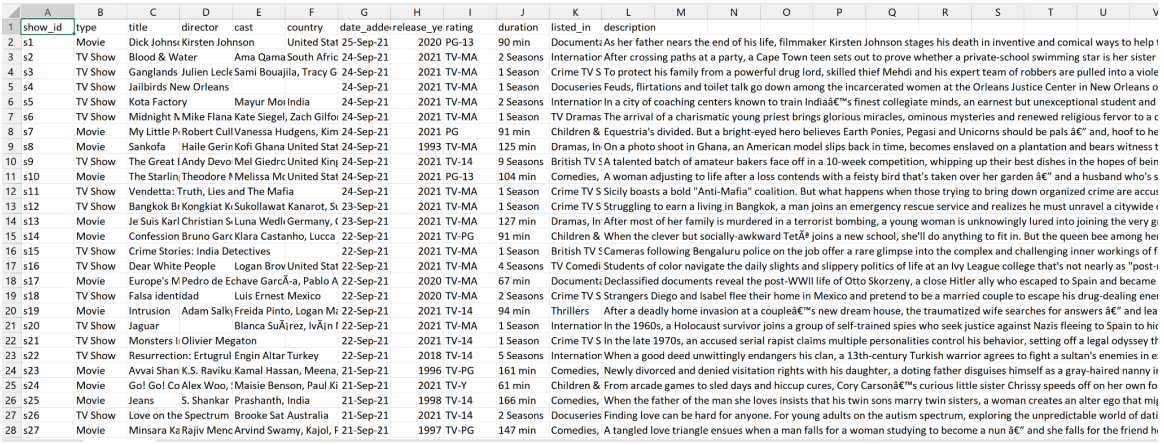
As we position our study within the broader discourse on network analysis in digital entertainment, we strive to contribute a piece to the puzzle of how modern streaming platforms are shaping the consumption and production of cinematic content. This, in turn, lays the groundwork for future explorations into the growing narrative of network structures, marked by real-world constraints and the proclivity for creative collaboration.

[13] https://en.wikipedia.org/wiki/List_of_streaming_media_services

UNIVERSITY OF CALGARY

Dataset description

The dataset analyzed in this project was taken from Kaggle[14]. It can be downloaded and used without any restrictions, as no permission is required to make use of it. The data itself is a single CSV (comma-separated values) file containing information on all movies and TV shows on Netflix up to 2021. The following is a screenshot of the raw CSV file opened with Excel:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	show_id	type	title	director	cast	country	date_adde	release_ye	rating	duration	listed_in	description											
2	s1	Movie	Dick Johns	Kirsten Johnson		United Stat	25-Sep-21	2020	PG-13	90 min	Documenta	As her father nears the end of his life, filmmaker Kirsten Johnson stages his death in inventive and comical ways to help t										
3	s2	TV Show	Blood & Water		Ama Qama	South Afric	24-Sep-21	2021	TV-MA	2 Seasons	Internatior	After crossing paths at a party, a Cape Town teen sets out to prove whether a private-school swimming star is her sister										
4	s3	TV Show	Ganglands	Julien Lecle	Sami Bouajila, Tracy G		24-Sep-21	2021	TV-MA	1 Season	Crime TV S	To protect his family from a powerful drug lord, skilled thief Mehdi and his expert team of robbers are pulled into a viole										
5	s4	TV Show	Jailbirds New Orleans				24-Sep-21	2021	TV-MA	1 Season	Docuseries	Feuds, flirtations and toilet talk go down among the incarcerated women at the Orleans Justice Center in New Orleans o										
6	s5	TV Show	Kota Factory		Mayur Mor	India	24-Sep-21	2021	TV-MA	2 Seasons	Internatior	In a city of coaching centers known to train Indiaâ€™s finest collegiate minds, an earnest but unexceptional student and										
7	s6	TV Show	Midnight M	Mike Flana	Kate Siegel, Zach Gilfo		24-Sep-21	2021	TV-MA	1 Season	TV Dramas	The arrival of a charismatic young priest brings glorious miracles, ominous mysteries and renewed religious fervor to a c										
8	s7	Movie	My Little P	Robert Cull	Vanessa Hudgens, Kim		24-Sep-21	2021	PG	91 min	Children &	Equestria's divided. But a bright-eyed hero believes Earth Ponies, Pegasi and Unicorns should be pals â€" and, hoof to he										
9	s8	Movie	Sankofa	Haile Gerin	Kofi Ghana	United Stat	24-Sep-21	1993	TV-MA	125 min	Dramas, In	On a photo shoot in Ghana, an American model slips back in time, becomes enslaved on a plantation and bears witness t										
10	s9	TV Show	The Great I	Andy Devo	Mel Giedrc	United Kinç	24-Sep-21	2021	TV-14	9 Seasons	British TV S	A talented batch of amateur bakers face off in a 10-week competition, whipping up their best dishes in the hopes of bein										
11	s10	Movie	The Starlin	Theodore M	Melissa Mc	United Stat	24-Sep-21	2021	PG-13	104 min	Comedies,	A woman adjusting to life after a loss contends with a feisty bird that's taken over her garden â€" and a husband who's s										
12	s11	TV Show	Vendetta: Truth, Lies and The Mafia				24-Sep-21	2021	TV-MA	1 Season	Crime TV S	Sicily boasts a bold "Anti-Mafia" coalition. But what happens when those trying to bring down organized crime are accus										
13	s12	TV Show	Bangkok Br	Kongkiat Kc	Sukollawat	Kanarot, Su	23-Sep-21	2021	TV-MA	1 Season	Crime TV S	Struggling to earn a living in Bangkok, a man joins an emergency rescue service and realizes he must unravel a citywide c										
14	s13	Movie	Je Suis Karl	Christian Sc	Luna Wedl	Germany, (	23-Sep-21	2021	TV-MA	127 min	Dramas, In	After most of her family is murdered in a terrorist bombing, a young woman is unknowingly lured into joining the very g										
15	s14	Movie	Confession	Bruno Garc	Klara Castanho, Lucca		22-Sep-21	2021	TV-PG	91 min	Children &	When the clever but socially-awkward TetÃ³ joins a new school, she'll do anything to fit in. But the queen bee among he										
16	s15	TV Show	Crime Stories: India Detectives				22-Sep-21	2021	TV-MA	1 Season	British TV S	Cameras following Bengaluru police on the job offer a rare glimpse into the complex and challenging inner workings of f										
17	s16	TV Show	Dear White People		Logan Brov	United Stat	22-Sep-21	2021	TV-MA	4 Seasons	TV Comedi	Students of color navigate the daily slights and slippery politics of life at an Ivy League college that's not nearly as "post-r										
18	s17	Movie	Europe's M	Pedro de E	chave GarcÃ-a, Pablo A		22-Sep-21	2020	TV-MA	67 min	Documenta	Declassified documents reveal the post-WWII life of Otto Skorzeny, a close Hitler ally who escaped to Spain and became										
19	s18	TV Show	Falsa identidad		Luis Ernest	Mexico	22-Sep-21	2020	TV-MA	2 Seasons	Crime TV S	Strangers Diego and Isabel flee their home in Mexico and pretend to be a married couple to escape his drug-dealing ene										
20	s19	Movie	Intrusion	Adam Salky	Freida Pinto, Logan Ma		22-Sep-21	2021	TV-14	94 min	Thrillers	After a deadly home invasion at a coupleâ€™s new dream house, the traumatized wife searches for answers â€" and lea										
21	s20	TV Show	Jaguar		Blanca SuÃ¡rez, IvÃ¡n I		22-Sep-21	2021	TV-MA	1 Season	Internatior	In the 1960s, a Holocaust survivor joins a group of self-trained spies who seek justice against Nazis fleeing to Spain to hic										
22	s21	TV Show	Monsters I	Olivier Megaton			22-Sep-21	2021	TV-14	1 Season	Crime TV S	In the late 1970s, an accused serial rapist claims multiple personalities control his behavior, setting off a legal odyssey th										
23	s22	TV Show	Resurrection: Ertugrul	Engin Altar	Turkey		22-Sep-21	2018	TV-14	5 Seasons	Internatior	When a good deed unwittingly endangers his clan, a 13th-century Turkish warrior agrees to fight a sultan's enemies in e										
24	s23	Movie	Avvai Shan	K.S. Raviku	Kamal Hassan, Meena,		21-Sep-21	1996	TV-PG	161 min	Comedies,	Newly divorced and denied visitation rights with his daughter, a doting father disguises himself as a gray-haired nanny i										
25	s24	Movie	Go! Go! Co	Alex Woo, :	Maisie Benson, Paul Ki		21-Sep-21	2021	TV-Y	61 min	Children &	From arcade games to sled days and hiccup cures, Cory Carsonâ€™s curious little sister Chrissy speeds off on her own fo										
26	s25	Movie	Jeans	S. Shankar	Prashanth, India		21-Sep-21	1998	TV-14	166 min	Comedies,	When the father of the man she loves insists that his twin sons marry twin sisters, a woman creates an alter ego that mi										
27	s26	TV Show	Love on the Spectrum	Brooke Sat	Australia		21-Sep-21	2021	TV-14	2 Seasons	Docuseries	Finding love can be hard for anyone. For young adults on the autism spectrum, exploring the unpredictable world of dati										
28	s27	Movie	Minsara Ka	Rajiv Menc	Arvind Swamy, Kajol, F		21-Sep-21	1997	TV-PG	147 min	Comedies,	A tangled love triangle ensues when a man falls for a woman studying to become a nun â€" and she falls for the friend h										

For the data cleaning process, two functions were created. The first one is *parseData*. This function opens the CSV file, reads each row as a dictionary, skips TV Shows, and deletes unnecessary fields (show_id, type, country, and description). It then returns the clean data as a list of dictionaries, where each dictionary represents a movie. The second function is *outputDataAsJSON*, which takes in the clean data, given by *parseData*, and creates a JSON file containing all the clean data into a specified output path. Called sequentially, these two functions clean the raw CSV data and yield a clean JSON file.

Once the data had been cleaned and formatted properly, a NetworkX graph was created by iterating over all the movies in the JSON file. The movie's name along with all its cast members and directors were added as nodes, and edges were added between movies and cast members, movies and directors, and between cast members and directors. This means that the network is tripartite, with the different types of nodes being "cast members", "directors", and "movies", that no edges exist within each type (i.e. movie to movie, director to director), and that the network is undirected (every connection between two nodes goes both ways).

Put formally, the network is undirected and tripartite, and its 3 different types of nodes are:

- **Movies:** All the movies in the dataset. Each one is represented with the movie's title.
- **Cast members:** The actors and actresses involved in each of the movies.
- **Directors:** The director(s) of each movie.

And edges exist only between:

- Movies and all their cast members.
- Movies and all their directors.
- Directors and all the cast members they have worked with.

[14] https://www.kaggle.com/datasets/shivamb/netflix-shows/data

Once the network was created, NetworkX's built-in function *create_gexf* was used to create a GEXF (Graph Exchange XML) file which stored the network along with all the information related to the nodes. The GEXF file was then used in Gephi to visualize and further analyze the network. It is important to note that the network does not contain any metadata.

Basic Statistics

Nodes and edges

The network contains 35,133 nodes and 95,500 edges. Out of all the nodes, 72.85% (25,594) are cast members, 15.71% (5,519) are movies, and 11.45% (4,022) are directors. The number of edges suggests a dense network of interactions, which is typical of creative industries (not unlike the film industry) where collaboration is key.

Number of connected components

The number of connected components in a network represents the count of distinct sub-networks where any two nodes within the same sub-network can reach each other through a path of edges, but there is no such path between nodes of different sub-networks.

In the Netflix movie network, each connected component represents a set of movies and the individuals involved in their creation that are interconnected. The presence of 512 such components paints a picture of a network with multiple independent or loosely connected sub-networks.

This can be attributed to several factors, such as:

- **Specialized productions**: There may be exclusive partnerships and productions limited to specific individuals, reflecting industry niches or unique cinematic ventures. For instance, independent films or projects spearheaded by a single auteur may result in distinct components within the network.
- **One-time collaborations**: The network might include one-off collaborations where an actor or director worked on a single project, leading to the formation of standalone components that do not link to the broader network.

This offers opportunities for strategic expansion for streaming platforms like Netflix, where identifying new connections that could bridge isolated components would promote cross-collaboration between different film communities.

Degree distribution

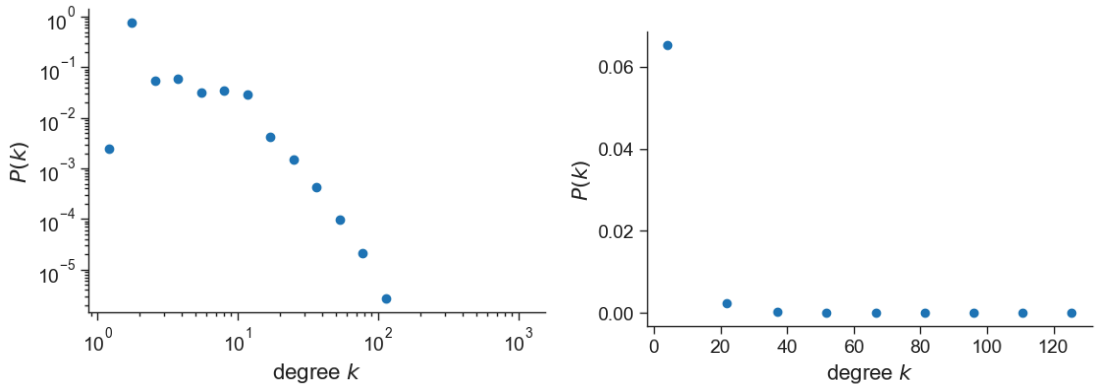
The degree distribution of a network illustrates the number of connections each node has within the network, which in this case ranges from 1 up to 133. The lower end of this spectrum typically represents standalone projects. For instance, the movie "WHAT DID JACK DO?" is a one-person movie with David Lynch taking on the role of both director and actor, creating an isolated component in the network with two nodes of degree one (the movie’s title and “David Lynch”).

Conversely, a node with a high degree indicates a prolific individual or project in the network. High-degree nodes often represent well-known entities, such as esteemed directors who have worked across numerous films or actors who have extensive filmographies within the dataset's scope. These nodes act as central hubs, influencing the network's structure and dynamics due to their extensive connections.

UNIVERSITY OF CALGARY

Analyzing the entire network's degree distribution graphs (shown below), reveal that the majority of nodes have fewer than five connections, indicating a large number of one-off or limited collaborations. In contrast, nodes with 40 or more connections are scarce, emphasizing the exceptional status of a few individuals or projects.

Overall average degree distribution graphs:



This distribution follows a pattern often observed in real-world networks, where many participants have limited connections, while a few have a vast number of links. This pattern is characteristic of a scale-free network, suggesting that the Netflix movie network may be influenced by preferential attachment[15] – a tendency for new nodes to connect to already well-connected ones.

However, due to the network's unique structure, it became apparent that a different approach would more accurately capture the web of connections. The network, inherently tripartite, requires a detailed examination to truly understand the relationship dynamics.

To provide a more detailed analysis, the network was segmented into three distinct bipartite networks: actors-directors, actors-movies, and directors-movies. This partitioning provided the average degree specific to each node type in their relevant contexts.

Average degree per type:

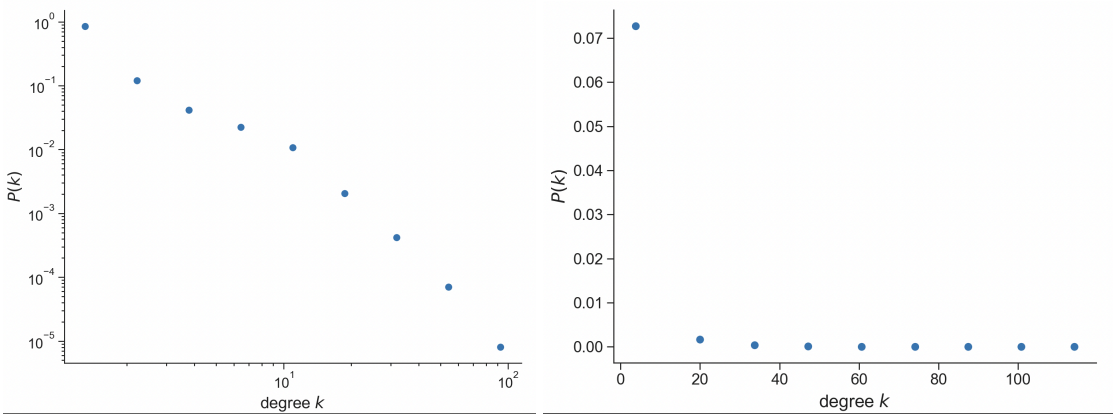
- **Movies:** Average degree of 8.28, indicating each movie is connected to approximately 8 actors and directors, reflecting the collaborative nature of film production.
- **Directors:** With an average degree of 10.99, directors are shown to engage with a high number of movies and actors, emphasizing their central role in the industry.
- **Cast members:** An average degree of 3.63 suggests that there may be less opportunity to participate in projects for cast members as compared to directors.

Focused Bipartite Network Insights:

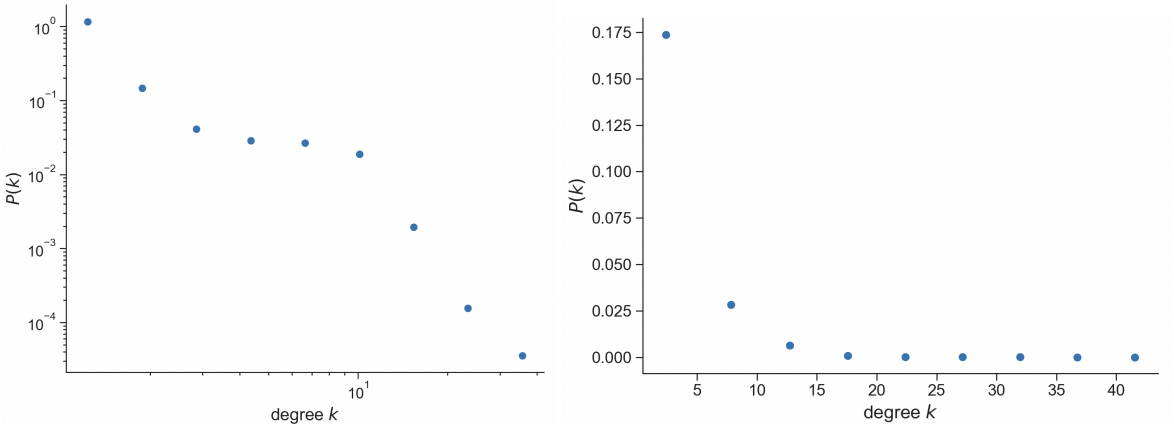
- **Cast-Directors Network:** Directors have an average degree of 9.57, demonstrating collaboration with a diverse range of cast members. Cast members have an average degree of 1.91, indicating more targeted collaborations, likely driven by specific

[15] https://en.wikipedia.org/wiki/Preferential_attachment

directorial visions or project requirements.

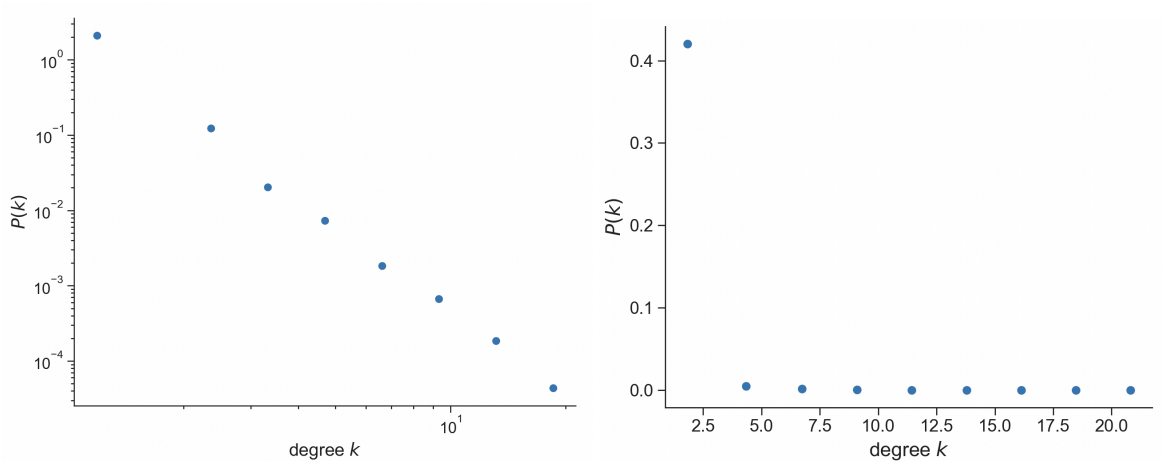


- **Cast-Movies Network:** Cast members hold an average degree of 1.71, signifying their involvement in a small number of movies, which could reflect selective project choices or competitive casting practices. Movies, with an average degree of 7.25, highlight the extensive cast typically involved in productions.



- **Directors-Movies Network:** Directors have an average degree of 1.39, pointing to a focused engagement with a limited number of movies, potentially showing the depth of involvement in each project. Movies, at an average degree of 1.08, suggest a predominantly singular directorial oversight, reaffirming the traditional model of film direction.

UNIVERSITY OF CALGARY



Clustering coefficient

The clustering coefficient is a critical metric, measuring the extent to which nodes in a network tend to cluster together. It quantifies the probability that two neighbors of a node are also neighbors of each other, forming a closed triangle of connections. This coefficient is particularly revealing in networks where the formation of collaborative groups is a fundamental characteristic of the system's architecture, as in the Netflix movie network.

Our analysis uncovered a stark difference in the clustering coefficients when comparing the Netflix movie network and the configuration null model. Specifically:

- Netflix movie network = 0.64482
- Configuration null model = 0.00072

The difference between the clustering coefficients of the Netflix movie network and the configuration null model shows the propensity for clustering in the actual network. With a clustering coefficient of 0.64482, the Netflix movie network exhibits a tendency towards tight-knit collaboration, a characteristic feature of creative industries where professionals frequently work in close cohorts.

In contrast, the configuration null model, with a clustering coefficient of 0.00072, indicates an almost negligible tendency for nodes to form triadic closures. The null model maintains the degree sequence of the original network but randomizes the connections, eliminating the structural dependencies and collaborative patterns that naturally arise in real-world settings.

So why do the clustering coefficients differ?

- **Preferential attachment[16]:** The high clustering coefficient in the Netflix movie network can be attributed to preferential attachment, where established actors and directors are more likely to collaborate with one another, forming a closely-knit community. This phenomenon is often driven by trust, shared artistic vision, and established reputation, leading to recurrent collaborations within a relatively exclusive group of individuals.

[16] https://en.wikipedia.org/wiki/Preferential_attachment

- **Industry constraints:** The film industry often operates within clusters defined by language, genre, and geography. These natural barriers promote higher clustering as individuals navigate within familiar and accessible circles.

The low clustering coefficient in the null model is expected, as it lacks any preferential connectivity that would lead to closed triangles. The random nature of link assignments in the null model serves as a baseline, highlighting that the clustering observed in the actual network is far from random and is instead likely driven by other factors.

The high clustering coefficient reveals the network's tendency toward cliquishness. It reflects a propensity for individuals to work within established circles, leading to concentrated clusters of collaboration. On one hand, this can foster in-depth collaboration and a consistent quality of work; on the other hand, it might limit the diversity of creative input by maintaining a relatively closed network of repeated collaborations. Understanding this balance is important for platforms, like Netflix, aiming to offer a rich and varied catalog while encouraging an innovative creative environment.

Path Length

Path length is a crucial measure in network analysis, denoting the average number of steps along the shortest paths for all possible pairs of network nodes. It's a measure of the network's efficiency in terms of information or relationship flow between nodes. In the Netflix movie network, this statistic reveals how closely connected the entities within the network truly are.

Upon comparing the average path length within the largest connected component of the Netflix movie network with that of the configuration null model, we observe a significant variation:

- Netflix movie network: 8.65
- Configuration null model: 5.24

So why do the path lengths differ?

- **Structural Constraints**: The longer path lengths within the network reflect the presence of structural and collaborative constraints. Unlike the null model, real-world networks like Netflix's are shaped by factors such as geographic distribution, language barriers, and genre-specific collaborations. These constraints naturally lengthen the path between nodes.

- **Exclusive Networks**: The film industry tends to form creative clusters where actors, directors, and producers repeatedly work within a select group, leading to exclusive networks. These clusters, while fostering in-depth collaboration, can result in longer paths as they limit cross-cluster interactions.

- **Preferential Attachment**: High-profile individuals often attract more collaborations, creating hubs that centralize the network's connections. While these hubs shorten paths within their vicinity, they can lengthen the overall average path length by creating several high-density regions loosely connected to each other.

UNIVERSITY OF CALGARY

This suggests that while there is a decent level of efficiency in the industry's connectivity, there's room for improvement, indicating potential for increasing cross-collaboration, which could enhance innovation and diversify the creative output.

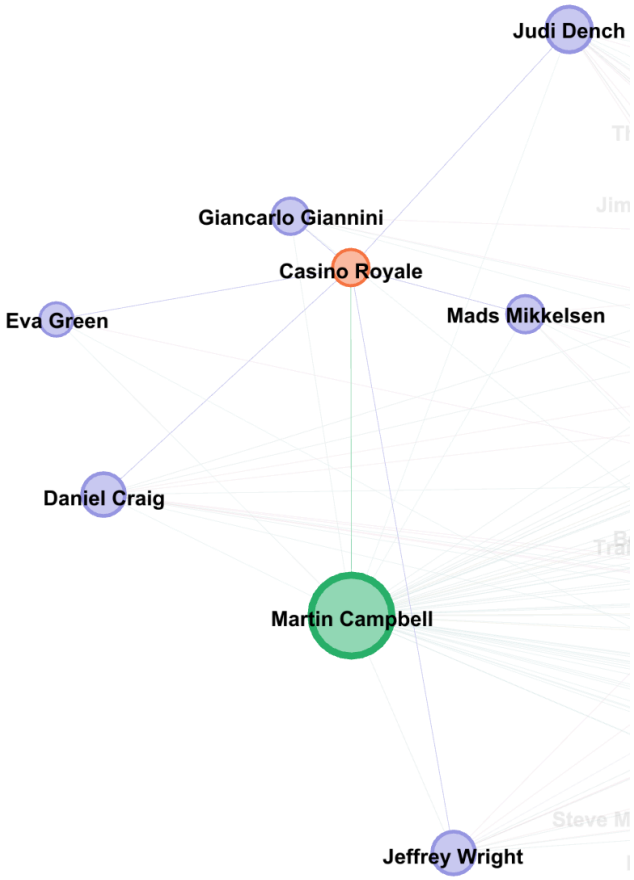
Understanding the factors that contribute to the increased path length in the network can provide valuable insights for Netflix and other content providers to encourage new collaborations that bridge existing clusters.

UNIVERSITY OF CALGARY

Network visualization

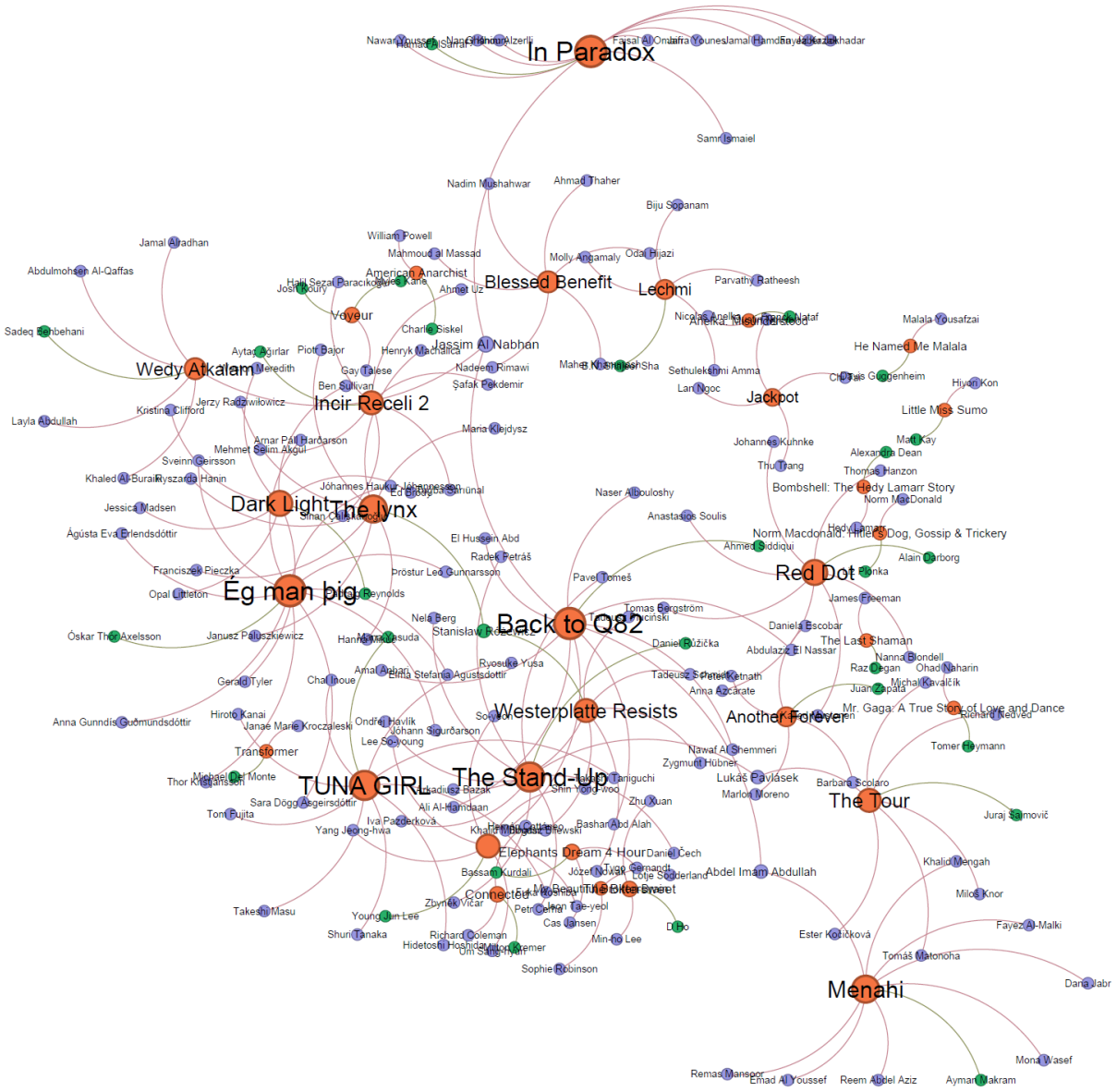
Gephi[17], an open-source software tool, was used for visualizing the network. The GEXF file was given as input and network visualizations were created. In the network visualizations, movies are represented with orange, directors with green, and cast members with violet.

The following is the subgraph of the movie “Casino Royale” (orange node). Martin Campbell, the director, is shown in green, while the cast is shown in violet. To make visualizations easier to understand, each node’s size varies according to its degree. The more connections a node has, the bigger it is.

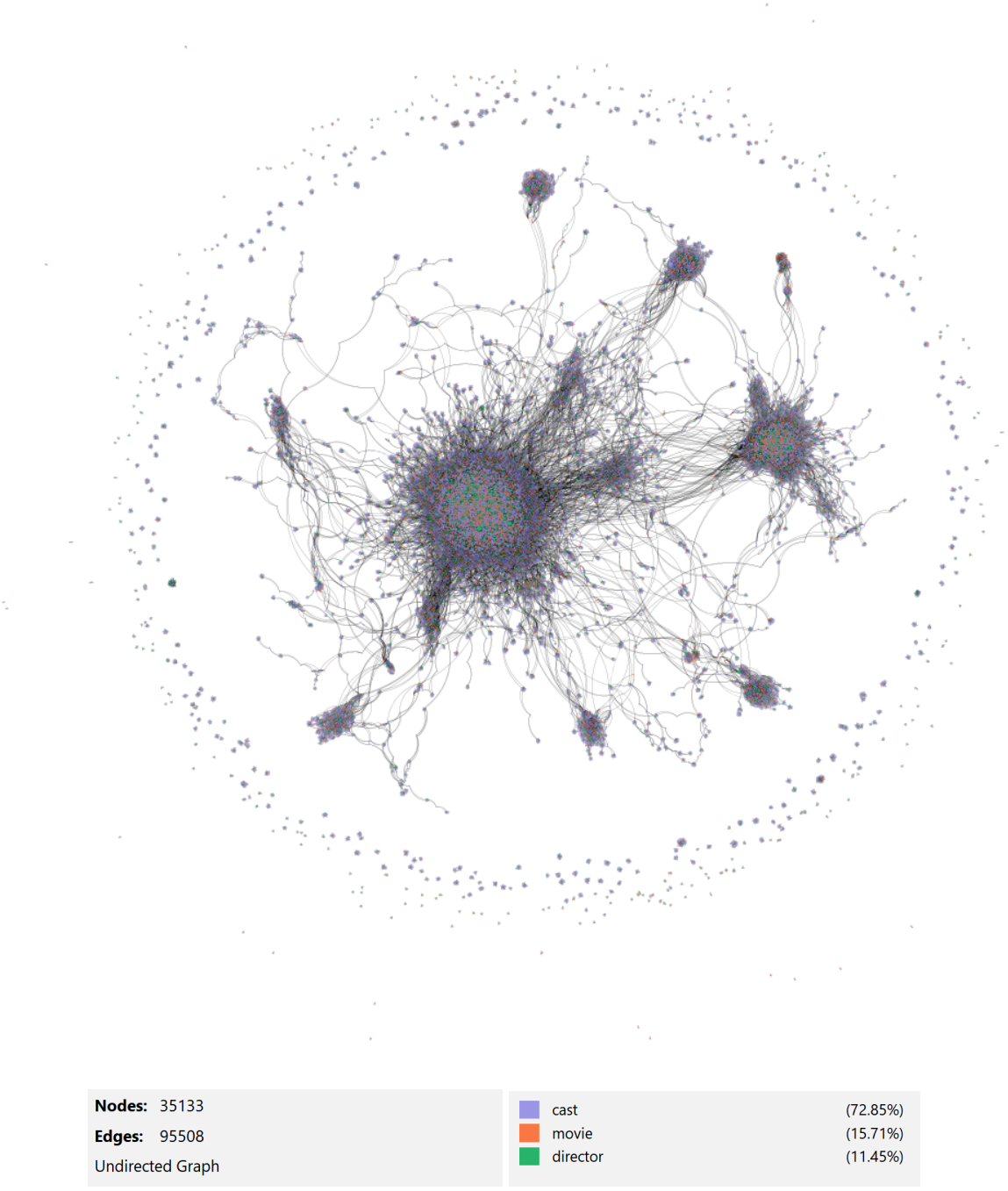


[17] https://gephi.org/

The image below shows the network at a higher level. Connections between people involved in different movies can be observed. The edges between actors and directors were omitted to simplify the image.



Visualization of the entire network:



At a glance, the network's structure reveals a series of dense clusters interspersed across a web of more sparsely connected nodes. These clusters, visually represented by nodes drawn closely together, suggest thriving communities within the network. Each community is likely defined by shared attributes — perhaps a common genre, production style, or geographical location. Such groupings are characteristic of homophily[18], the principle that entities tend to associate with others that are similar to themselves. In this context, it might mean that individuals tend to collaborate with the same people over and over (because of shared

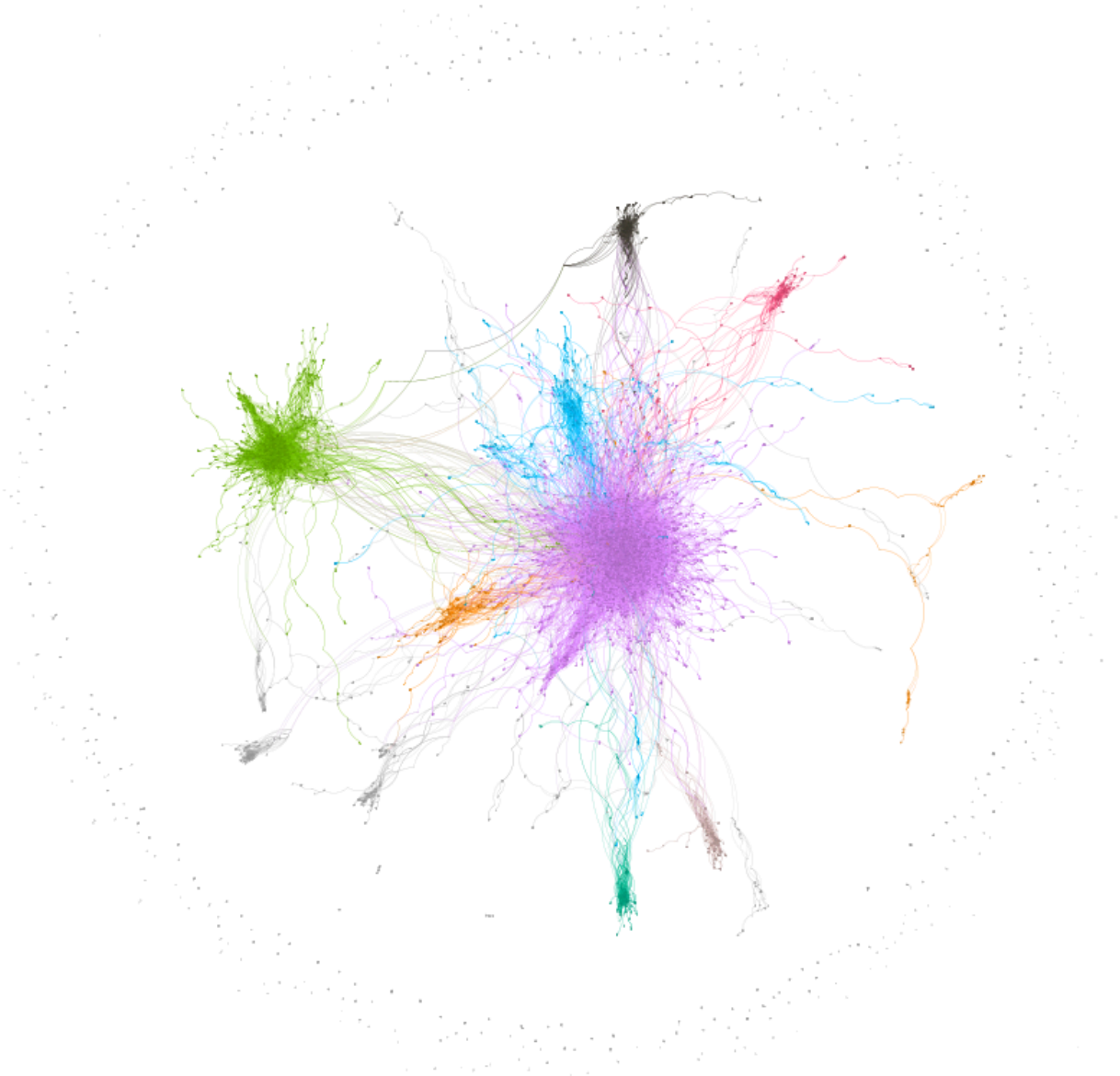
[18] https://en.wikipedia.org/wiki/Homophily

experiences, interests, etc.) or within the same genre, creating discernible 'neighborhoods'.

In each communities' heart are the densely packed nodes, hubs of activity where numerous movies, directors, and actors interconnect. These high-degree nodes often represent important movies or influential industry figures that have worked with a wide array of actors and directors.

As previously discussed, the network has 512 connected components, meaning that the dots on the periphery are the other 511 connected components (besides the largest one). These are typically one off or debut projects, where the cast and director were relatively unknown at the time.

The apparent emergence of different communities in the network's visualization suggests that the answer to the first research question, whether different communities exist and if they can be identified, is yes, and prompted the creation of a visualization where each community is clearly differentiated from the rest. The image below is said visualization:



Results

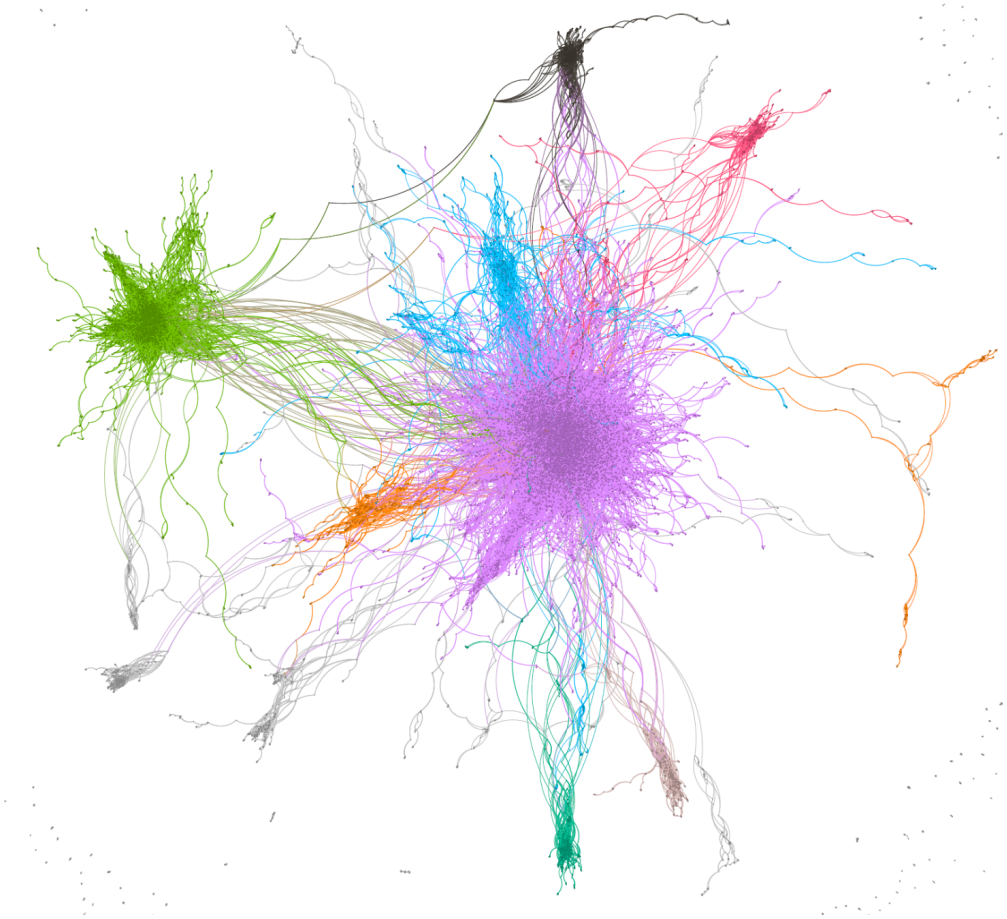
In order to address the first research question, the Louvain method was employed to detect communities and differentiate them from each other. Because of the large number of nodes and edges, it was unfeasible to run other community detection algorithms. The Louvain Method for community detection is a type of greedy optimization algorithm that runs in O(n * log(n)) time, where n is the number of nodes in the graph. This made it suitable for the Netflix movie network. Due to the size of the network, the Louvain method was executed 100 times, and the modularity was recorded in each attempt. The average modularity of the network over these 100 iterations was 0.8351.

Modularity refers to the strength of the division of a network into smaller modules, these modules are also referred to as communities[19] [20]. The modularity value provides an estimate on how dense the connections are between the nodes within these communities, compared to the connections between different communities. A high modularity indicates a strong community structure, with clear and well-defined communities that are highly connected internally and relatively disconnected from each other.

The average modularity value, 0.8351, indicates a high modularity, meaning that there are clear-cut communities within the network. This was expected, since in the film industry individuals tend to or want to collaborate with well known figures that have had prolific careers, which points to the existence of preferential attachment.

The following is a zoomed in image presented in the previous section of the network after running the Louvian method.

[19] https://en.wikipedia.org/wiki/Community_structure
[20] https://en.wikipedia.org/wiki/Modularity_(networks)



From this image, it is clear that multiple important communities exist within the network. The nodes with the highest degrees of each community were analyzed to identify the characteristics defining these communities. After going through the 40 most connected nodes in each of the six biggest communities, it became clear that the factor distinguishing each community was its geographical location. The most important factor determining how individuals decide to collaborate with each other and on which movies they choose to or get the opportunity to participate in, turns out to be the film industry in which they participate, i.e. Nigerian film industry, Indian film industry, Latin American film industry, etc.

Each of the colors in the image shown above can be mapped to a different film industry as follows:

Color	Film Industry	% of nodes
Pink	United States (Hollywood)	51.46%
Light Green	India (Bollywood)	14.50%
Blue	Latin America	5.19%
Orange	East Asia/Hong Kong	2.63%
Black	Nigeria (Nollywood)	2.63%
Dark Green	Egypt	2.11%

Having identified the most prolific communities and answered the first question, the second research question arises naturally. Are there any key individuals facilitating connections between different communities?

A way to interpret this question is to look at which individuals have the highest betweenness centrality[21] in each community. Betweenness centrality is a measure of a node's importance in a network, calculated by the number of shortest paths that pass through that node. It indicates the node's role as a bridge within the network, potentially connecting various parts of the graph[22]. Thus, the individuals with the highest betweenness centrality are the key ones that facilitate the most connections between different communities, since the greatest amount of shortest paths flow through them.

Below are given the five individuals with the highest betweenness centrality and the highest degree per community. This is to offer a comparison and show that a higher degree does not equate to higher centrality. This is an important distinction to make for individuals looking to transfer from one film industry to another, or to gain more connections in their current one, it is more beneficial to approach different people depending on the expected outcome.

To the left are the individuals with the highest betweenness centrality, which is referred to only as 'Centrality'. The tables to the left answer the second research question. The tables to the right show the individuals with the highest degree. Both tables are sorted from highest to lowest centrality and degree, respectively.

Pink → American film industry (Hollywood)

Person	Role	Degree	Centrality
Steven Spielberg	Director	126	0.0578
Robert Rodriguez	Director	82	0.3551
Lasse Hallström	Director	74	0.0315
Don Micheal Paul	Director	88	0.0235
Quentin Tarantino	Director	80	0.0214

Person	Role	Degree	Centrality
Martin Scorsese	Director	133	0.0174
Steven Spielberg	Director	126	0.0578
Steve Brill	Director	97	0.0193
Don Micheal Paul	Director	88	0.0235
Robert Rodriguez	Director	82	0.3551

Green → Indian film industry (Bollywood)

Person	Role	Degree	Centrality
Anupam Kher	Cast	71	0.0375
Om Puri	Cast	55	0.0338
Priyanka Chopra	Cast	31	0.0287
Amrish Puri	Cast	32	0.0287
Leena Yadav	Director	28	0.0148

Person	Role	Degree	Centrality
Anurag Kashyap	Director	89	0.0055
Dibakar Banerjee	Director	72	0.0064
Anupam Kher	Cast	71	0.0375
David Dhawan	Director	67	0.0082
Ram Gopal Varma	Director	63	0.0022

Blue → Latin American film industry

[21] https://en.wikipedia.org/wiki/Betweenness_centrality
[22] https://neo4j.com/docs/graph-data-science/current/algorithms/betweenness-centrality/

Person	Role	Degree	Centrality
Maria Ripoli	Director	33	0.0109
Guillermo D. Toro	Director	37	0.0086
Diego Luna	Cast	19	0.0082
Jordi Sánchez	Cast	6	0.0071
Fernando Ayllón	Director	51	0.0045

Person	Role	Degree	Centrality
Fernando Ayllón	Director	51	0.0045
Jan Suter	Director	42	0.0023
Raúl Campos	Director	39	0.0021
Maria Ripoli	Director	33	0.0109
Alfonso Cuarón	Director	33	0.0043

Orange → East Asian/Hong Kong film industry

Person	Role	Degree	Centrality
Jon Lucas	Cast	24	0.2000
Iko Uwais	Cast	15	0.0164
Sahajak Boontha	Cast	14	0.0111
Song Kang-ho	Cast	6	0.0071
Timo Tjahjanto	Director	27	0.0071

Person	Role	Degree	Centrality
Johnnie To	Director	52	0.0054
Wong Jin	Cast	45	0.0017
Wilson Yip	Director	43	0.0044
Banagjong Pisantha	Director	37	0.0027
Dante Lam	Cast	29	0.0013

Black → Nigerian film industry (Nollywood)

Person	Role	Degree	Centrality
Adze Ugah	Director	31	0.0041
Kunle Afolayan	Cast	84	0.0041
Hamisha Daryani	Cast	13	0.0040
Ramsey Nouah	Cast	48	0.0038
Omoni Oboli	Cast	63	0.0032

Person	Role	Degree	Centrality
Kunle Afolayan	Cast	84	0.0041
Omoni Oboli	Cast	63	0.0032
Ramsey Nouah	Cast	48	0.0038
Niyi Akinmolayan	Director	48	0.0018
Kayode Kasum	Director	42	0.0016

Dark green → Egyptian film industry

Person	Role	Degree	Centrality
Omar Sharif	Cast	7	0.0154
Youssef Chahine	Director	88	0.0150
Ismail Farouk	Director	25	0.0077
Asghar Farhadi	Director	25	0.0043
Yousra El Lozy	Cast	8	0.0032

Person	Role	Degree	Centrality
Youssef Chahine	Director	88	0.0150
Sameh Abdulaziz	Director	56	0.0027
Khaled Marei	Director	35	0.0011
Wael Ehsan	Director	34	0.0029
Ahmed Nacier	Director	33	0.0009

Taking a closer look at some of these “bridge” nodes and digging deeper into their career paths, some interesting stories that highlight these findings can be found. Take, for example, the actor Om Puri, with a betweenness centrality of 0.0338. He is a Bollywood actor that appeared mostly in Hindi/Bollywood films. However, at some point he acted in a British Movie named “The Parole Officer”, which gained him many connections in the British Film community. This made him a key individual between two communities, as reflected by the network.

A different example is the case of Anupam Kher, another Bollywood actor. While he was involved mostly in Bollywood films, he also acted in “Silver Linings Playbook” and “Bend it Like Beckham”, which are Hollywood and British productions, respectively. This resulted in Anupam acting as a bridge between the Bollywood, Hollywood, and British communities. This explains why he has a betweenness centrality of 0.037, which is the second highest in the entire network.

UNIVERSITY OF CALGARY

Null Model analysis

Null Model Network/Graphs[23] refers to the type of Network graphs that offer a simplified version of a network and serves as a baseline for comparison for the network under analysis. Null model analysis can be used to show that the degree sequence, clustering, and community structure do not exist based on mere coincidence. Different types of null models exist, such as configuration model, degree-preserving null model, etc.

For our analysis, we used the configuration null model[24]. Our network holds most of the information on the connections between different nodes, and in the configuration null model, the degree sequence is preserved. By maintaining the same number of edges connected to each node, the configuration model ensures that the degree distribution remains unaltered. In addition to preserving the degree sequence, the configuration model provides scalability, it is much more efficient to produce large networks with a configuration null model. For a network as large as the Netflix movie network (35,500 nodes and 96,500 edges), having a null model that can quickly produce multiple null models is an advantage. In our analysis, ten configuration models were created using NetworkX. For each null model, the average clustering coefficient and average shortest path lengths were calculated. Finally, the results of all iterations were averaged. The results were the following:

Null Model Values:

- Clustering coefficient = 0.00072
- Average shortest path length = 5.24

Netflix movie network:

- Clustering coefficient = 0.64482
- Average shortest path length = 8.65

The network was compared with the null model's values in the "Basic Statistics" section, and the corresponding interpretations were also given there.

Clustering coefficients of the ten null models (avg. 0.0007178121):

[0.0007096905206, 0.0006449054248, 0.0006673019698, 0.0006775797337, 0.0007306867269, 0.0007519477948, 0.0008838090696, 0.0006641447473, 0.0008142893214, 0.0006337660061]

Average shortest path of the ten null models (average = 5.239819193258):

[5.241908916838, 5.237568756169, 5.238186291380, 5.246008539500, 5.24011344972, 5.24399179521, 5.238039126074, 5.236645781995, 5.2378870038608, 5.2378422718323]

[23] https://en.wikipedia.org/wiki/Null_model
[24] https://www.cs.cornell.edu/courses/cs6241/2020sp/readings/Fosdick-2018-configuration.pdf

Discussion

Our investigation into the Netflix movie network's topology revealed a complex web of interactions among movies, directors, and actors. The pronounced clustering coefficient and the sizable average path length of 8.65 distinguish the actual network from its randomized counterparts, akin to a creative domain.

We discovered that the network's structure is markedly non-random, characterized by an inherent preferential attachment and substantial clustering. These insights resonate with existing studies on social networks within creative industries, where professional ecosystems often revolve around a few individuals and tight production groups. Our findings suggest that similar mechanisms operate within the Netflix movie network, where certain actors and directors serve as central hubs, fostering clusters of collaboration that are both robust and, at times, insular.

Our research questions sought to unravel the community dynamics and identify the key individuals within the network. The application of the Louvain method confirmed the existence of distinct communities, primarily delineated by geographical lines—a reflection of the global yet compartmentalized nature of the film industry. Our second question, focusing on key individuals, was tackled with the use of betweenness centrality, revealing the pivotal roles played by certain individuals who bridge diverse film traditions.

However, our exploration was not without its limitations. The dataset's end at the year 2021 creates a temporal boundary to our conclusions, leaving recent shifts and emerging trends unaccounted for. Also, the lack of comprehensive metadata limited the depth of our network analysis, restricting the range of attributes that could have enriched our understanding of the complex interactions within the network.

Future endeavors should aim to incorporate more recent data, ideally spanning various forms of content beyond movies, to construct a more current and holistic picture of Netflix's creative landscape. There exists, as well, a compelling need for cross-platform analyses that benchmark the observed network characteristics against other streaming giants[25], contributing to a broader discourse on digital content ecosystems.

Our study has effectively addressed the research questions, enriching the discourse on network science within the realm of digital entertainment. Our findings highlight the need for continued exploration into the evolving narrative of network structures, where real-world constraints and cultural idiosyncrasies weave a complex web of relationships that define and drive creative industries forward.

[25] https://en.wikipedia.org/wiki/List_of_streaming_media_services

Methods

In order to analyze the network, a variety of software tools and libraries were used. Microsoft Excel was initially used to better understand the dataset and quickly find any imperfections in it. Python[26] was used due to its great support with network graphs and ease of use. Python provides an extensive set of libraries that were used in this project. NumPy[27] was used for efficient numerical computations and data manipulation. NetworkX[28] provided great handling and analysis of network data structures. In addition to that, Matplotlib[29] played an important role in generating insightful visualizations showing the results of our analysis. Later on, Gephi[30], an open source visualization tool, was used to visualize the network. With the help of these tools and libraries, we were able to conduct comprehensive statistical analyzes, perform complex network calculations and effectively visualize our findings.

In order to analyze the community structure of the network, the Louvain Algorithm was used. This algorithm is well suited for networks with a large number of nodes, it optimizes the modularity[31] value in each of its iterations, and results in a group of nodes with dense internal connections. The algorithm reveals the underlying community structures. It is implemented in NetworkX, providing a robust framework for executing the algorithm and interpreting the results. Community detection with the help of the Louvain algorithm resulted in slightly different communities each time, thus in order to get an accurate measurement of the modularity, the community detection algorithm was executed multiple times and the values were recorded. Overall, the application of the Louvain Algorithm proved instrumental in uncovering the intricate community structure within the network, contributing to a deeper understanding of its underlying structure.

[26] https://www.python.org/
[27] https://numpy.org/
[28] https://networkx.org/
[29] https://matplotlib.org/
[30] https://gephi.org/
[31] https://en.wikipedia.org/wiki/Modularity_(networks)

Code

To provide a centralized repository for storing and managing the project's source code, Github[32] was used. Github provided seamless collaboration among team members and provided easy tracking of changes in the code base while helping better organize the code.

The code repository can be found on the following link:
https://github.com/simratbenipal/572-Project-Netflix

[32] https://github.com/