Network Science Project - Building Real World Applications

Hemalata Nayak

Supervisor: Dr. Alexander Nwala

Department of Applied Science
College of William and Mary

This report is submitted for the final project in "Network Science" course
PhD First Semester

April 18, 2014

Abstract

Table of contents

List of figures

List of tables

Chapter 1

Milestone - 1

1.1 Project Task

Our project aims to analyze the collaborative networks of renowned film directors, focusing on key collaborators and their roles. We will investigate patterns of collaboration, such as directors consistently working with the same individuals, and explore how these collaborations contribute to the creative process and the overall quality of the films. Additionally, we will examine how some directors prioritize diversity and inclusion by hiring collaborators from historically marginalized groups, such as African Americans and women, to broaden the perspectives and voices in the U.S. film industry. Through network analysis, we aim to gain insights into the dynamics of these collaborations and their impact on the industry's creative labor pool and representation.

It involves four steps as following:

- Data Extraction/cleaning
- Network Generation
- Network Visualization
- Analysis

1.2 Data Extraction

The project starts by extracting relevant information (all details of directors and their movies) from a CSV file (‘100 film directors.csv‘) containing data on directors and their movies. The snapshot of the CSV file is shown in Figure 1.1. The following list of directors includes the last name, first name, sex, ethnicity/race (A=Asian, Asian American (incl India), B=Black, I=Indigenous (Native American, Maori), L=Latin American, W=White), labels (H=top 25 highest grossing directors (excluding animation directors) and Q=identifies as LGBTQ), and IMDb URIs of 101 directors.

1	LastName	FirstName	Sex	Ethnicity_Race	Labels	IMDb_URI
2	Abrams	J.J.	M	W	H	https://www.imdb.com/name/nm0009190/
3	Allen	Woody	M	W		https://www.imdb.com/name/nm0000095/
4	Anderson	Paul Thomas	M	W		https://www.imdb.com/name/nm0000759/
5	Anderson	Wes	M	W		https://www.imdb.com/name/nm0027572/
6	Araki	Gregg	M	A	Q	https://www.imdb.com/name/nm0000777/

Fig. 1.1 The csv file which contains director and movie information

I have extracted all the IMDb URIs of each director from the csv file and stored it in a directory. The data are organized as shown in the Figure 1.2. The directory structure follows the format where "film-directors" is the main directory, and each sub-directory is named according to the unique director ID for each director. In each sub-directory, we will find a credit.json file containing all the details of the movies as shown in Figure 1.3 the director has worked on.

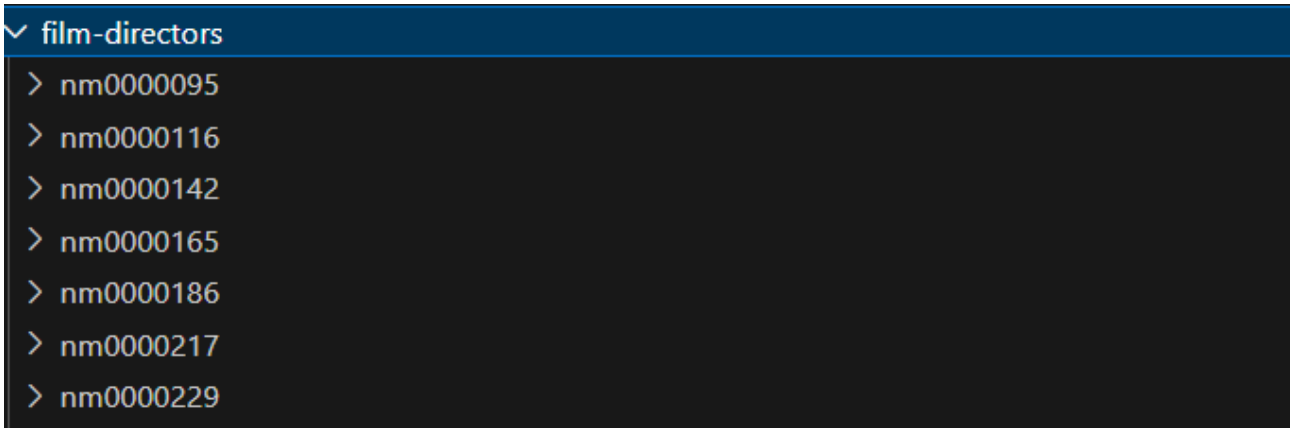

Fig. 1.2 The directory structure

I then scraped the movie URIs in the credit.json file for all directors, extracting the crew information for each movie and storing it in the respective directory.

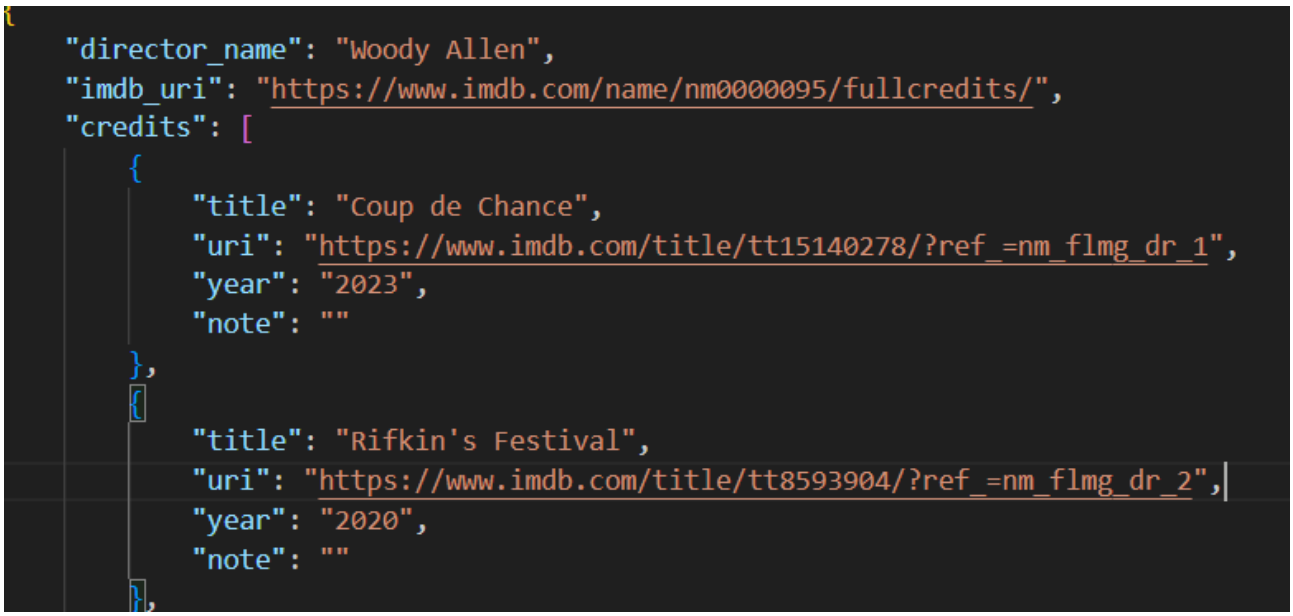

Fig. 1.3 The structure of the credit.json file

After extracting the crew information, the next step involved normalizing all variants of cast and writing credit to their root roles. This normalization process ensured consistency in how roles were represented across different movies. The normalized information was then stored in a directory named "Normalized_data" within each director's directory. The differences in file content before and after normalization are illustrated in Figure 1.4.

So far data extraction and cleaning is completed. The next steps involve analyzing the extracted data to create the collaborative network and visualize the network to find the relationships among directors and their collaborators.

(a) Crew details for a movie before normalization

(b) Crew details for a movie after normalization

Chapter 2

Milestone 2

In the previous section I had extracted the data and normalized it. But there were issues with downloading some files, hence they were empty. So I modified that code to handle such issues and re downloaded data and then I normalized roles and sub roles only for the movies which are longer than 70 mins and have saved it in the directory named "Normalized data".

2.1 Distribution of entities in the director-crew dataset:

Then I addressed some of the research questions for sanity check. Let's see below.
Q1. Number of featured movies?
Ans: The total no. of featured movies is found to be 1383 (excluding co-directors).
Q2. Number of directors, number movies per director, and average number of movies per director?
Ans: There are 101 directors. Average number of movies per director is: 13.693. The no. of movies per director is given in the table 2.1.
Q3.Number of crews?
Ans: The total number of crews is 7577.
Q4.Number of roles and frequency of each role?
Ans: There are total 17 roles.Out of which 8 are main roles and others are subroles. The frequency of each role is given the table 2.2.

The next step is generating the network and visualize it. And then I will address the research question.

Table 2.1 Number of movies per director ID

Director ID	Number of Movies	Director ID	Number of Movies	Director ID	Number of Movies
nm0000095	53	nm0000116	12	nm0000142	41
nm0000165	36	nm0000186	21	nm0000217	40
nm0000229	38	nm0000231	25	nm0000233	14
nm0000318	20	nm0000338	28	nm0000343	23
nm0000361	32	nm0000386	18	nm0000399	20
nm0000464	17	nm0000487	14	nm0000490	41
nm0000500	24	nm0000517	10	nm0000520	13
nm0000600	17	nm0000631	30	nm0000709	25
nm0000759	9	nm0000777	11	nm0000876	13
nm0000881	15	nm0000941	10	nm0001005	11
nm0001054	21	nm0001060	16	nm0001068	9
nm0001081	11	nm0001331	11	nm0001392	15
nm0001631	4	nm0001741	11	nm0001752	37
nm0001814	24	nm0002132	9	nm0004716	8
nm0005069	12	nm0009190	6	nm0027572	12
nm0036349	6	nm0037708	11	nm0122344	12
nm0138927	9	nm0160840	12	nm0169806	8
nm0190859	9	nm0200005	5	nm0269463	9
nm0281945	8	nm0298807	21	nm0327944	9
nm0336620	14	nm0336695	8	nm0362566	11
nm0366004	7	nm0392237	7	nm0420941	4
nm0426059	6	nm0476201	6	nm0501435	6
nm0510912	11	nm0570912	11	nm0583600	6
nm0590122	7	nm0619762	16	nm0634240	11
nm0668247	12	nm0697656	6	nm0716980	8
nm0751102	10	nm0751577	9	nm0796117	15
nm0853380	8	nm0868219	12	nm0893659	10
nm0898288	13	nm0905152	7	nm0905154	8
nm0911061	23	nm0946734	12	nm1119645	11
nm1148550	8	nm1218281	6	nm1347153	43
nm1443502	3	nm1490123	12	nm1503575	3
nm1560977	5	nm1716636	4	nm1802161	5
nm1883257	11	nm1950086	4	nm2011696	3
nm2125482	4	nm3363032	4		

Table 2.2 Number of crews per role/subrole

Role/Subrole	**Number of Crews**
Produced by - producer	3046
Produced by - producer produced by	592
Produced by - producer produced by pga	211
Writing Credits	3814
Sound Department - rerecording mixer	2106
Sound Department - sound designer	535
Sound Department - supervising sound editor	848
Directed by	1415
Cinematography by	1932
Casting By	1912
Music by	1521
Production Design by	1359
Costume Design by	1266
Makeup Department - hair department head	384
Makeup Department - makeup department head	461
Special Effects by - special effects supervisor	624
Special Effects by - visual effects supervisor	1

Chapter 3

Final Report

3.1 Network Visualization

Up to this point, data has been collected for each director, and roles have been normalized. A preliminary check has been conducted to analyze the number of crews, movies, and the frequency of roles. The next step is network visualization before addressing the research questions.

A director-crew network is constructed, with directors and crew members as nodes and edges connecting them if they have collaborated on a movie. Since this network is large and represents real-world connections, it is built using NetworkX and visualized using Gephi. The director-crew network is illustrated in Figure 3.1 and figure 3.3 using different layout.

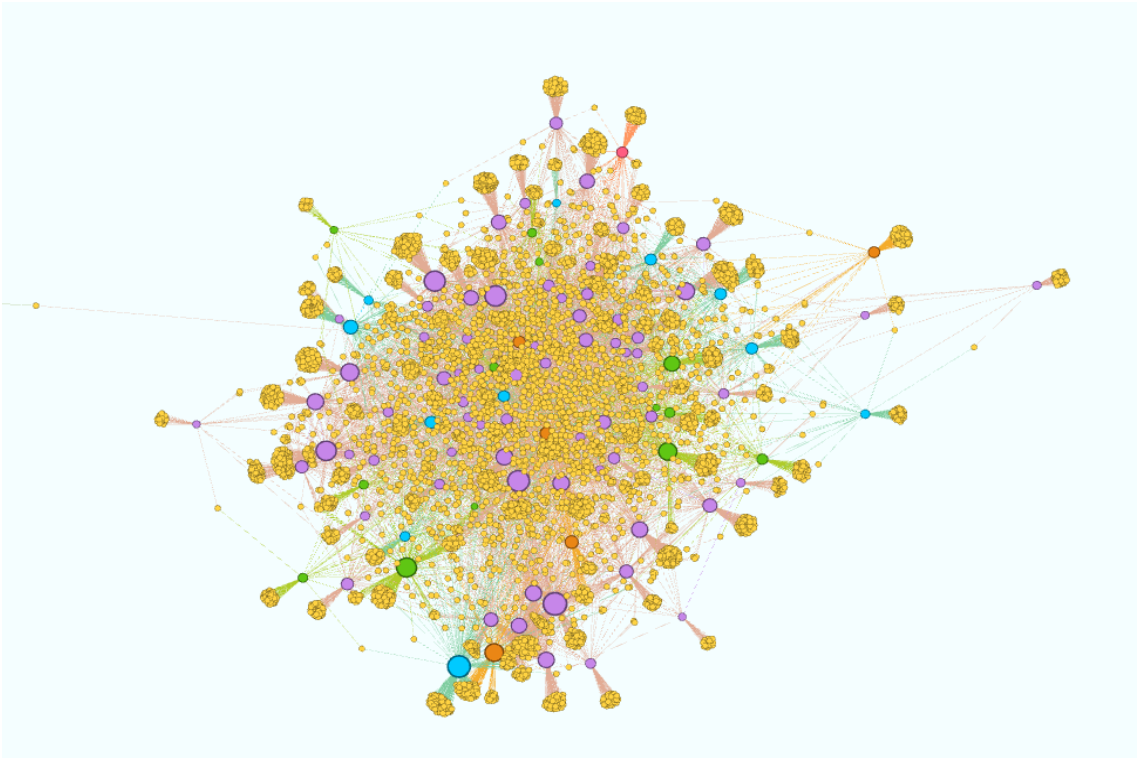

Fig. 3.1 Director-Crew Network visualized using Gephi with force-atlas2 layout

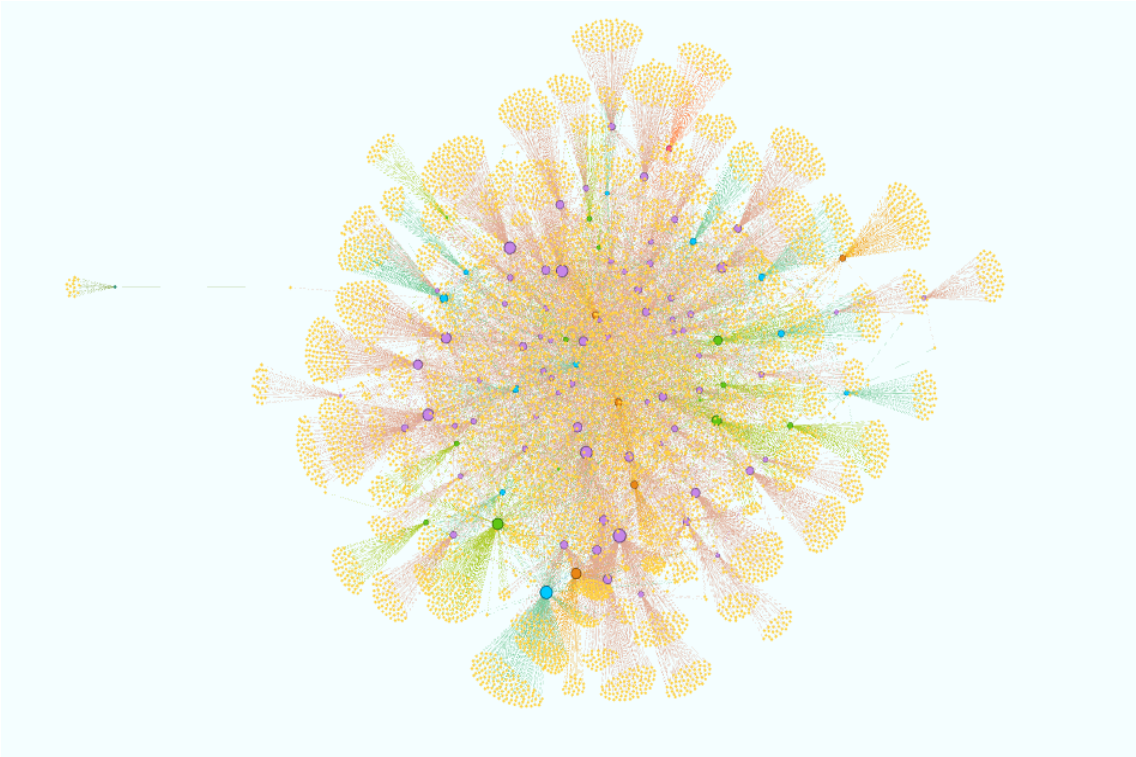

Fig. 3.2 Director-Crew Network visualized using Gephi with a different layout

The director ids are coloured based on their ethnicity race. And all crews have same color. Figure 3.3 shows what each color represents where A=Asian, Asian American (incl India), B=Black, I=Indigenous (Native American, Maori), L=Latin American, W=White).

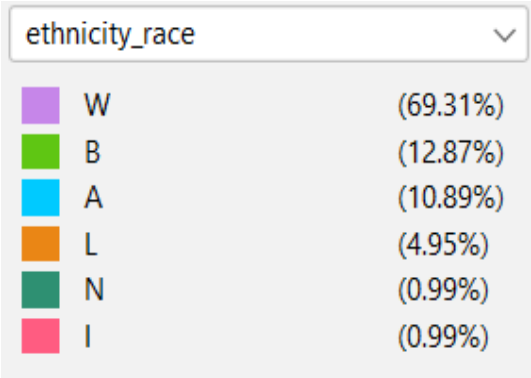

Fig. 3.3 Different colors for different directors according to their ethnicity race

As it can be seen in figure 3.3, The number of white directors is more and number of Indigeneous directors is the least.

3.2 Characterization of Director-Crew Network

- **What are the properties (degree distribution, average shortest path length, triangles aka clustering coefficient, density/sparcity)?**

Degree Distribution: In the study of graphs and networks, the degree of a node in a network is the number of connections it has to other nodes and the degree distribution is the probability distribution of these degrees over the whole network. Degree-distribution of the Director - Crew network is shown in figure 3.4. The plot of degree distribution indicates that the majority of crews (95%) have relatively low collaboration connections (degree) with other nodes (directors or crews), typically between 10 to 20 collaborations. However, there is a smaller subset (5%) of crews that have significantly higher collaboration connections, ranging from 50 to 250.

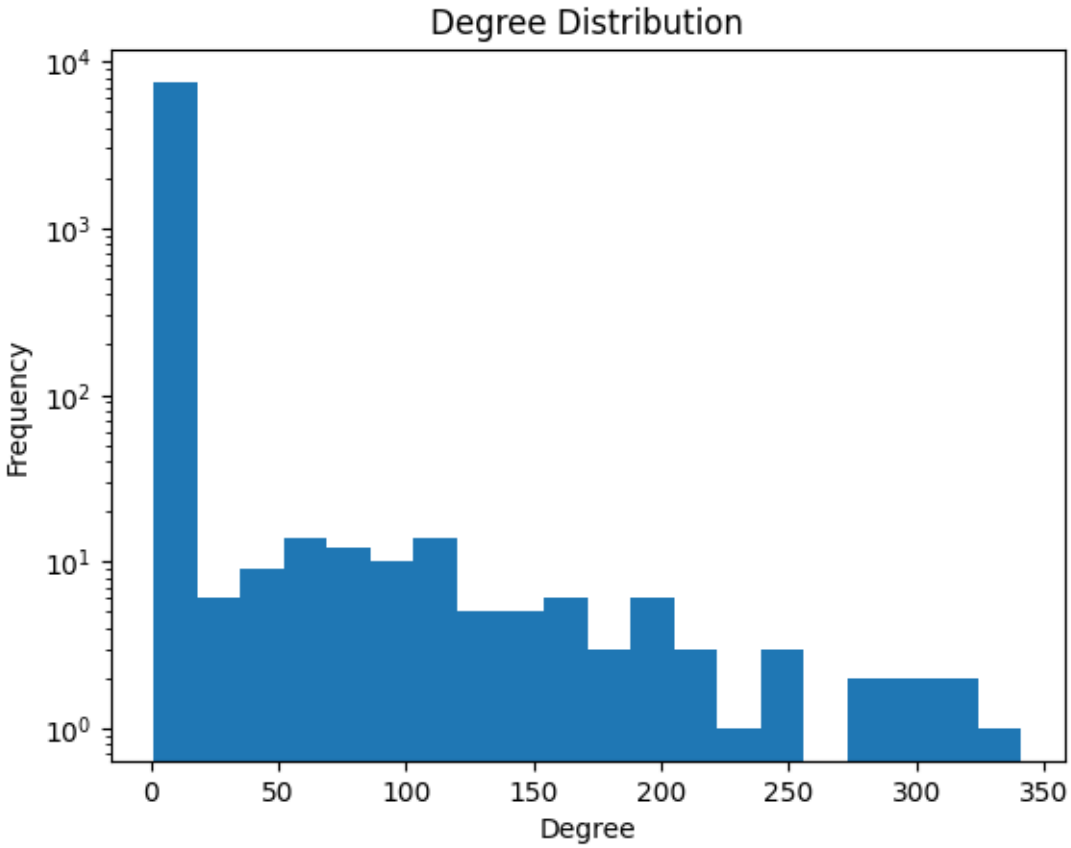

Fig. 3.4 Degree distribution of the director - crew network in log-scale

Average Shortest Path Length (APL): Average path length, or average shortest path length is defined as the average number of steps along the shortest paths for all possible pairs of network nodes. Consider an unweighted directed graph G with the set of vertices V. Let $d(v_1, v_2)$, where $v_1, v_2 \in V$ denote the shortest distance between v_1 and v_2. Assume that $d(v_1, v_2) = 0$ if v_2 cannot be reached from v_1. Then, the average path length l_G is:

$$l_G = \frac{1}{n \cdot (n-1)} \cdot \sum_{i \neq j} d(v_i, v_j),$$

where n is the number of vertices in G.

The average shortest path length was found to be 4.056 which indicates that, on average, it takes approximately 4 steps (or collaborations) to connect any two nodes (director or crew) in the network which represents Small World property.

Clustering Coefficient: A clustering coefficient is a measure of the degree to which nodes in a graph tend to cluster together or in another words how many triangles can be formed in the network. In case of director - crew network the clustering coefficient was found to be 0.052. The low clustering coefficient suggests that the network is less clustered, meaning nodes are less likely to form tightly interconnected groups or triangles. Here we have low clustering coefficient because of the way we have created the network. I am connecting edges if the director and crew have collaborated together. But if along with this I also include edges when two crews have collaborated together, in that case clustering coefficient will be high which is the property of a real-world network.

Density : The density of a network is determined by its ratio of links to nodes. The higher the ratio, the denser the network. In my network, the density was found to be 0.00043. The low density network indicates that the network is less sparse.

As the network characterization is done, now I am going to address the research questions.

3.3 Research Questions

Q1. Investigate and quantify how directors have influenced the careers of crew members.

To answer this, I explored the influence of directors on the careers of crew members by assessing changes in the number of roles crew members secured before and after initiating collaborations with specific directors. To achieve this, I constructed an interaction matrix of order M N to represent the collaborations between directors and crew members, with dimensions corresponding to the number of directors (M) and the number of crews (N). Then I computed the total collaborations for each crew member and identified the number of roles held by each crew member both before and after their initial collaborations with directors. The impact estimate was calculated as,

$$\text{Impact Estimate} = \text{Number of Roles After Collaboration} - \text{Number of Roles Before Collaboration}$$

Then the average impact estimate for each director was determined , providing a metric to gauge the overall influence of directors within the industry. Higher impact estimate indicates a more significant

impact of the director on the crew member's career. The plot in figure 3.5 shows the top 10 directors with highest average impact estimate. The name of these directors are provided in table 3.1. Joel Coen has the highest impact on carreer of crews with average impact estimate of 5.4.

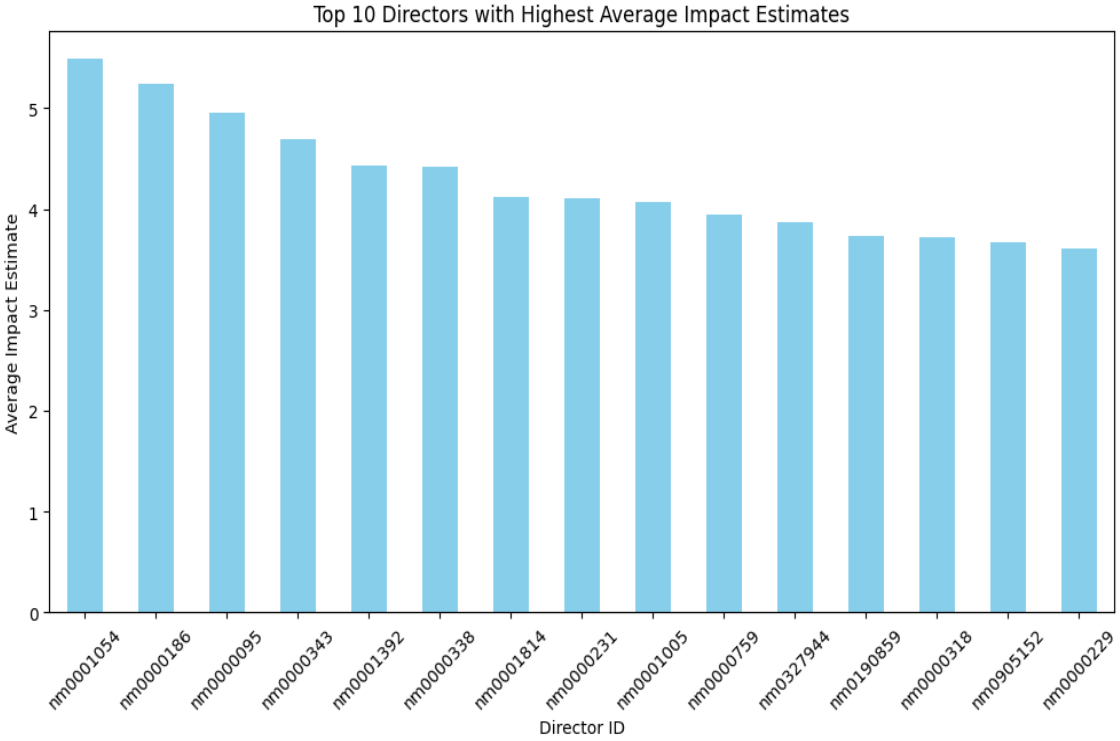

Fig. 3.5 Top 15 directors with highest average impact estimate where M stands for male directors and F stands for female directors.

Director ID	**Name**
nm0001054	Joel Coen (M)
nm0000186	David Lynch (M)
nm0000095	Woody Allen (M)
nm0000343	David Cronenberg (M)
nm0001392	Peter Jackson (M)
nm0000338	Francis Ford Coppola (M)
nm0001814	Gus Van Sant (M)
nm0000231	Oliver Stone (M)
nm0001005	Jane Campion (F)
nm0000759	Paul Thomas Anderson (M)
nm0327944	Alejandro G Inarittu (M)
nm0190859	Alfonso Cuaron (M)
nm0000318	Tim Burton (M)
nm0905152	Lilly Wachowski (F)
nm0000229	Steven Spielberg (M)

Table 3.1 List of top 15 Directors who has highest impact on career of crew and their IDs

But our dataset contains a wide range of crews. Some crews have started their career as early as 1978 and some have started their career just recently like in 2019. So considering that into account I decided to change the way I calculate Impact estimate. The new formula is given as follows:

$$\text{Impact Estimate} = \frac{\text{Number of Roles After Collaboration} - \text{Number of Roles Before Collaboration}}{\text{Total Number of Years the crew has worked}}$$

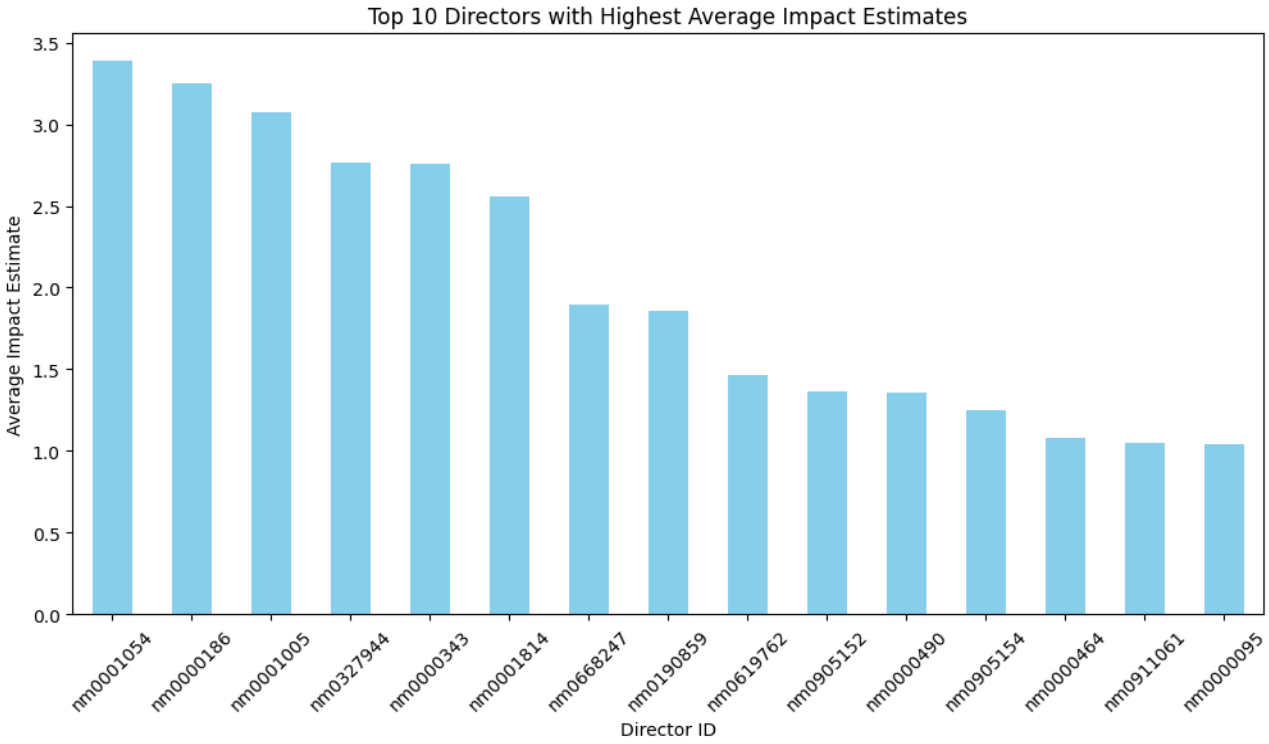

Fig. 3.6 Top 15 directors with highest average impact estimate

The plot in Figure 3.6 shows the top 15 directors with the highest average impact estimate based on the new approach. The names of these directors are given in Table 3.2.

If we look at both Tables 3.1 and 3.2, Joel Coen and David Lynch are the top directors who have the highest impact on the crew. However, there are some changes in the rankings of other directors. Most of the directors in the top 15 in Table 3.1 are the same as in Table 3.2, but the rank is slightly different.

It is concluded that Joel Coen has the highest impact on career of crews.

Director ID	Name
nm0001054	Joel Coen (M)
nm0000186	David Lynch (M)
nm0001005	Jane Campion (F)
nm0327944	Alejandro G. Iñárritu (M)
nm0000343	David Cronenberg (M)
nm0001814	Gus Van Sant (M)
nm0668247	Alexander Payne (M)
nm0190859	Alfonso Cuarón (M)
nm0619762	Mira Nair (F)
nm0905152	Lilly Wachowski (F)
nm0000490	Spike Lee (M)
nm0905154	Lana Wachowski (F)
nm0000464	Jim Jarmusch (M)
nm0911061	Wayne Wang
nm0000095	Woody Allen (M)

Table 3.2 Top 15 directors with the highest average impact estimate (considering the number of years the crew has worked)

Q2. Measure how roles of crew members fluctuate.

To measure the fluctuation in roles among crew members, I first calculated the number of different roles held by each crew member over time. This provided a detailed view of the variety of roles each individual took on throughout their career. Next, I calculated the frequency distribution of these unique roles to understand how commonly crew members shifted between different roles. To gain a broader perspective, I then calculated the total number of crew members, represented as 'total crew', by counting the unique entries in the role stability dataset. Finally, I determined the percentage of total crew members for each unique role count. The formula is as follows:

$$\text{Percentage of crews} = \left(\frac{\text{Role frequency}}{\text{Total crew}}\right) \times 100$$

This comprehensive approach allowed for a thorough analysis of role fluctuations and provided insights into the stability and versatility of crew member's careers.

Figure 3.7 shows the percentage of crews with how many roles they have played in their entire career. The *Normalized role* column shows different type of roles the crew has played in their entire career and. As it can be seen from the figure nearly 96% crews have worked only in one unique role throughout their entire career. Very few crews have more fluctuation in their role. **David Lynch** is the only crew who has worked in **6** unique roles throughout his career.

The visual depiction in Figure 3.8 reinforces this observation, demonstrating minimal role variation across the dataset. Thus, it can be inferred that role consistency prevails among the sampled crews.

Normalized Role	Number of Crews	Percentage of Crews (%)
1	7263	95.855880
2	218	2.877128
3	70	0.923848
4	20	0.263957
5	5	0.065989
6	1	0.013198

Fig. 3.7 Percentage of crews with no. of roles

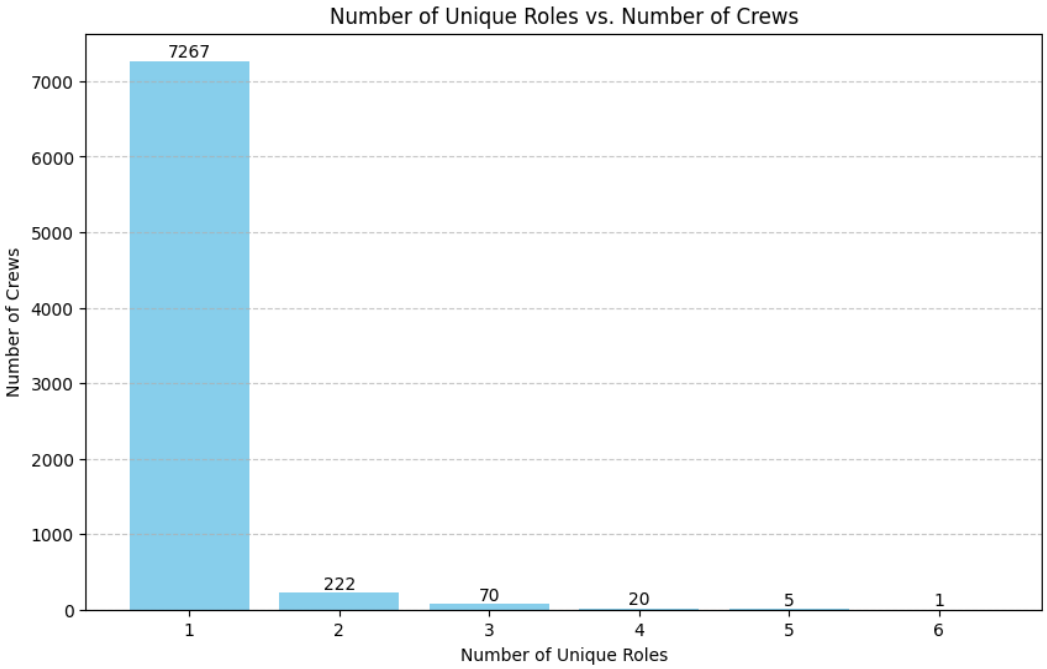

Fig. 3.8 Number of unique roles vs number of crews

3.4 Conclusion

In this study, we analyzed the Director-Crew network and confirmed that it exhibits characteristics of a real-world and small-world network. Upon addressing the research question, it was observed that while not all, certainly some directors significantly influence the careers of their crew members, as anticipated. Furthermore, the findings indicate that the majority of crew members typically engage in only one unique role throughout their careers. However, very few crews, such as David Lynch, have diversified their career paths by undertaking various unique roles, distinguishing them from their peers.