Group Coursework Submission Form

Specialist Masters Programme

Please list all names of group members: (Surname, first name) 1. Sabbagh Dit Hawasli, Tiana 2. Abu Dayyeh, Rand 3. Sinkov, Evgeny	4.Liu, Xiaotong **GROUP NUMBER:** **7**
MSc in: Business Analytics	
Module Code: SMM638	
Module Title: Network Analytics	
Lecturer: Dr Simone Santoni	**Submission Date:** **16 November 2021**

Declaration:

By submitting this work, we declare that this work is entirely our own except those parts duly identified and referenced in my submission. It complies with any specified word limits and the requirements and regulations detailed in the coursework instructions and any other relevant programme and module documentation. In submitting this work we acknowledge that we have read and understood the regulations and code regarding academic misconduct, including that relating to plagiarism, as specified in the Programme Handbook. We also acknowledge that this work will be subject to a variety of checks for academic misconduct.

We acknowledge that work submitted late without a granted extension will be subject to penalties, as outlined in the Programme Handbook. Penalties will be applied for a maximum of five days lateness, after which a mark of zero will be awarded.

Marker's Comments (if not being marked on-line):

Deduction for Late Submission:

Final Mark: **%**

1. Introduction

This project sought to analyse a proportion of the actor network embedded in the American film industry. Data used in the network analysis was scraped from a list on IMDb titled '500 Top-Rated Features Since 2000'. The scope of this project includes only US movies released between 2000 and 2012, and the 10 main actors from each film. The Hollywood movie industry is characterized by short-term relationships based on collaboration. A set of actors co-starring in a movie will interact for the duration of the film's production process. Moreover, the relationships among actors across movies are the product of an enduring network whereby mutual trust and reputations have been established over the years (Cattani and Ferriani, 2008).

2. Network characteristics

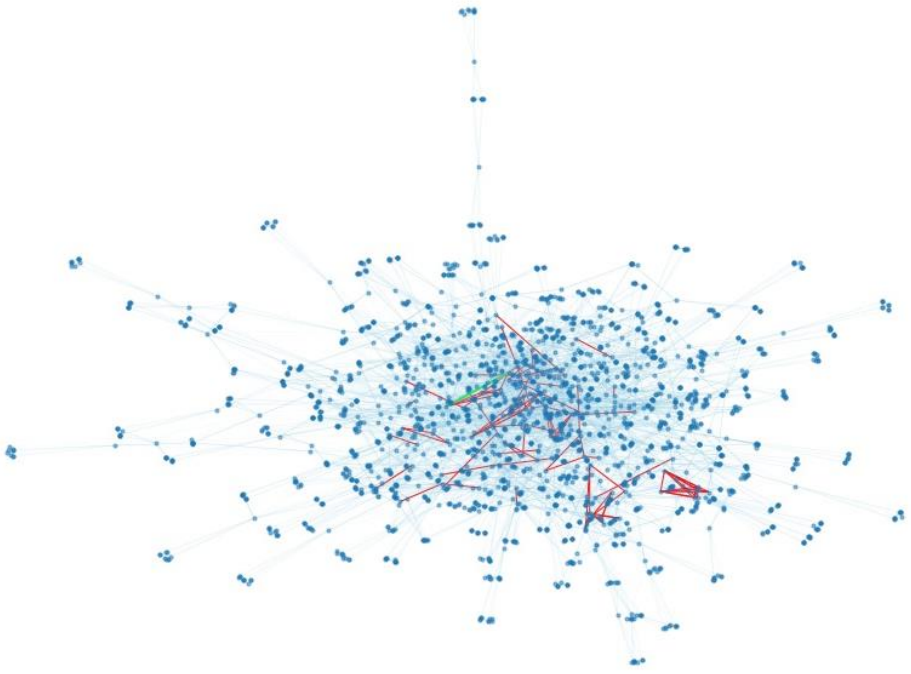
In the graph representing the co-stardom network, each movie actor is illustrated as a point (node) and each collaborative relation between a pair of actors is illustrated as a line (edge). The network is characterized as one-mode, weighted, and undirected. One-mode because only the set of movie actors is considered. Weighted as actors who have collaborated in more movies together, will have stronger connections. Undirected because of the collaborative nature of the relationships.

3. Representing the network as a graph

When the network was initially drawn out with all 2708 actors, it appeared to be disconnected. Therefore, for the analysis that followed, we derived a sub-graph 'Gcc' which comprised solely the connected components i.e., 2503 nodes. The colour of the edge corresponds to its weight (blue: 1, red: 2-3, green: >3).

Upon first inspection, the graph exhibits many clusters of nodes which indicates a small-world network system. It may be inferred, each cluster in the graph represents a community of nodes, which suggests the network is divided into modules. This outcome is perhaps the result of specifying the number of actors per movie, which automatically groups actors together (as a community) based on the movie they co-starred in.

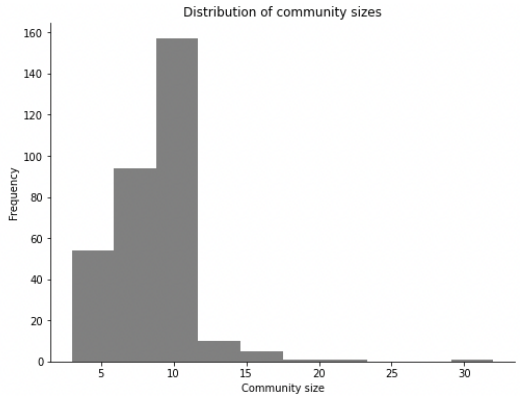
Figure 1: Gcc graph for connected components of the movie network



4. Community Detection

Due to issues with applying the Girvan-Newman algorithm conventionally used for community-detection, we resorted to the Kojaku-Masuda algorithm. The number of communities detected by Kojaku-Masuda algorithm for the whole network is 298 with group sizes ranging from 3 to 33. As such, the Hollywood movie actor network has high modularity which is denoted by the dense connections of nodes contained within communities.

Figure 2: Community detection



5. Degree distribution of the network

The degree distribution for Gcc reveals some unconventional patterns. Since we have specified that there be 10 actors per movie, the most frequent degree is 9 (10 actors per movie minus ego node), and most degree values are 9n for actors starring in multiple movies.

Some hubs are evident in the graph as there are few nodes with degrees above 9, signifying they are the most connected in the network. In the actual network, these hubs correspond to actors that starred in the greatest number of movies from the IMDB list.

Figure 3: Degree distribution scatterplots

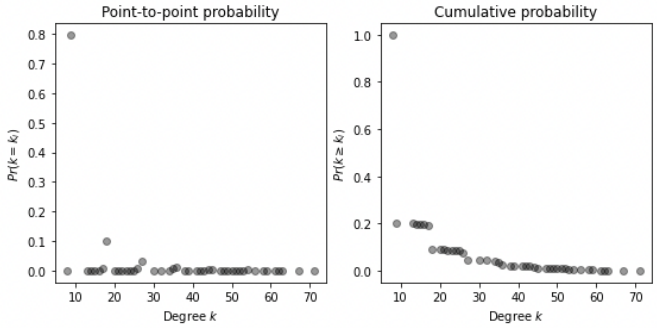
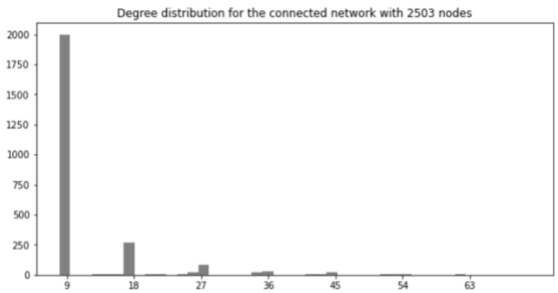


Figure 4: Degree distribution bar chart

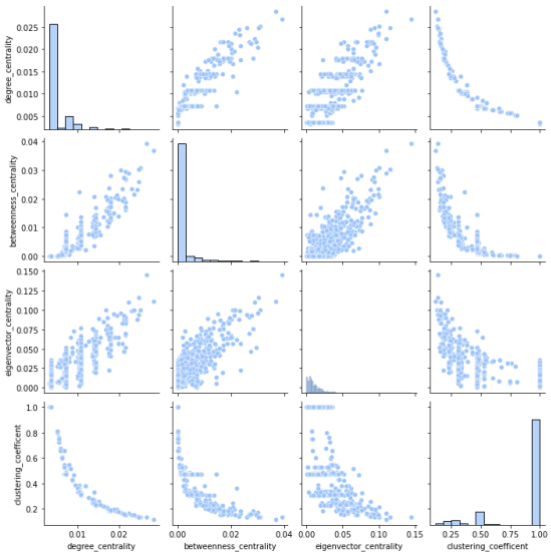


6. Network Attributes

Network 'diameter' which indicates the actual size of the network, was found to be 9 meaning that the longest shortest path in this network constitutes 9 consecutive nodes. The average shortest path length was found to be 4, which signifies a difference of 5 compared to the network diameter. Moreover, the average degree of connectivity in the graph was found to be ≈7.

A correlation matrix for the node centrality measures was generated, which revealed a great deal of variance among the nodes. This variance points to the activity of homophily mechanisms in the network.

Figure 5: Correlation matrix for node measures



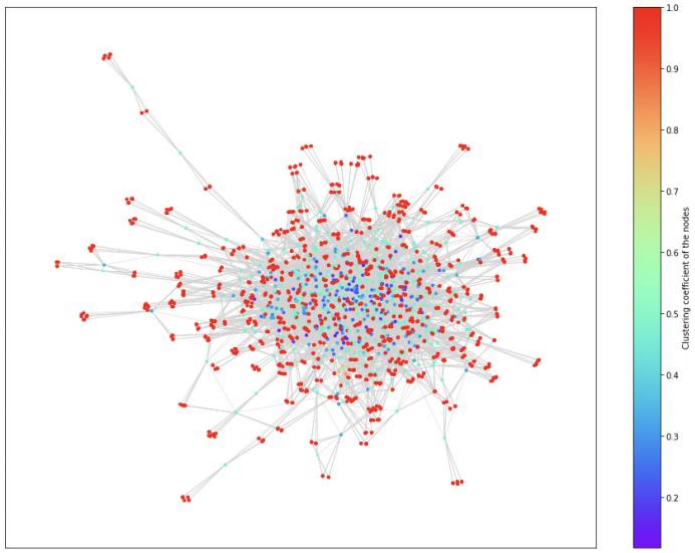
An important inference is that the network demonstrates small-world properties as it fits the characteristics of having a high mean clustering coefficient (0.88) and a small value for average shortest path (4). It may also be noted that most nodes have a clustering coefficient equal to 1, which indicates the network comprises many cliques. Our analysis revealed that 1997 nodes are in cliques in this network which signify the dense clusters in the network graph.

Figure 6: Descriptive statistics for centrality measures and clustering

	degree_centrality	betweenness_centrality	eigenvector_centrality	clustering_coefficent
count	2503.000000	2503.000000	2503.000000	2503.000000
mean	0.004896	0.001216	0.013140	0.877181
std	0.003212	0.003720	0.015065	0.250525
min	0.003197	0.000000	0.000039	0.117505
25%	0.003597	0.000000	0.004321	1.000000
50%	0.003597	0.000000	0.008259	1.000000
75%	0.003597	0.000000	0.016794	1.000000
max	0.028377	0.039297	0.144938	1.000000

These cliques were identified by generating a graph which manipulated the colours of nodes based on their clustering coefficient scores, with the colour red indicating a clustering coefficient of 1. Accordingly, all the cliques can be spotted from the red clusters of nodes in the graph; and we find that cliques are overwhelmingly prevalent in the network.

Figure 7: Graph of Gcc with nodes distinguished by clustering coefficient



7. Centrality Measures

To determine the most important actors based on different premises, we used the following centrality measures:

a. Degree Centrality

The degree centrality of the node refers to the fraction of degrees this node has in comparison to the whole network, thus it is a measure of how many connections the node has. The highest degree centrality values were attributed to the following movie actors:

Actor / Actress	Degree Centrality
John C. Reilly	0.02837729816147082
Joseph Gordon-Levitt	0.026778577138289367
Laura Linney	0.025179856115107913
Jim Broadbent	0.02478017585931255
Philip Seymour Hoffman	0.02478017585931255

The actors/actresses mentioned in the above have the highest numbers of connections in the network. In the graph, each of these actors' representative nodes would signify a hub. We may infer that they have acted in the greatest number of the top US movies rated in IMDB.

b. Betweenness Centrality

Betweenness centrality conveys the fraction of times the node appears on the shortest paths in the network. Accordingly, a node (actor/actress) with a high betweenness centrality will be

situated centrally in the network and regarded as mediators. The below table shows the actors/actresses with the highest betweenness scores:

Actor / Actress	Betweenness Centrality
Joseph Gordon-Levitt	0.03929718505621907
John C. Reilly	0.03672232626912827
Laura Linney	0.03078987307250857,
Philip Seymour Hoffman	0.030387999735242803
Tommy Lee Jones	0.030081000858188104

It may be noted that the same highly connected actors are in a central position in the network. Therefore, there is a correlation between the betweenness and degree centrality measures in this network. The mean betweenness centrality reflects a relatively low value due to movies being selected based on their IMDB rating rather than common denominators (e.g. genre, common actors); thus the collaborations and shortest paths across the network have been artificially reduced.

c. Eigenvector Centrality

Eigenvector centrality depicts the strength of a node's position and the strength of its neighbors' position. If an actor holds a relatively middling position, but their neighbouring actors they have acted with are important in the network their score will be higher. The table below lists the actors/actresses with the highest Eigenvector centrality:

Actor / Actress	Betweenness Centrality
Joseph Gordon-Levitt	0.14493803385784115
Philip Seymour Hoffman	0.11571640006707758
John C. Reilly	0.11067849795135917
Christian Bale	0.10028216017700342
Leonardo DiCaprio	0.10004558896096256

The table introduces two new actors who are Christian Bale and Leonardo DiCaprio to the nodes with the highest centralities. This can be explained by the fact both mentioned actors have neighbours with relatively high degrees (connections) which was reflected in the eigenvector centrality value.

8. Network Outcomes

The network demonstrates a small-world network, as any two nodes (actors) can reach each other through a small sequence of nodes given that the network diameter is 9. This outcome may have been produced due to the characteristics we ascribed to the network by selecting the top 10 actors from the top-rated US movies. Accordingly, we end up with an actor set comprising the most eminent actors.

A possible outcome of this network is homophily based on popularity of the actor. The most famous actors will likely co-star with very famous actors as well. As mentioned in the introduction, movie actors that are well-established in the industry have built strong reputations over the years. The more movies an actor has starred in, the more they have built trust with other actors in the network, which leads to them being selected for more popular movie roles in the future and starring aside other big actors.

Finally, due to the short average path of the graph, an actor can very easily connect with another actor in the network. For example, we found that Tom Hardy only needs 2 steps to reach Lindsay Lohan. In a sense this occurrence clearly conveys the small world phenomenon, aka the '6 degrees of separation' theory, which we would inevitably find in a network of the most prominent Hollywood stars.

9. References

Cattani, G., & Ferriani, S. (2008) 'A Core/Periphery Perspective on Individual Creative Performance: Social Networks and Cinematic Achievements in the Hollywood Film Industry', *Organization Science*, *19*(6), pp. 824–844.