Available online at
https://jcrinn.com/
https://crinn.conferencehunter.com/

Journal of Computing Research and Innovation 9(2) 2024

Journal of Computing Research and Innovation

Exploring Film Industry Dynamics: A Network Science Approach to Internet Movie Database Analysis

Muhammad Izzat Farid Musaddin[1*]

[1]*ICT Management Centre, MARDI Headquarters, Serdang, Malaysia*

ARTICLE INFO

Article history:
Received 23 June 2024
Revised 18 August 2024
Accepted 27 August 2024
Online first
Published 1 September 2024

Keywords:
Network Analysis
Centrality Measure
IMDB
Network Science
Graph Theory
Igraph

DOI:
10.24191/jcrinn.v9i2.455

ABSTRACT

Throughout the history of the film industry, many people have been involved in roles like acting, directing, or even writing the storyline of a TV show or movie. A question arises: Who is the most influential person among all those people? The objective of this study is to provide an answer to this inquiry. Firstly, the Internet Online Movie Database (IMDb) was selected as the data source for this study due to its vast data volume. Furthermore, we employed network science methods to study the social networks of the film industry. To be precise, we performed network analysis where we gained valuable information from properties that relate to influence, which is called centrality measures. Three commonly used centrality measures were chosen to provide different perspectives based on the IMDB dataset, namely betweenness, closeness, and degree centrality. Moreover, we want to identify individuals with the highest scores for all centrality measures tested. In addition, the KNIME Analytics Platform tool was used to preprocess the IMDB data by implementing data integration and transformation. Subsequently, the Igraph package available in Python was utilised to obtain the centrality measure scores. The results from these methods pointed to specific nodes, which were then compared with the rating table of the IMDB dataset.

1. INTRODUCTION

When discussing films, one would wonder who the most influential person is. Focus cannot be directed only towards actors; other job roles like directors, producers, and writers can also have immense influence in the film industry, depending on how many people they have worked with on different projects. Who is the most influential person in the film industry? Which job categories have the most influence on others? Among all the directors in IMDB, who is the most influential? The primary objective of this study is to attempt to explore these questions. To understand the influence of the IMDB dataset, we employed the Network Science approach, specifically performing network analysis to gain insights on the scores by the network-related properties called centrality measures.

[1*] Corresponding author. *E-mail address*: izzatf@mardi.gov.my
https://doi.org/10.24191/jcrinn.v9i2.455

There is a buzz in the world of technology, where the vast majority of people are talking about artificial intelligence and machine learning as ways to analyse and make predictions based on the data they have. However, when greater emphasis is placed on exploring the social aspect of data, this then involves an in-depth exploration of network analysis. The reason is that by using Network Science methods, we can identify the relationships and interactions between the nodes that exist within a network.

2. LITERATURE REVIEW

This field of research focuses on understanding natural and artificial networks through the development of theoretical and practical methods and techniques (Sadaf et al., 2024) Specifically, Network Science aids in identifying connections between nodes, and comprehending these connections is crucial for determining the type of graph and the value it represents. However, Network Science did not emerge independently. When exploring its history, Network Science is recognised as a field of study that originated from Graph Theory, a branch of mathematics that focuses on networks (Henning & van Vuuren, 2022). It was a significant moment for researchers as they succeeded in finding practical solutions to real-world problems that emerged during the late 1960s and 1970s when social networks were modeled using Graph Theory (Henning & van Vuuren, 2022). The primary objective of their studies was to gain insight into the complex relationships and behaviors of network nodes. In addition, both computer scientists and mathematicians regarded Graph Theory as the standard approach for managing graphs and their characteristics due to the helpfulness of available algorithms (Zhu et al., 2022). Subsequently, scientists extended their efforts to apply Graph Theory to other fields, including biological phenomena.

On the other hand, Machine Learning involves implementing procedures to make predictions based on data sources and preferred algorithms depending on the research objectives. What sets Network Science apart from Machine Learning is that the former gathers detailed information, such as network properties, through examination of the network rather than making predictions. An illustrative example of this concept is our ability to identify the most influential person within a social network. By doing so, we can gain insights into the reasons for their influence and how they manage to establish numerous connections with others (Abbasi & Fazl-Ersi, 2022). Furthermore, through analysis, researchers can forecast outcomes based on the accessible datasets, thus facilitating decision-making.

A network is essentially a collection of vertices connected by edges representing the relationships between them within the same plane, depending on the type of network being observed. As an illustration, in a social network, individuals are treated as vertices, and connections between them are represented by edges. When performing an analysis on a selected network, attention is directed towards network parameters known as centrality measures, which prove helpful when we want to understand the network in depth. Further studies have revealed the presence of parameters within Network Science known as centrality measures (Hernández & Sánchez, 2020). Centrality measures are properties studied during network analysis, identifying the importance of nodes or vertices based on their ranking within a given network. Network analysis has been conducted since the 1950s. During that period, communication networks were tested to assess their performance against corporate standards, including error rate and time to problem resolution (Centola, 2022). Centrality measures are typically based on fundamental network properties and are used to identify prominent or key nodes that play pivotal roles in maintaining connectivity of influencing the flow of information in a network (Bloch et al., 2023). They are commonly used in various fields, including social network analysis, transportation planning, and the analysis of biological networks. One of the most commonly used measures is degree centrality, which considers the number of adjacent node neighbours.

In addition, the centrality measure enables researchers to comprehend the connections between nodes within a network, gauging the influence on other connected nodes. Various types of centrality measures exist to provide a deeper understanding of a network. For example, some measures assess the importance

https://doi.org/10.24191/jcrinn.v9i2.455

of vertices based on the number of links or edges that they have with other nodes, known as degree centrality.

Betweenness centrality measures how often a node lies on the shortest path between other nodes in the network, indicating which nodes act as bridges to connect different groups. Essentially, the betweenness centrality calculation yields a score based on the percentage of shortest routes that traverse through a particular node (Bacsi et al., 2023). Nodes with high betweenness centrality act as bridges or intermediaries between different parts of the network, playing a crucial role in maintaining connectivity. In a transportation network, for instance, a node with high betweenness centrality might be a critical intersection. On the other hand, closeness centrality calculates the shortest path between a node and all other nodes in the network, assigning them a score based on the sum of these shortest paths. In simpler terms, a network's closeness centrality reveals the proximity of a node to every other node (Naderi & Shojaei, 2023). Previous research on closeness centrality also identifies it as farness (Mann et al., 2022). Nodes with high closeness centrality are central because they can reach other nodes quickly. In a social network, a person with high closeness centrality can quickly reach others through short paths. Furthermore, degree centrality emphasises the count of out-degree for a node in strongly connected components and the count of degrees for an undirected network (Qazi et al., 2021). In other words, degree centrality is a simple measure that counts the number of edges connected to a node (degree) in a network. Nodes with a high degree of centrality are well-connected to many other nodes in the network. In social networks, for example, a person with a high degree of centrality has many connections.

The objectives of this study were to delve into how Network Science influences social networks at a deeper level and what insights can be gained from performing network analysis, particularly when implementing centrality measures. This is aimed at quantifying key nodes that play crucial roles in the structure and function of a network involving actors, directors, and writers.

For the purpose of this study, three different centrality measures, betweenness, closeness, and degree centralities, were chosen to understand and determine the influence a node has on the network of the Internet Movie Database (IMDB) dataset. They were selected for this study because they were among the most commonly used measures of centrality in network analysis. Incorporating the study of social networks, we explored the following list of research questions for the present study: 1. Who is the most influential person in the IMDB dataset? 2. What is the most influential job type in the IMDB dataset? and 3. Does a director have any influence on the success of a movie/ TV show?

3. RESEARCH MOTIVATION

This research aims to understand the Internet Movie Database (IMDB) from a social point of view. Due to its large volume, which keeps increasing to this day, there is a need to develop a feasible method to analyse and make efficient interpretations of the network. Specifically, we want to identify the influence of a social network. Thus, further investigation of one particular network science property called Centrality Measure during the network analysis would help achieve the goal. To be more precise, the measure looks into the nodes or individuals deemed necessary in influencing its surroundings. We can also comprehend the network behavior by conducting this research.

4. METHODOLOGY

4.1 The Internet Movie Database (IMDB)

In this study, Centrality Measures such as betweenness, closeness and degree centralities analysis were performed on the Internet Movie Database (IMDB). The IMDB has evolved into one of the most sought-after repositories throughout the world, where users search for information on movies. The database

https://doi.org/10.24191/jcrinn.v9i2.455

contains data, including movies, TV shows, actors, and even reviews for each show. This information helps users decide what show to watch next. IMDB was founded in 1990 by a software engineer named Col Needham, who was located in Bristol at Hewlett-Packard (Lewis, 2024). He organised all the movies he had watched into a discussion group related to films (Lewis, 2024). It subsequently expanded and became one of the early pioneers of the World Wide Web (www) (Lewis, 2024). The growth during this period was facilitated by collaborative strategies, allowing individuals with movie information to insert data after registering on the site. By 2008, IMDB had evolved into a more sophisticated platform through multiple enhancements and acquisitions of other film-related sites, such as Box Office Mojo, ensuring steady growth (Lewis, 2024).

As one of the most popular websites for storing data on movies and TV shows, IMDB has become a go-to source for users to seek information and a valuable resource for researchers to analyse data using various methods and models. This is possible because IMDB provides links to its database, allowing people to download the datasets. However, it is essential to note that this is permissible only in accordance with their terms of service. For instance, users are allowed to use the data solely for personal, and non-commercial purposes (IMDb, n.d.). Specifically, making any alterations towards the data are prohibited (IMDb, n.d.). By offering this functionality, IMDB retains control over usage while actively contributing to the broader data community.

4.2 Data Preprocessing: KNIME Analytics Platform

To implement the Network Science method, a preliminary data cleansing process was carried out, involving the removal of unwanted rows and standardisation of data format to better suit the network analysis. After analysing the IMDB datasets and their structure, data cleansing and transformation were necessary steps before conducting network analysis. To be more specific, when using algorithms for network analysis, it is essential to ensure that the data is structured in a way that aligns with the method. Because the data in the IMDB dataset is stored in separate tables, making it challenging to use them directly in network analysis algorithms, a transformation of the data is necessary. For example, information about actors is stored in a table named 'name.basics.tsv.gz' while movie data is stored in 'title.basics.tsv.gz'.

In an effort to address this issue, data integration is performed, which is a subset activity of a more comprehensive process called ETL. ETL, which stands for Extract, Transfer and Load, encompasses the process of retrieving and consolidating transactional data from various data sources, including Excel and text files (Aaltonen et al., 2021). The gathered data is then converted into databases tailored for reporting and analytics (Kinast et al., 2023). While the overall purpose of ETL aligns with the development of a data warehouse involving the handling of massive amounts of data, the primary focus of this study is on data extraction, integration, and transformation. There is no emphasis on developing a data warehouse, as the intention is solely tailored for the study. To facilitate this process, the KNIME Analytics Platform was chosen to perform the necessary tasks. KNIME, in short, the Konstanz Information Miner Analytics Platform, is a tool that enables the integration of new algorithms, data manipulation, and visualisations using nodes (Gladysz et al., 2023). KNIME is a graphical user interface (GUI)-based tool that utilises a node-based approach to data analysis. In addition to being free-to-use, KNIME offers a user-friendly experience, allowing users to simply drag and drop components, known as nodes, into the provided view. While KNIME minimises the need for coding, there are instances where code writing is still necessary. For example, when implementing string manipulation for specific data columns, all required functions must be written as code in the provided space; otherwise, the procedure will result in an error. The following is an example of a screenshot of what a KNIME workplace looks like.

https://doi.org/10.24191/jcrinn.v9i2.455

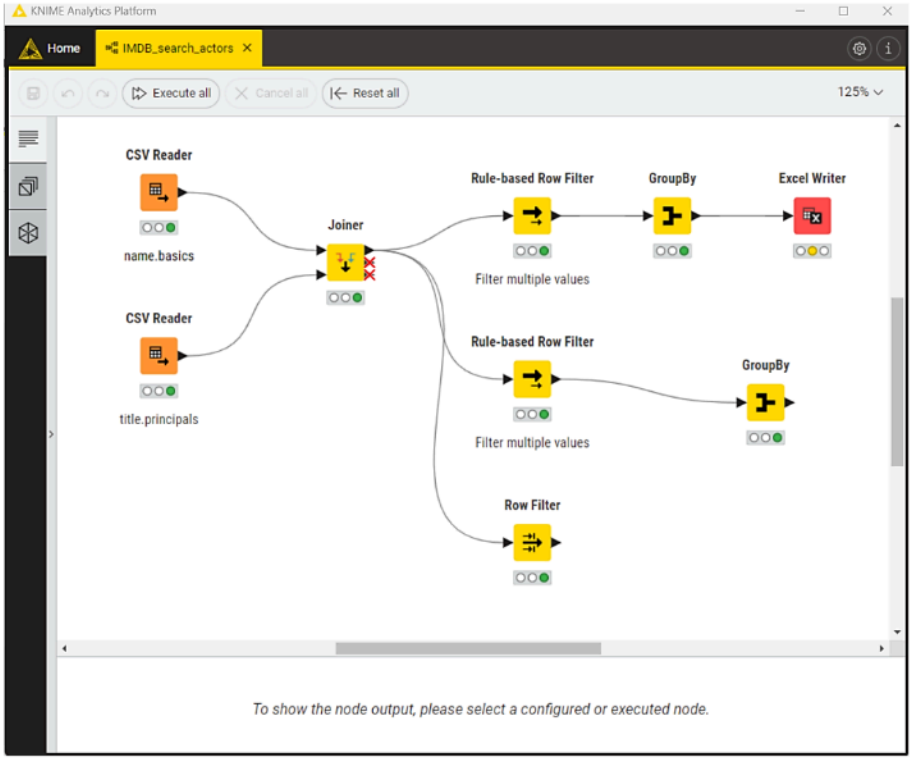
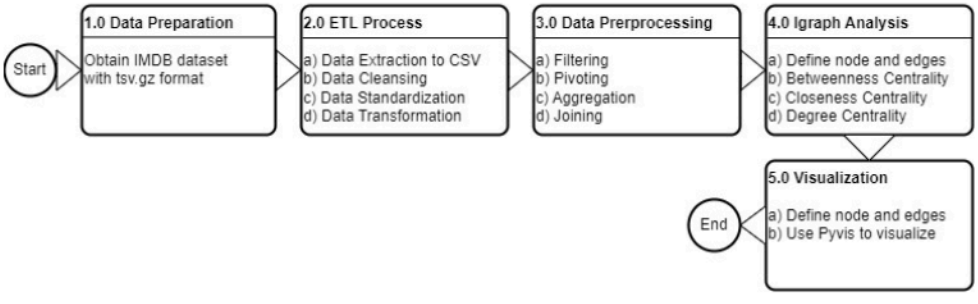


Fig 1. View of KNIME workspace

Upon opening the KNIME software, three main panels are displayed on the screen: the project panel, Node repository, workspace panel, and console panel. After selecting preferred nodes and establishing connections, the project can be executed, enabling the production of outputs such as writing results into a CSV file.

In support of the Extract, Transform, Load (ETL) process, KNIME provides various procedures, including data cleansing, standardisation, and transformation, which were implemented for the IMDB dataset selected earlier. In the present study, the use of KNIME enables the execution of all these activities, facilitating data manipulation. For example, preprocessing the input data can be accomplished through methods such as filtering, pivoting, aggregation, and joining. This is crucial to ensure that the data is prepared for analysis. The following Fig. 2 shows the complete process flow of this study.



https://doi.org/10.24191/jcrinn.v9i2.455

Fig. 2. Process flow conducted in this study

In this study, the KNIME software was utilised in both the ETL process in phase 2.0 and data preprocessing in phase 3.0, as outlined in Fig. 2. During this stage, data was transferred from the original IMDB data source to a newly created CSV file with a structure designed to align with the standard format for the subsequent data analysis phase. Specifically, this structure was tailored to accommodate analysis using Igraph, a Python library that supports Network Analysis. To elaborate, the dataset was crafted to suit Igraph analysis, considering elements such as nodes and edges derived from the IMDB dataset. In this context, nodes represent individuals involved, such as actors, directors, and writers, while edges denote the relationship between these nodes. For example, an actor node may be connected to a director node if they have collaborated on the same movie. The data transformation process was repeated for each research question, adjusting nodes and relations to match the specific algorithms involved, which are different from each other. After completing the transformation process, analysis was performed using Igraph algorithms. This can be observed in phase 4.0 of Fig. 2, where the selected centrality measures were employed when analysing the network.

4.3 Network Analysis: Igraph

Igraph is a freely accessible network analysis software package for anyone interested in graph analysis (Askar, 2021). Beyond its cost-free availability, Igraph empowers scientists to generate and manipulate their graphs, offering functionality beyond fundamental analysis (Askar, 2021). Operations like graph manipulation utilise graph theory algorithms to gain a deeper understanding of the network. An essential aspect of graph analysis involves uncovering communities within the graph, exemplifying one of the crucial activities performed when exploring a network. This purpose may involve creating subgraphs from the main graph to apply community detection algorithms like the Louvain Method. Additionally, Igraph includes built-in functions to facilitate centrality measure calculations. This functionality is evident in functions like "betweenness ()", "closeness ()", and "eigen_centrality ()". Furthermore, Igraph is versatile, supporting multiple programming languages like Python and R (Askar, 2021). Its capability extends to handling graphs with millions of vertices and edges, showcasing scalability (Mostafavi et al., 2022).

4.4 Network Visualization: Pyvis

After performing the analysis using Igraph, a different Python package, namely Pyvis, is used to visualize network-based and graph data, where users can interact with them (Perrone et al., 2020). To use Pyvis, we must create graphs using the code statements provided in the library by manually inserting the specific nodes and edges. For instance, we use the "add_node()" function to create nodes, whereas the "add_edges()" function creates edges that connect two nodes. However, when dealing with a large data volume like the IMDB dataset, using the original dataset file with a TSV file format and looping through each row to create the nodes and edges is preferable. Doing this prevents the hassle of calling the code statements stated earlier for every node and edge creation. Upon finishing the coding for the loop function, we can view how the nodes behave within the network. Turning off the "fixed()" function would enable the movements of the nodes where vigorous movement defines how connected a node is to other nodes in the network.

5. RESULTS AND DISCUSSION

Initially, the raw dataset obtained from the IMDB repository was not compatible with KNIME because it lacked the support to read TSV-type files. Consequently, the datasets containing the IMDB tables were converted into CSV format. As the research progressed, it became evident that the IMDB database volume was enormous, reaching millions of data entries. This scale of the dataset posed challenges. It became impractical to execute network analysis as both the Igraph and KNIME code execution environments

https://doi.org/10.24191/jcrinn.v9i2.455

became unresponsive after prolonged execution times. Therefore, a decision was made to narrow down the data analysis process to focus exclusively on the data covering information for 2021. Although the 2023 data was complete, it had fewer data than 2021; thus, the network analysis revolved around using the latter.

A cleansing process was initiated during which unwanted data was removed from the datasets. For example, records of individuals without assigned job categories were filtered out and excluded from the analysis. It was discovered that the job category column, instead of containing meaningful values, had instances with '\N'. Additionally, there were records with multiple copies. To address this, KNIME's 'Duplicate Row Filter' node was employed, successfully eliminating the duplicate entries within the dataset, as illustrated in Fig. 3 below.

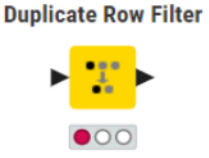


Fig. 3. Application of KNIME's Duplicate Row Filter

Fig. 4 illustrates the application of KNIME's Duplicate Row Filter node to remove duplicates. Following the utilisation of this node, there was a significant reduction in the amount of data used in this study from 702165 to 202018 rows.

Rows: 702165 | Columns: 2

#	RowID	source String	target String
1	Row0	nm0749914	nm0031652
2	Row1	nm0765384	nm0031652
3	Row2	nm0749914	nm0250185
4	Row3	nm0765384	nm0250185
5	Row4	nm0749914	nm15107981
6	Row5	nm0765384	nm15107981
7	Row6	nm0749914	nm0539221
8	Row7	nm0765384	nm0539221
9	Row8	nm0079782	nm0425543
10	Row9	nm0079782	nm0061442
11	Row10	nm0079782	nm0773694

Rows: 202018 | Columns: 2

#	RowID	source String	target String
1	Row0	nm0749914	nm0031652
2	Row1	nm0765384	nm0031652
3	Row2	nm0749914	nm0250185
4	Row3	nm0765384	nm0250185
5	Row4	nm0749914	nm15107981
6	Row5	nm0765384	nm15107981
7	Row6	nm0749914	nm0539221
8	Row7	nm0765384	nm0539221
9	Row8	nm0079782	nm0425543
10	Row9	nm0079782	nm0061442
11	Row10	nm0079782	nm0773694

Fig. 4. Reduction in the dataset size after removing duplicates

These data were removed not only to uphold the accuracy and quality of output produced by the network analysis models but also for the practical benefit of reducing the number of rows, thus lessening execution time.

To address the research questions outlined earlier, we implemented specific centrality measure functions available in Igraph, encompassing Betweenness, Closeness and Degree centrality. The code snippet of the functions used is shown below in Fig. 5.

https://doi.org/10.24191/jcrinn.v9i2.455

```
degree = Graph.degree(g)
max_degree = max(degree)
closeness = Graph.closeness(g)
max_closeness = max(closeness)
betweenness = Graph.betweenness(g)
max_betweenness = max(betweenness)
```

Fig. 5. Python code related to Igraph Centrality Models

The results obtained from executing the respective functions reflect only the node with the highest centrality value, determined by the utilisation of the "max ()" function. Table 1 displays the outcomes of all centrality functions for each research question, specifically focusing on the data from 2023.

Table 1. Results of the centrality measure score for each question

Questions	Betweenness	Closeness	Degree
Q1 (Most influential person)	S' Cephas, 80108510.3	Paul Keyes, 1.0	Dick Wolf, 307
Q2 (Job category)	Actor, 0.0	Actor, 0.649285	Actor, 449867
Q3 (Director)	Seth Gamble, 2619.167	John Carpenter, 1.0	Mike Koslov, 199

It was observed that the nodes in the IMDB network represent individuals without categorisation based on their job descriptions, such as actors or directors. Therefore, the analysis included everyone in the dataset. In response to question 1, which was to determine the most influential person in the IMDB dataset, the focus was placed on the degree of nodes, where the maximum degree corresponds to the individual with the most connections to other people. According to the results in Table 1, Dick Wolf had the highest degree, signifying connections to other people totalling 307 unique connections. Moreover, it is observed that the resultant scores for the centrality measures in questions 1 and 3 are unique to each other, making it challenging to distinguish who the most influential person and director is. This was evident as Paul Keyes attained the highest closeness score of 1 while S' Chephas had the highest betweenness score, amounting to 80108510.29.

The results obtained in order to answer Question 2, which focused on determining the most influential job category, identified that the actors have the most influence, obtaining the highest score for each centrality measure. To elaborate, actors emerged as the most popular nodes within the network, achieving the maximum degree centrality score. Moreover, actors are considered significant, surpassing other job categories in closeness centrality. This implies that actors can easily reach other people. One could speculate that this is due to their work in movies or TV shows, which closely interconnect with other individuals. Consequently, information can be disseminated more rapidly by actors than by individuals in any other occupation. This underscores the prominent role that actors play in the movie industry. However, relying on only one centrality measure would be insufficient to prove the influence of a person.

The application and utilisation of the Python code for Igraph were further enhanced to facilitate the process of obtaining data on the top 5 highest scores for each centrality measure. This was conducted to ascertain whether individuals who obtained the highest score for specific centrality measures also secured the top spots in other centrality measures. The responses to Question 1 are shown in Table 2, which suggests that the node with the highest degree of centrality does not appear in the top 5 for both betweenness and

https://doi.org/10.24191/jcrinn.v9i2.455

closeness centrality. This reveals that having many connections does not necessarily mean that the individual can reach other people faster.

Table 2. Top 5 Highest Degree Centrality for Q1

Name	Degree Centrality
Dick Wolf	307
Midnight	295
Tony Warren	258
Ricky Greenwood	254
Lapis Afterglow	248

Table 3. Top 5 Highest Closeness Centrality for Q1

Name	Closeness Centrality
Paul Keyes	1
Jack Arnold	1
Antoni Krauze	1
Edwared Zebrowski	1
Max Varnel	1

Table 4. Top 5 Highest Betweenness Centrality for Q1

Name	Betweenness Centrality
S' Cephas	80108510.29
Cameron Dixon	76814345.13
Shotaro Ishinomori	75837316.69
Reg Watson	73460596.55
Chiqui Carabante	61545917.24

Upon analysing the data presented in Tables 2, 3, and 4, it became evident that no individual attained high scores for different centrality measures. Instead, each individual's high score remained exclusive to a particular centrality measure. This observation was consistent across the data sets. Similar efforts were made in the context of Question 3 to identify any recurring individuals in alternative centrality measures yielding comparable results.

Table 5. Top 5 Highest Degree Centrality for Q3

Name	Degree Centrality
Mike Koslov	199
Ricky Greenwood	191
Chad Cunningham	184
Matthew Winters	163
Rhiannon Anatomik	157

https://doi.org/10.24191/jcrinn.v9i2.455

Table 6. Top 5 Highest Closeness Centrality for Q3

Name	Closeness Centrality
John Carpenter	1.0
Jean-Luc Godard	1.0
Jean-Paul Battaggia	1.0
Bille August	1.0
Neil LaBute	1.0

Table 7. Top 5 Highest Betweenness Centrality for Q3

Name	Betweenness Centrality
Seth Gamble	2619.16
Mick Blue	2164.16
Dana Vespoli	1574.91
Whitney Wright	1468.08
Casey Calvert	1293.33

Similar to the outcomes presented in Tables 2, 3, and 4, executing the Igraph algorithm on the director dataset revealed that a node tends to possess the highest value for only one respective centrality measure. We can identify that the individuals with the highest scores for each centrality, Mike Koslov, John Carpenter, and Seth Gamble, differ from the results for question 3 in Tables 5, 6, and 7. It is also the same case for the remaining individuals within the Top 5 highest scores, where we do not see the same names achieving high scores for other centrality measures.

When looking into the perspective of visualizing the network using Pyvis, we gain valuable insights to help make decisions. Attempts to develop visualizations of the network graph based on the three research questions were made. However, after executing the developed Python codes with the Pyvis package, results were only returned for Question 3. This is because the IMDB dataset was too large, exceeding 100 million records.

Fig. 6 shows the visualization of the most influential director which answers Question 3.

https://doi.org/10.24191/jcrinn.v9i2.455

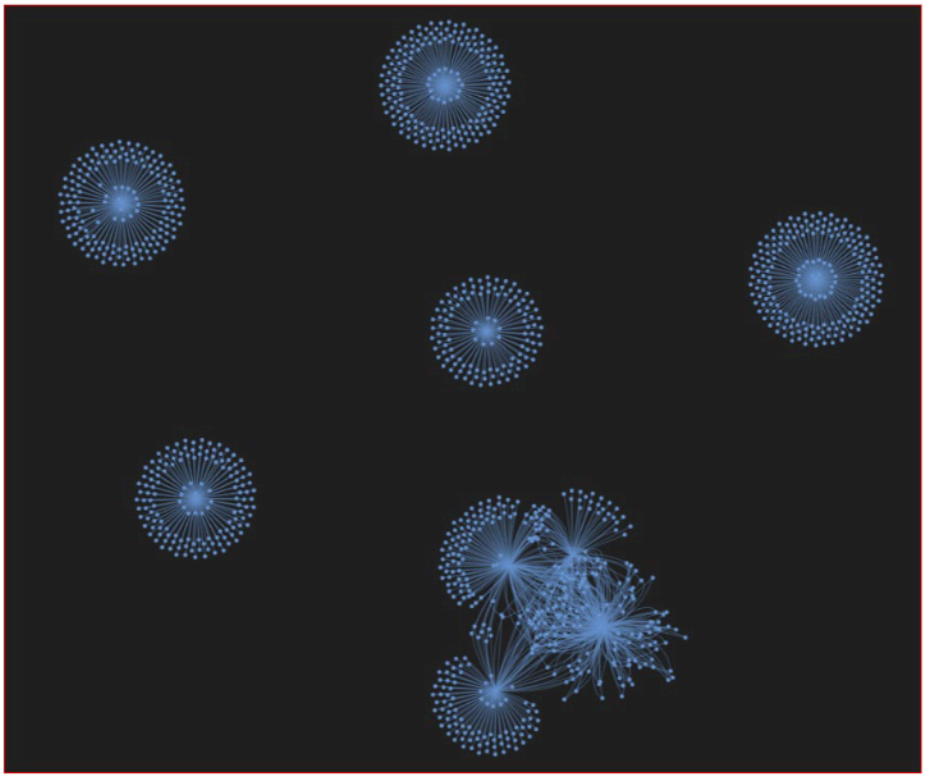


Fig. 6. The overall network of directors

Based on Fig. 6 above, we can see several node groups in the network, five of which are isolated, and only one connected component is seen interacting with other node groups. When diving deep into the individuals that obtained the highest score for Degree Centrality, Closeness, and Betweenness Centrality, like in Tables 5,6 and 7, namely Mike Koslov, John Carpenter, and Seth Gamble, we can see the different behaviors that each of them has. Fig. 7 shows how Mike Koslov interacts with other nodes within the network.

https://doi.org/10.24191/jcrinn.v9i2.455

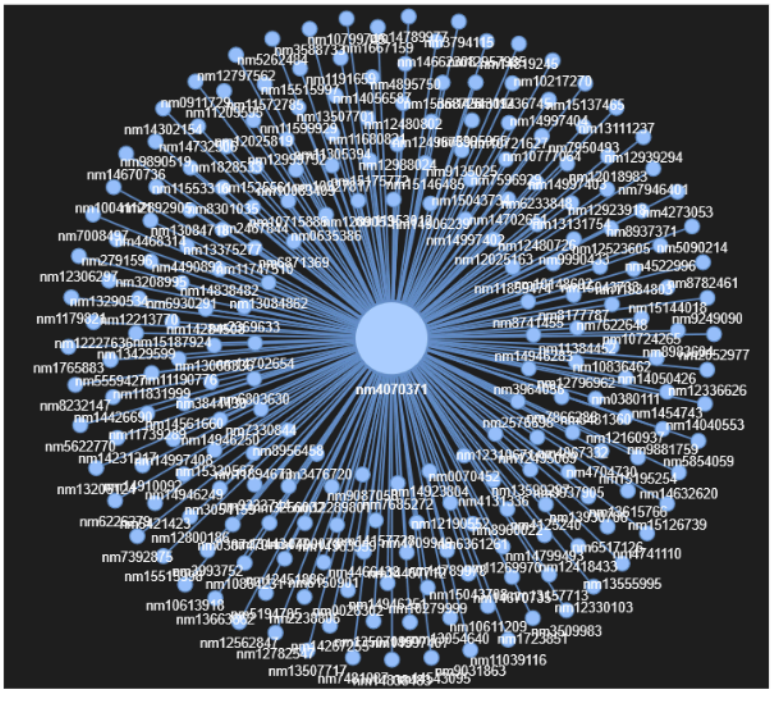


Fig. 7. Graphical view of the interactions of Mike Koslov (nm4070371) with other nodes

The centered node is Mike Koslov with an ID of nm4070371. He is the most critical person in this group, connected to everyone on the graph. Furthermore, this is validated by the fact that he obtained the highest score of 199 for degree centrality. It was unexpected that the closeness centrality aspect of the network analysis did not return any visualization. We looked back at the scores for the measure and identified that everyone obtained a score of 1.0. As Closeness centrality calculates the sum of the distance from one node to all other nodes in a network, it would mean that the distance for each node is equal (Pendong et al., 2024). Next, we look at the individual with the highest betweenness centrality, Seth Gamble, who scored 2619.16. Fig. 8 shows Seth Gamble's interaction with the people around him.

https://doi.org/10.24191/jcrinn.v9i2.455

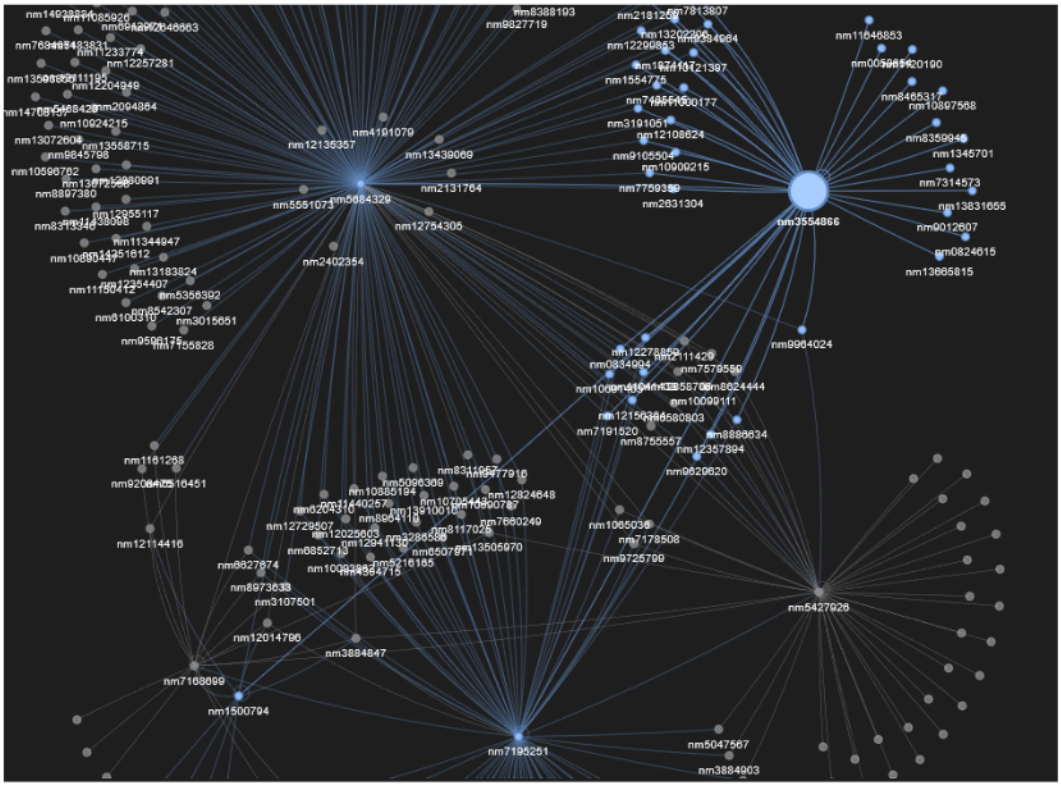


Fig. 8. The interaction between Seth Gamble (nm3554866) with the people around him

Based on Fig. 8, due to having the highest score for the betweenness centrality, it can be seen that Seth Gamble with ID nm3554866 is directly connected to two prominent nodes within their respective groups. They are Bree Mills (nm7195251) and Siouxsie Q (nm5684329). Since Seth Gamble has the highest betweenness centrality of 2619.16, which directly connects the two nodes representing their groups, it indicates that they rely on him for more efficient communication within the network. This occurrence can be treated as Seth being the bridge between the three groups. In addition, the blue lines link together with Seth Gamble, showing the connectivity he has with other nodes. So, disseminating information to the groups is crucial to have Seth involved because, without him, the groups will have no connection with each other. His absence thus disrupts the communication between the respective node groups.

Knowing that a person like Mike Koslov or Seth Gamble is vital within their respective groups, however, it did not reflect the entirety of the IMDB network. Since both of them only obtained the highest scores for unique centralities, it is not evident that they are the most influential in the dataset. This study illustrated that the likelihood of a node in a network achieving high scores across multiple centralities is slim after the network analysis. Consequently, a pertinent question arose regarding determining the most influential person in the 2023 IMDB dataset. Another aspect that can be explored is the ratings of each TV show or movie in the IMDB dataset. Ratings are a way to determine the success of a TV show or movie based on consumers' reactions to whether they feel satisfied or unsettled with what they have watched. Concerning this, the results retrieved from the execution of Igraph centrality measure functions in Tables 1 to 7 were compared with the ratings based on the different settings for each question. For instance, in Network Science terminology, it can be assumed that an actor has the most influence in the dataset if he has high scores for any of the centrality measures and high ratings for the TV shows or movies the actor has worked in. To assess the impact that directors have on the success of movies or TV shows, we examined

https://doi.org/10.24191/jcrinn.v9i2.455

the highest attainable degree centrality from the user node. We compared the movie they worked on with the ratings for the respective show dataset. Table 8 presents the top 5 individuals with the highest degree centrality, along with the number of TV shows they were involved in and the average ratings for 2023.

Table 8. Number of Shows and their Average Ratings for top the 5 Highest Degree Centrality

Actor	TV Shows / Movies	Average Ratings
Mike Koslov	7	5.57
Ricky Greenwood	25	6.728
Chad Cunningham	1	7.4
Chad Cunningham	53	7.9
Matthew Winters	5	5.22
Rhiannon Anatomik	2	6.4

Upon examining Table 8, we observed that the number of shows did not significantly impact the average ratings of the shows an actor has worked in. Additionally, it came to our attention that two individuals with identical names were identified as Chad Cunningham. Further exploration of the dataset revealed that the first Chad was an actor, while the other served in multiple roles as a director, producer, and writer. This finding underscores the notion that nodes with higher degrees did not necessarily equate to high influence in a network. The question at hand was how we determine the most influential person in the 2023 IMDB dataset. Returning to the definition of the centrality measures, the choice depends on the specific objective of measuring influence. Suppose the aim is to identify an individual with many connections while neglecting the proximity of those connections. In that case, Degree Centrality, counting the number of edges a node has, would be the most suitable option. On the other hand, if the goal is to reach other people quickly, prioritizing Closeness Centrality would be appropriate. Additionally, if the focus is on maintaining connections between different groups of people, then Betweenness Centrality becomes more beneficial. Furthermore, it could be that all three centralities measured in this study were insufficient to provide a clear understanding of the influence of nodes within the IMDB dataset. Therefore, employing other various centrality measures might demonstrate varied results.

6. CONCLUSION

In summary, the research highlights that different centrality algorithms provide different results in determining influence in a network. Moreover, from the results obtained, there is a high tendency for nodes to consistently receive high scores for only one measure rather than for the other measures. Based on this, we cannot confidently say that a person like Mike Koslov is the most influential Director, although he received the highest Degree score. Furthermore, even with having the most connections to other people, it is still uncertain if the individual can reach other people that fast since the connections can be far apart from each other. From here, the centrality measure score becomes less meaningful. This study highlights that influence cannot be accurately measured solely by obtaining high centrality scores for a single measure. If the research objective focuses solely on understanding one type of centrality measure, such as the degree of nodes, then degree centrality would be the most appropriate measure to utilise. Apart from that, the comparison between the results of executing the Igraph centrality measure functions with the rating data in IMDB showed that having the highest degree of centrality score did not significantly impact the average ratings of the shows that the actor has worked in.

Several apparent limitations were encountered during this study. The sheer volume of the IMDB datasets significantly impacted the performance of the Data Processing and Model Implementation phase in KNIME. Not only that, but due to the data structure of the IMDB dataset, the KNIME job became complex as the preprocessing phase, which includes data integration, cleansing, and transformation, involved multiple operations. Next, we stumbled upon another limitation: executing the Igraph code

https://doi.org/10.24191/jcrinn.v9i2.455

required approximately 4 hours to complete the task. Specifically, there was a period during the code execution process when the kernel stopped working due to heavy usage of RAM and CPU.

Recognising the limitations that were encountered, we are committed to improving our study in the future. We believe that exploring other Network Science related packages or methods like NetworkX, and diving into other centrality measures like Page Rank and Eigenvector Centrality, is crucial. These efforts will not only enhance the understanding of the influence of nodes in the IMDB dataset but also contribute to the advancement of network analysis and data processing.

7. ACKNOWLEDGEMENTS/FUNDING

The authors would like to acknowledge the support of the Malaysian Agricultural Research and Development Institute (MARDI) for providing insights for this study. Additionally, I would like to express my gratitude to the JCRINN Editorial Board for publishing this manuscript in JCRINN.

8. CONFLICT OF INTEREST STATEMENT

The authors agree that this research was conducted in the absence of any self-benefits or commercial or financial conflicts and declare the absence of conflicting interests with the funders.

9. AUTHORS' CONTRIBUTIONS

Muhammad Izzat Farid Musaddin carried out the research, wrote and revised the article.

10. REFERENCES

Aaltonen, A., Alaimo, C., & Kallinikos, J. (2021). The making of data commodities: Data analytics as an embedded process. *Journal of Management Information Systems*, *38*(2), 401–429. https://doi.org/10.1080/07421222.2021.1912928

Askar, M., Cañadas, R. N., & Svendsen, K. (2021). An introduction to network analysis for studies of medication use. *Research in Social and Administrative Pharmacy*, *17*(12), 2054-2061. https://doi.org/10.1016/j.sapharm.2021.06.021

Abbasi, F., & Fazl-Ersi, E. (2022). Identifying influentials in social networks. *Applied Artificial Intelligence*, *36*(1). https://doi.org/10.1080/08839514.2021.2010886

Bacsi, Z., Fekete-Farkas, M., & Ma'ruf, M. I. (2023). A graph-based network analysis of global coffee trade—The impact of COVID-19 on Trade Relations in 2020. *Sustainability*, *15*(4), 3289. https://doi.org/10.3390/su15043289

Bloch, F., Jackson, M. O., & Tebaldi, P. (2023). Centrality measures in networks. *Social Choice and Welfare*, *61*(2), 413–453. https://doi.org/10.1007/s00355-023-01456-4

Centola, D. (2022). The network science of collective intelligence. *Trends in Cognitive Sciences*, *26*(11), 923–941. https://doi.org/10.1016/j.tics.2022.08.009

Gladysz, B., Tran, T.-A., Romero, D., Van Erp, T., Abonyi, J., & Ruppert, T. (2023). Current development on the Operator 4.0 and transition towards the Operator 5.0: A systematic literature review in light of Industry 5.0. *Journal of Manufacturing Systems*, 70, 160–185. https://doi.org/10.1016/j.jmsy.2023.07.008

Hernández S., D., & Sánchez G., D. (2020). Centrality measures in simplicial complexes: Applications of

topological data analysis to Network Science. *Applied Mathematics and Computation*, 382, 125331. https://doi.org/10.1016/j.amc.2020.125331

Henning, M. A., & van Vuuren, J. H. (2022). *Graph and network theory: An applied approach using Mathematica*. Springer International Publishing.

IMDb. (n.d.). *General Information*. https://help.imdb.com/article/imdb/general-information/can-i-use-imdb-data-in-my-software/G5JTRESSHJBBHTGX

Kinast, B., Ulrich, H., Bergh, B., & Schreiweis, B. (2023). Functional requirements for medical data integration into knowledge management environments: requirements elicitation approach based on systematic literature analysis. *Journal of Medical Internet Research*, 25, e41344. https://doi.org/10.2196/41344

Lewis, R. (2024). *IMDb. Encyclopedia Britannica*. https://www.britannica.com/topic/IMDb

Liu, Y., & Ma, Y. (2022). Quantifying award network and career development in the movie industry. *Frontiers in Physics*, 10. https://doi.org/10.3389/fphy.2022.902890

Mann, C. F., McGee, M., Olinick, E. V., & Matula, D. W. (2022). Flowthrough centrality: A stable node centrality measure. *Journal of Data Science*, 696–714. https://doi.org/10.6339/22-jds1081

Mostafavi, H., Spence, J. P., Naqvi, S., & Pritchard, J. K. (2022). *Limited overlap of eQTLs and GWAS hits due to systematic differences in discovery*. bioRxiv. https://doi.org/10.1101/2022.05.07.491045

Naderi, H., & Shojaei, A. (2023). Digital twinning of civil infrastructures: Current state of model architectures, interoperability solutions, and future prospects. *Automation in Construction*, 149, 104785. https://doi.org/10.1016/j.autcon.2023.104785

Perrone, G., Unpingco, J., & Lu, H. M. (2020). *Network visualizations with Pyvis and VisJS*. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2006.04951

Pendong, C. H. A., Suoth, E. J., Fatimawali, F., & Tallei, T. E. (2024). Network pharmacology approach to understanding the antidiabetic effects of pineapple peel hexane extract. *Malacca Pharmaceutics*, *2*(1), 24–32. https://doi.org/10.60084/mp.v2i1.162

Qazi, A., Hardaker, G., Ahmad, I. S., Darwich, M., Maitama, J. Z., & Dayani, A. (2021). The role of information & communication technology in elearning environments: A systematic review. *IEEE Access,* 9, 45539–45551. https://doi.org/10.1109/access.2021.3067042

Sadaf, A., et al. (2024). A bridge between influence models and control methods. *Applied Network Science*, *9*(1). https://doi.org/10.1007/s41109-024-00647-x

Zhu, J., Chong, H.-Y., Zhao, H., Wu, J., Tan, Y., & Xu, H. (2022). The application of graph in BIM/GIS Integration. *Buildings*, *12*(12), 2162. https://doi.org/10.3390/buildings12122162

https://doi.org/10.24191/jcrinn.v9i2.455