**Facebook TV Show Page Network Analysis**

**Abstract**

TV SHOWS

This case study leverages a large dataset of Facebook pages and their interactions (likes) from November 2017. We analyze this data as a network, where pages are represented by nodes and edges depict connections between pages with mutual liking. It contains 3,892 nodes and 17,262 edges. The study focuses on understanding the structure and characteristics of the network, identifying key nodes and communities, exploring the potential implications for social media and content distribution.

Various graph analysis techniques and machine learning models were applied to extract insights and predict viewer engagement levels within the network. Additionally, a Graph Convolutional Network (GCN) model was implemented to leverage the structural properties of the graph data for deeper analysis and prediction.

**About the dataset**

The dataset contains information about the relationships between the characters in the network. The dataset includes 8 different distinct types of pages, with 3,892 nodes and 17,262 edges specifically for the TV Shows category. Each row in the file represents an edge between two Facebook pages of TV shows with mutual likings.

-Each node represents individual Facebook pages. Each page is a unique identifier within the dataset.

-Each edge represents the relationship between pages established by "likes." An undirected edge connects two pages if page A likes page B, and vice versa. This means information flow and influence can potentially go in both directions within the network.

**Objective**

The objective of this project is to analyze the network structure of Facebook TV show pages, identify influential nodes, and predict viewer engagement levels based on mutual liking connections. Our aim is to understand the structure and characteristics of the network, identify key nodes and communities, and explore the relationship between page content and user engagement. By analyzing these aspects, we hope to extract actionable insights for content creators, marketers, and platform administrators to optimize audience engagement and network growth strategies, gain valuable insights into how information flows through the network and how users interact with different types of content.

**Insights**

1. Network Structure: Understanding the overall structure of the network, including the distribution of nodes and edges, density, and transitivity.

2. Key Nodes and Communities: Identifying the most influential nodes and communities within the network, which could help in understanding the popularity and distribution of content.

3. Implications for social media and Content Distribution: Exploring the potential implications of the network structure and key nodes for social media platforms, content creators, and marketers.

4. Potential for Misinformation and Echo Chambers: Analyzing the potential for misinformation and echo chambers within the network, which could have significant consequences for social media users and society.

5. Genre Preferences: By studying the mutual likings between TV show pages, network analysis can uncover patterns related to genre preferences among Facebook users. This information can be valuable for content creators, networks, and advertisers to tailor their strategies.

6. Comparative Analysis: Network analysis allows for the comparison of different TV show categories, networks, or platforms based on mutual likings. This comparative analysis can highlight trends, similarities, and differences in audience preferences.

7. Predictive Insights: By studying mutual likings over time, network analysis can provide predictive insights into the popularity of TV shows, potential trends, and audience engagement levels.

8. Potential Applications: Exploring potential applications of the findings, such as improving content distribution strategies, detecting misinformation, and mitigating echo chambers.

**Visualizing the Network:**



**General Observations**

● Type: Graph

● Number of nodes: 3,892

● Number of edges: 17,262

● Maximum degree: 126

● Minimum degree: 1

● Average degree: 8

**Graph Connectivity:**

**1**. In a Graph the path between every pair of vertices is called edges.

**2**. In our graph we have links between all the nodes.so it helps us to

understand that each tv show page is related to every other tv show page directly

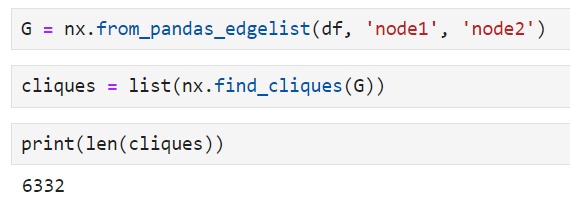
or indirectly.

**Clique:**

● In Graph theory a clique is defined as a maximal subgraph of a graph.

● A clique would represent a subset of pages where every pair of distinct pages is mutually liked, forming a complete subgraph.

● In python we use nx.find\_cliques to find the clique in the graph.



Number of cliques in the Network:6332

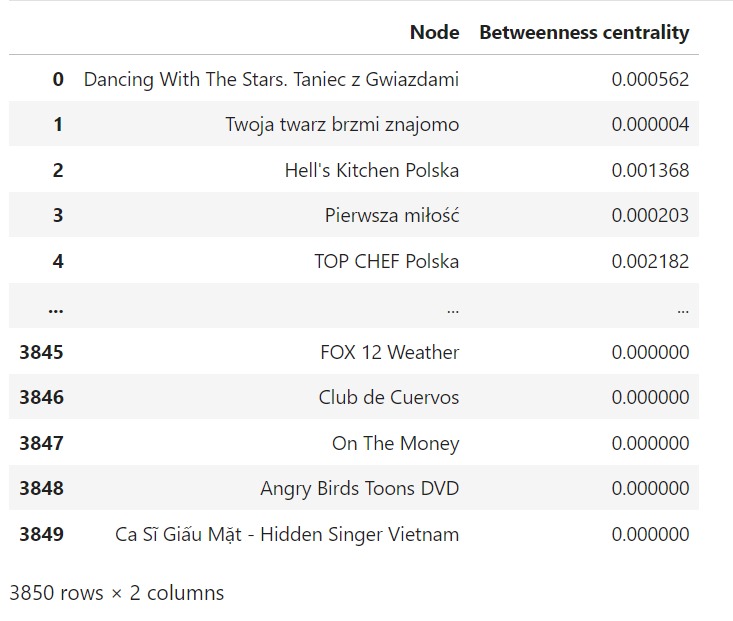
**Centrality Measures**

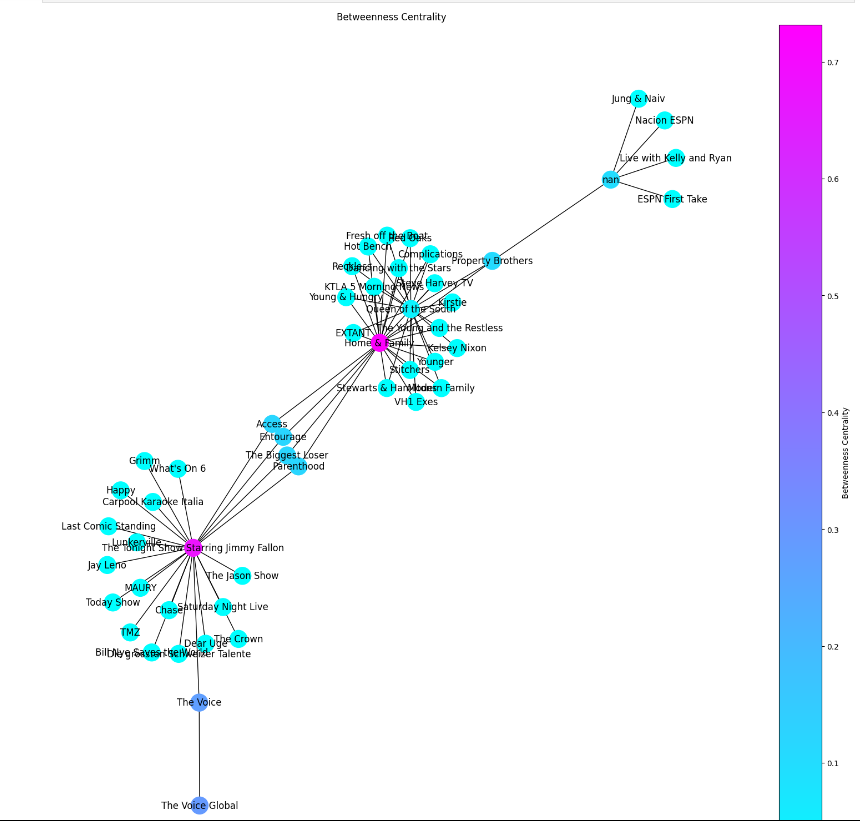
Centrality measures are metrics used to quantify the importance or influence of nodes in a network. Here are some common centrality measures:

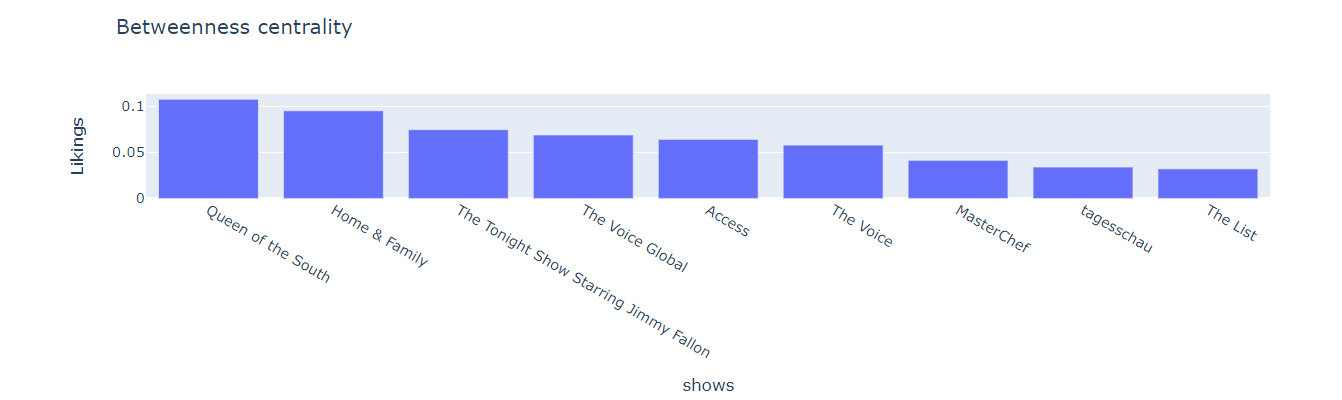
Betweenness, Closeness, Eigenvector, Degree and Harmonic.

**1. Betweenness Centrality:**

This measures the number of shortest paths that pass through a node. A node with a high betweenness centrality acts as a "bridge" between different parts of the network.







Top 10 TV Shows:

1.Queen of the South

2.Home & Family

3.The Tonight Show Starring Jimmy Fallon

4.The Voice Global

5.Access

6.The Voice

7.MasterChef

8.tagesschau

9.The List

10.top chief

Inference:

The table shows a list of tv shows with their corresponding betweenness centrality scores.

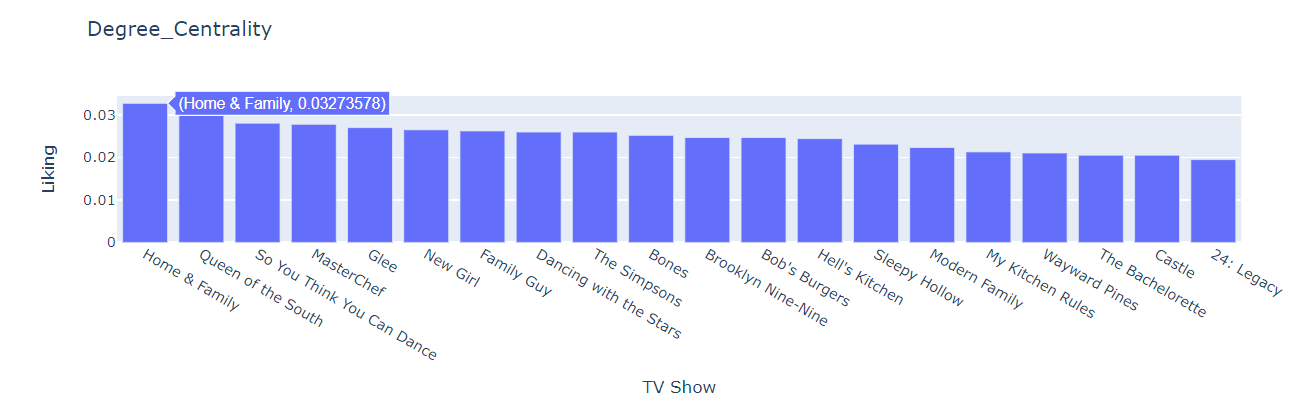
The first tv show, "Queen of the South", has the highest betweenness centrality of 0.1076496. This indicates that this video is frequently watched by users who also watch other diverse shows on Facebook.

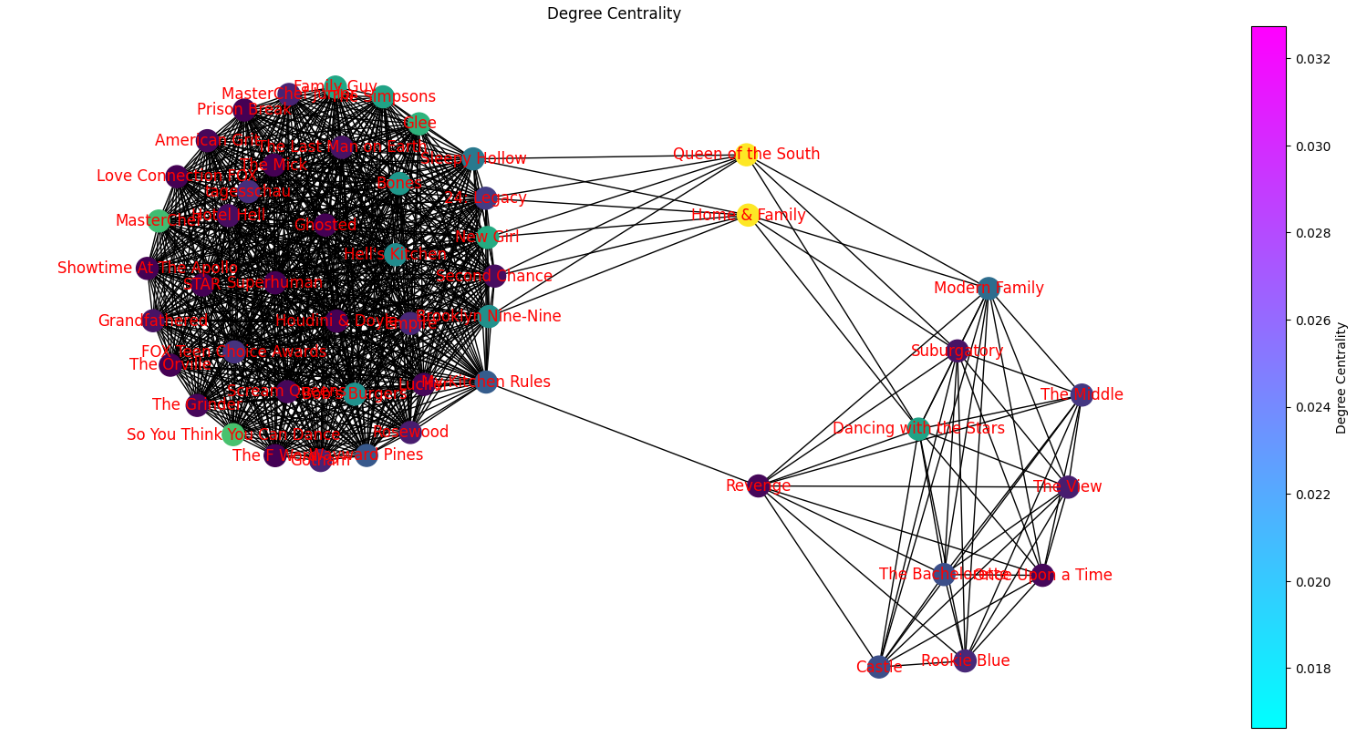
Other shows in the table likely have varying betweenness centrality scores, suggesting their positions as bridges between different viewing interests on the platform.

**2. Degree Centrality:**

This measures the number of edges that are incident to a node. Degree centrality is a measure of a node's importance in a network based on the number of connections it has to other nodes. In the tv shows network, degree centrality can provide insight into which characters are the most connected and potentially the most influential.







Inference:

TV shows with a high degree-centrality is one that is connected to many other shows. This could be because the show has been referenced in a lot of other shows, or because it has appeared in crossover episodes with a lot of other shows. Shows with high degree centrality might belong to the same genre or share a fictional universe. High degree centrality could also indicate a show's broader cultural impact.

TV shows with the highest degree centrality according to the graph:

• Queen of the South

• Home & Family

Shows with the lowest degree centrality include:

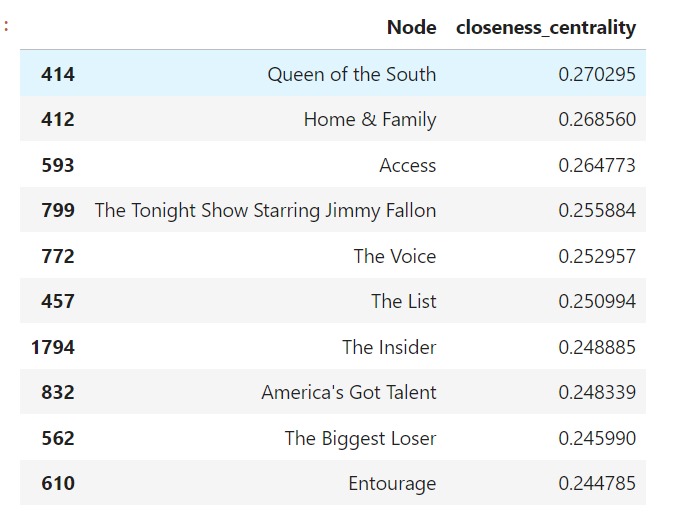
• 24: Legacy

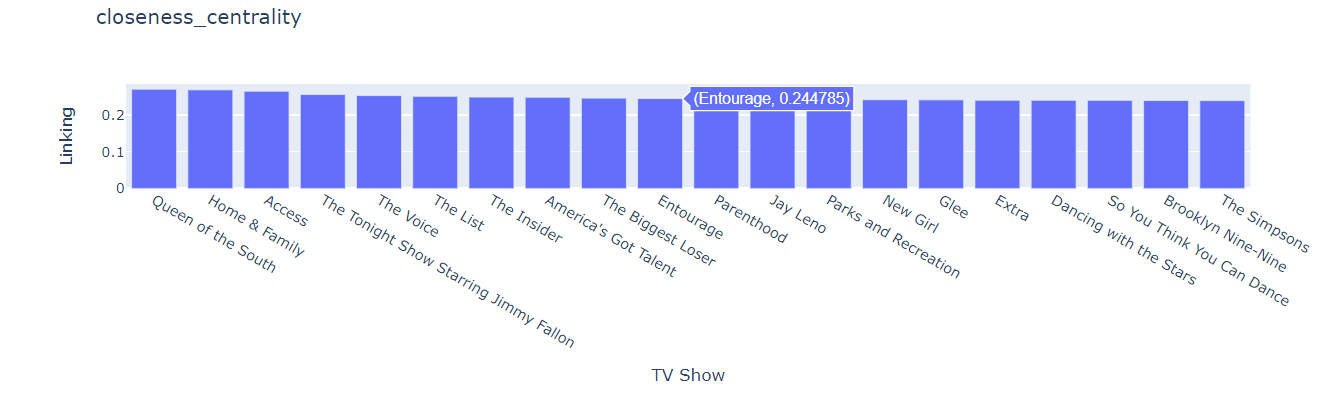
• The Bachelorette

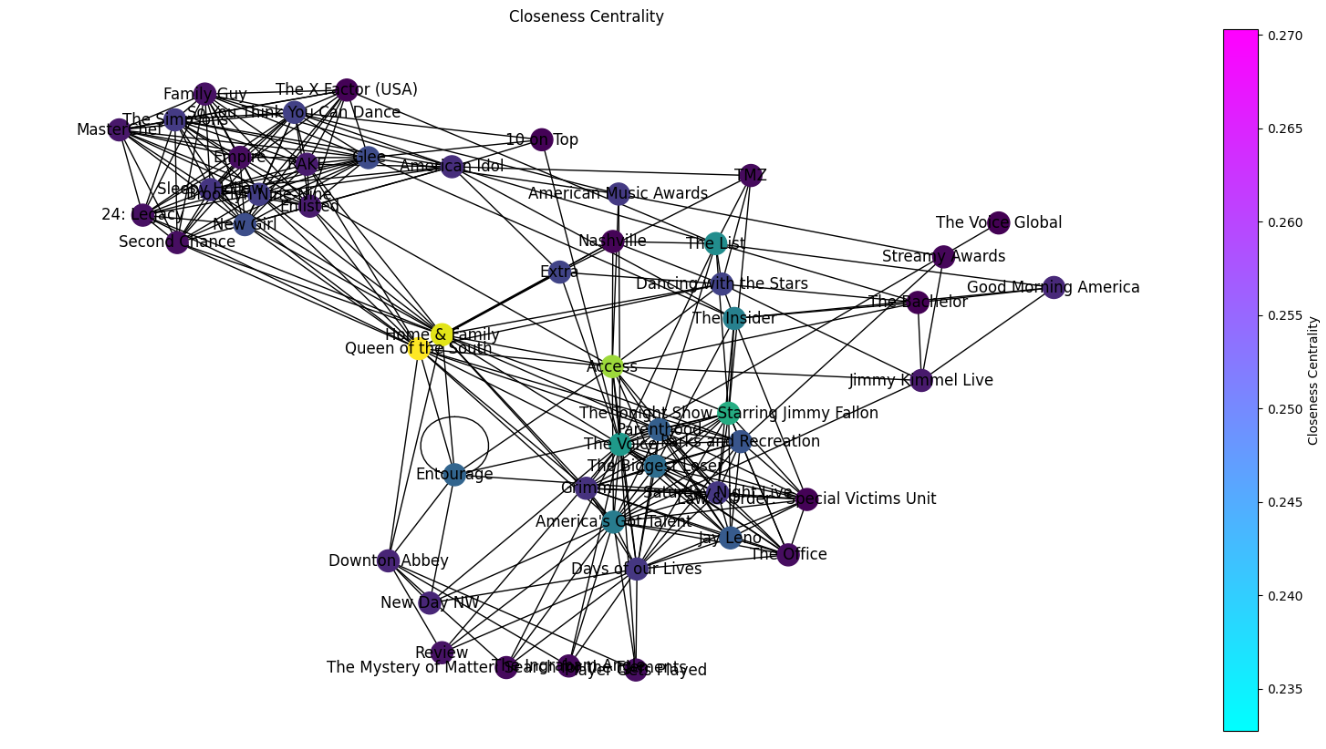
• Castle

**3.Closeness Centrality**

This measures the distance of a node to all other nodes in the network. Closeness centrality is a measure of how close a node is to all other nodes in the network. A node with a high closeness centrality can communicate with many other nodes in a short time.







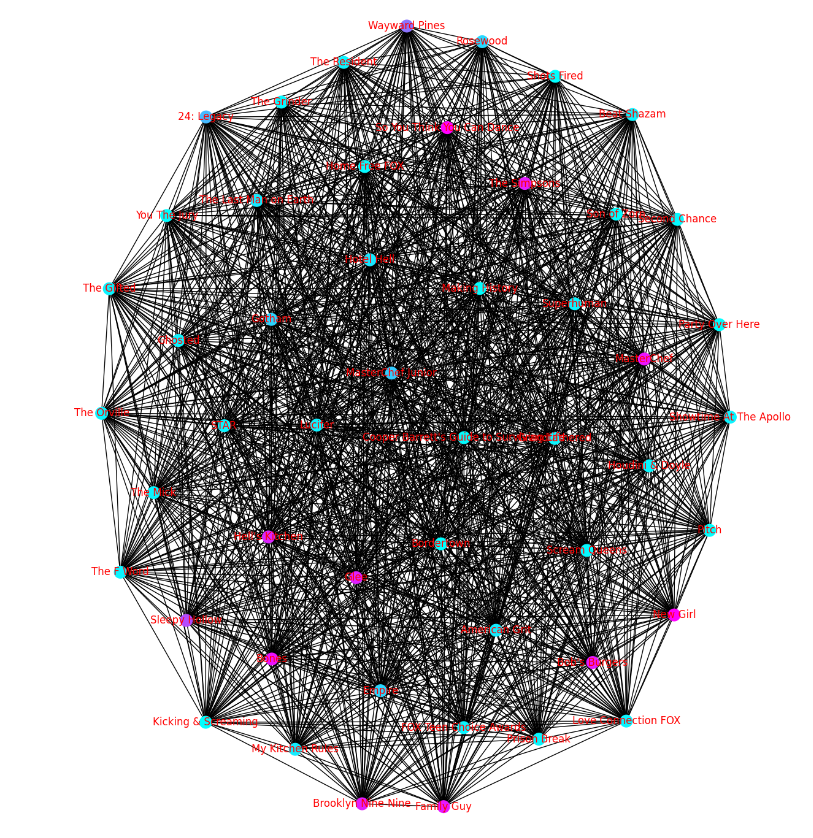
Inference:

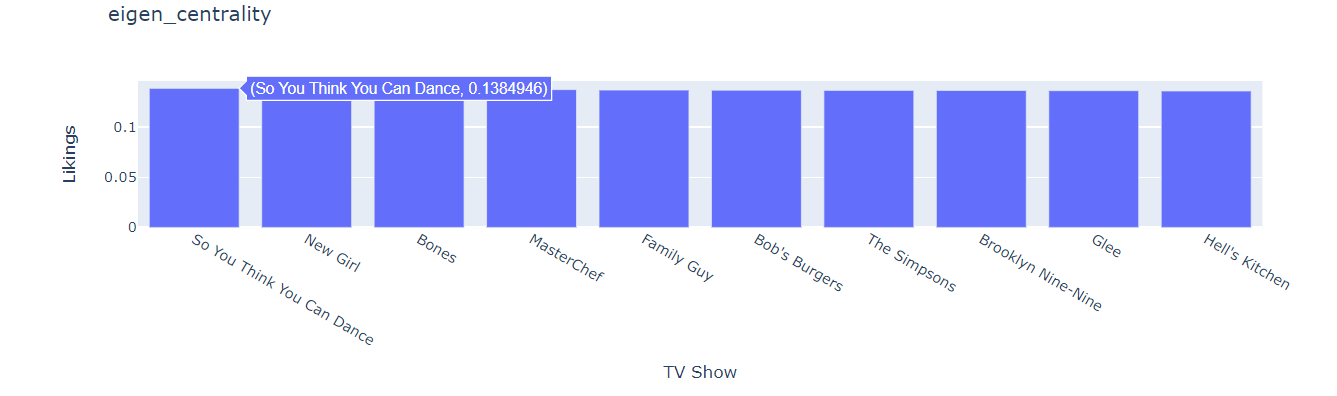
TV shows with a high closeness centrality are the ones that are well-connected to many other shows and can be reached quickly from any other show in the network. This suggests that Queen of the South is a central show in the network of TV shows and may have connections to many other shows either through direct references or by being referenced through common connections. Other shows with high closeness centrality include Home and Family and Access.

Shows with lower closeness centrality, such as The Simpsons, Brooklyn Nine-Nine, So You Think You Can Dance, and Dancing with the Stars, may not be as well-connected to the overall network of shows. This could mean that they are niche shows that appeal to a specific audience, or that they are newer shows that have not yet had the opportunity to build connections with other shows.

**4. Eigenvector Centrality:**

Eigen vector centrality is a network centrality measure that assigns importance scores to nodes based on their connections to other highly connected nodes in the network.





Inference:

Eigenvector centrality is a measure of a node's influence in a network. It considers not only the number of connections a node has, but also the importance of the nodes that it is connected to.

According to the graph, shows like So You Think You Can Dance, New Girl, Bones have the highest eigenvector centrality. This suggests that these shows are influential within the network of TV shows and are likely to be referenced or parodied by other shows. These shows may be trendsetters or cultural touchstones within their genre.

Shows with a lower eigenvector centrality, such as My Kitchen Rules, The Last Man on Earth, and Hotel Hell may not be as influential within the network. This does not necessarily mean that they are low-quality shows, but rather that they may not be as widely referenced or well-connected to other shows in the network...

**5.Harmonic Centrality**

Harmonic centrality is a measure used in network analysis to determine the importance of a node within a network. It is based on the concept of harmonic mean and considers the distances between a node and all other nodes in the network.

A blue squares with white text

Description automatically generated

A map of a network

Description automatically generated with medium confidence

Inference:

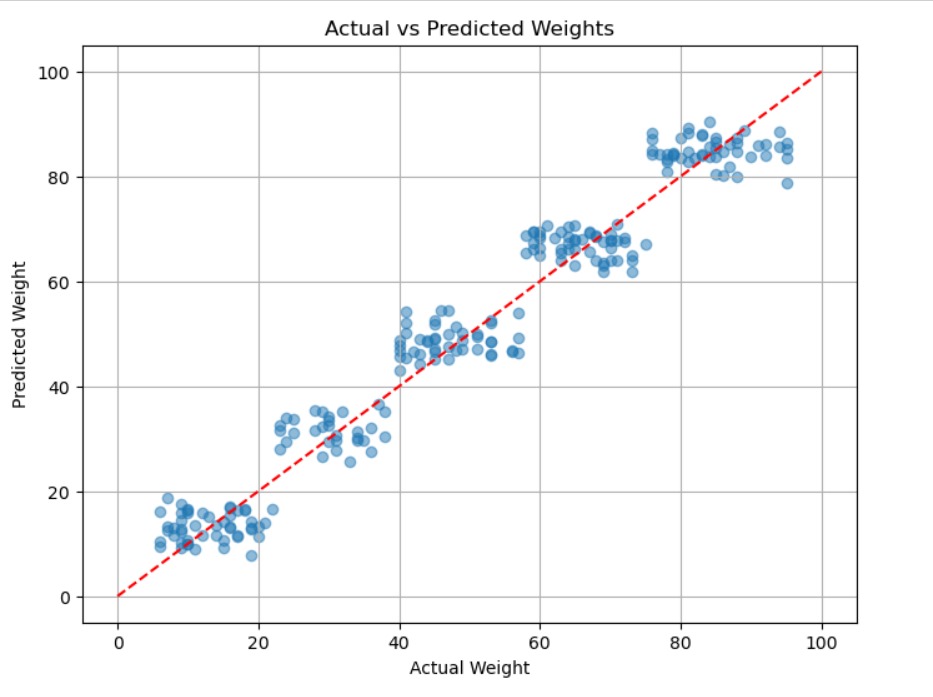
Analysing this social network graph of TV shows reveals interesting connections. A prominent cluster emerges, containing shows like The Bachelor, American Idol, and Dancing with the Stars. These tv show’s thick connecting lines suggest a strong relationship, possibly due to a shared genre (reality TV) or target audience. Conversely, shows on the periphery, like The Mystery of Matter: Searthy for that kinyents, have just a few connections, indicating niche appeal or lower popularity.

The line thickness between shows acts as a gauge as well. The Bachelor's thicker connection to The Bachelorette compared to So You Think You Can Dance implies a greater similarity between the Bachelor-Bachelorette duo.

Overall, this graph provides a glimpse into the interconnected world of television, where genres, themes, and target demographics bind various shows together.

**ML Models:**

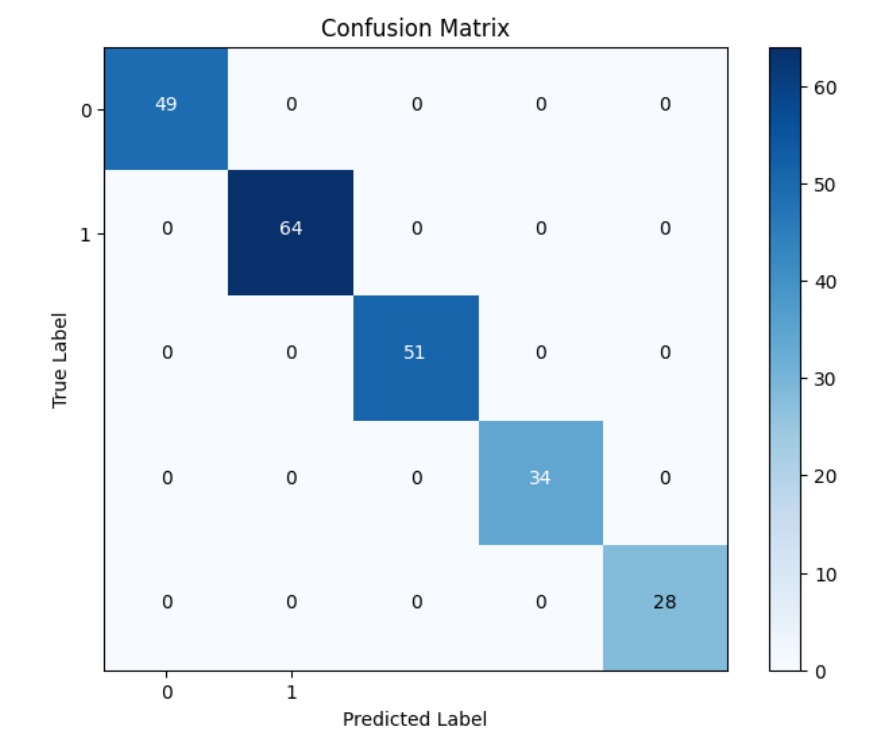
**1.ML Model using Random Forest Regressor:**

****

Inference:

TV Shows with a higher predicted-weights of mutual likes also tend to have a higher actual-weights of mutual likes. There is a positive correlation between actual and predicted weights, though the data points are scattered. The model might be good at identifying some features of shows that tend to be liked by similar audiences. People's taste in TV shows can be subjective and vary considerably, making it difficult to predict perfectly.

**2. ML Model using Decision Tree Classifier**



A white background with black and white text

Description automatically generated with medium confidence

Inference:

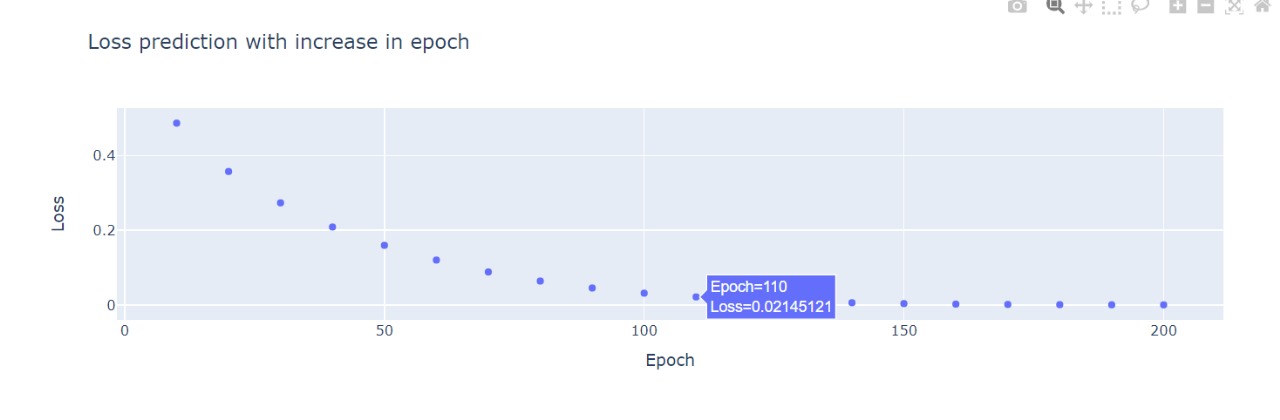
we can infer the following about the model's performance:

* Accuracy: The accuracy of the model can be calculated by dividing the sum of all true positives (49 + 64 + 51 + 34 + 28 = 226) by the total number of samples. This would give an accuracy of 226/226 = 1.0, indicating a high level of correctness in the model's predictions.
* Precision: The precision of the model can be calculated for each class by dividing the number of true positives for that class by the sum of true positives and false positives for that class. For example, for the first class, the precision would be 49/ (49+0).
* Recall: The recall of the model can be calculated for each class by dividing the number of true positives for that class by the sum of true positives and false negatives for that class. For example, for the first class, the recall would be 49/ (49+0).
* F1 Score: The F1 score of the model can be calculated for each class by taking the harmonic mean of precision and recall for that class. This would give an overall measure of the model's performance in terms of both precision and recall.
* Class-wise Performance: The confusion matrix provides a class-wise breakdown of the model's performance, allowing for the evaluation of the model's performance on each class individually. This can be useful for identifying classes where the model is performing poorly and for adjusting the model accordingly.

The fact that all off-diagonal entries are 0s indicates that the model is perfect in its predictions, with no false positives or false negatives. This is an ideal scenario, and the model's performance is as good as it can get

**3. Graph Convolutional Neural Network**





Based on the training progress and loss values provided, we can make the following inferences about the GCN model:

* The model is effectively learning the patterns in the input graph data: The steadily decreasing loss values over the 200 epochs indicate that the GCN model is successfully capturing the underlying structure and relationships within the graph. The final low loss value of 0.000173 suggests the model achieved high accuracy in representing the input graph.
* The model is converging towards an optimal solution: The consistent decrease in loss values without any significant fluctuations suggests that the model is converging towards an optimal solution during the training process. This is a positive sign, as it means the model is learning the graph patterns effectively.
* The model can be used for various graph-based tasks: The trained GCN model can now be used for a variety of tasks on the given graph dataset, such as node classification, link prediction, or graph-level prediction. The learned node representations and graph embeddings can be leveraged for these downstream applications.
* The model has potential for generalization: The successful training of the GCN model on the provided graph dataset indicates that the model architecture and training approach are suitable for this type of graph-structured data. This suggests that the model may also perform well on other similar graph datasets, demonstrating its potential for generalization.

In summary, the GCN model has been trained effectively on the given graph dataset, as evidenced by the steadily decreasing loss values and the final low loss. This implies that the model has learned to capture the underlying patterns and relationships within the graph and can now be used for various graph-based tasks and potentially generalized to other similar datasets.

**Conclusion:**

1. Graph Analysis Insights:

- Through visualizing the graph and computing centralities such as degree centrality, betweenness, closeness, eigen vector, and harmonic centrality, we gained a deep understanding of the network's structure and identified key nodes or TV show pages that are central to the network.

- The addition of the weight column, representing the percentage of viewers who liked both TV show pages, provided valuable information on the strength of connections between nodes. This allowed us to discern the existence of connections and their significance in viewer engagement.

2. Machine Learning Modeling:

- We employed Random Forest Regressor to predict the weight values based on the features extracted from the graph data. The regression plot provided insights into the model's performance and the relationship between predictor variables and the weight.

- Further enriching the dataset with a categorical column based on weight ranges enabled us to implement a Decision Tree Classifier. This model effectively categorized the strength of connections between TV show pages, allowing for more granular analysis and targeted interventions based on viewer engagement levels.

3. Graph Convolutional Network (GCN):

- Implementing a GCN model on the network data provided a sophisticated approach to analyzing the graph structure and predicting relationships between TV show pages. This deep learning technique leverages the inherent structure of the graph to make predictions, potentially offering more nuanced insights compared to traditional machine learning models.

4. Overall Insights and Contributions:

- By combining graph analysis and machine learning techniques, we gained a holistic understanding of the Facebook TV show pages network. We identified influential nodes, measured viewer engagement, and made predictions about the strength of connections between TV show pages.

- Our findings can inform content creators, marketers, and platform administrators about viewer behavior, preferences, and potential strategies for enhancing viewer engagement and expanding audience reach.

5. Future Directions:

- To enhance the predictive capabilities of the models, future research could explore incorporating additional features such as demographic data or content attributes.

- Further experimentation with advanced graph neural network architectures and ensemble learning techniques could lead to improved model performance and more accurate predictions.

- Continuous monitoring and analysis of the network data will enable us to adapt strategies in response to evolving viewer preferences and behaviors, ensuring sustained engagement and growth within the Facebook TV show pages ecosystem.

**Tools and software used:**

Language-Python

Platform-Jupyter Notebook

**Libraries used:**

1. Pandas

2. Networkx

3. Matplotlib

4. Plotly

5. Numpy

6. Seaborn

7. json

**References:**

networkrepository.com

https://pytorch-geometric.readthedocs.io/en/latest/generated/torch\_geometric.nn.models.GCN.html

http://www.orgnet.com/sna.html?ref=dataroots.ghost.io