TECHNICAL REPORT ON NETWORK ANALYSIS ON WIKIPEDIA'S REQUEST FOR ADMINSHIP VOTES

INTRODUCTION

Network analysis is a powerful technique used to study relationships and interactions among entities. In simple terms, it helps us understand how people, objects, or even events are connected. In the context of this task, the focus was on social networks within a voting system.

Many people confuse network analysis with network marketing, which is an entirely different concept. Network marketing is a business model that relies on a network of distributors to grow sales, whereas network analysis is a data-driven approach to studying connections between nodes (individuals, organizations, etc.). The focus here is purely on network analysis to gain insights from the dataset.

Data Source

The dataset used in this analysis originates from the Stanford Large Network Dataset Collection. It represents Wikipedia's request-for-adminship (RfA) votes, where users vote on whether another user should be granted administrative privileges.

Why This Dataset Was Chosen

The dataset used in this analysis comes from the Wikipedia voting dataset, which contains voting interactions between Wikipedia administrators during elections. This dataset was chosen because:

1. It represents a real-world decision-making process: Each vote represents a directed connection between two individuals.
2. It allows us to study influence and trust:Analyzing the voting network enable us to identify key influencers and patterns of approval or disapproval.
3. It provides a foundation for social network analysis: The dataset structure is ideal for studying network metrics and community formation.

E-voting systems play a crucial role in modern decision-making, enabling transparent and efficient elections. Applying network analysis to voting data, we can:

- Identify key influencers in the voting process.
- Understand the flow of voting decisions.
- Detect patterns and anomalies in voting behavior.
- Evaluate voter engagement trends over time.

DATA EXPLORATION

The dataset originally came in a TXT format, which contained raw voting records with user interactions. To make the dataset more structured and easier to analyze, it was converted to CSV format, allowing for better readability and manipulation using Python libraries

Dataset Structure

The dataset consists of the following key columns:

- SRC (Source Voter): The user casting the vote.
- TGT (Target User): The user being voted for.
- VOT (Vote Value): The nature of the vote (1 for support, -1 for oppose, 0 for neutral).
- RES (Election Result): Indicates whether the target user was successfully granted admin status (1) or not (-1).
- Date: The date the vote was cast.
- Time: The exact time of voting.
- YEA (Year of Voting): The year the voting took place.

Steps taken:

- Python libraries such as pandas and numpy were imported. Pandas helps load and manipulate the dataset efficiently. NumPy is useful for handling numerical data and performing mathematical operations.
- Since the dataset had ambiguous column names (SRC, TGT, RES, etc.), the columns were renamed for better readability.

- The first few rows and dataset structure were checked. The dataset contains 198,275 entries with 9 columns. Some columns have missing values, particularly in 'Voters', 'Time', 'Date', and 'Review'.

DATA CLEANING

Cleaning ensures that the data is usable for analysis. address missing values, standardize text formats, and remove duplicates.

Steps taken:

- The 'Voters' column has missing values. They were replaced with anonymous.
- Text data were made clean by stripping unwanted spaces. This removes leading and trailing spaces, ensuring consistency.
- Since date and time are crucial for voting analysis, rows where either 'Date' or 'Time' is missing were removed.
- Column with reviews and vote_type were dropped.

NETWORK ANALYSIS

Since this dataset represents voting as a relationship between voters and candidates, we analyze it as a graph. Networkx library was imported. NetworkX is a powerful Python library for analyzing graphs and networks. It helps measure centrality, connectivity, and influence in the voting system.

Each vote represents a connection (edge) between a voter (node) and a candidate (node). To create a voting network, we represented:

- Voters (SRC) as nodes.
- Candidates (TGT) as nodes.
- Edges as voting interactions, directed from the voter (SRC) to the candidate (TGT).

Research Questions

The analysis was structured around six key questions.They are as follows:

1. Who are the most central actors in the network based on degree centrality?
2. How has voting activity changed over time, and what patterns can be observed?
3. How is the network structured in terms of communities, and what are the dominant groups?
4. Which nodes play the most crucial role in information flow and connectivity within the network?
5. How does degree centrality influence election success?
6. Which node pairs are most likely to form connections in the future based on prediction scores?

Tools and Techniques Used

To perform the analysis, Python and specialized libraries were used:

- Python: Primary programming language for analysis.
- NetworkX: Used for building and analyzing the voting network.
- Pandas: Data preprocessing and manipulation.
- Matplotlib & Seaborn: Visualization of network structures.
- Scikit-learn: Applied for link prediction models.
- Louvain Algorithm: Used for community detection.

Methodologies

This study utilizes Social Network Analysis (SNA) to examine the structure, influence, and evolution of the network. SNA is a method used to study relationships and interactions among individuals, groups, or entities in a network by analyzing the connections (edges) between them. The methodology involves multiple techniques, each used to answer a specific research question.

1. Degree Centrality Analysis

Definition: Degree centrality measures how well-connected a node (individual or entity) is within a network. It is calculated as the number of direct links (edges) a node has to other nodes. A higher degree centrality suggests greater influence and connectivity.

Application:

- A bar chart ranking users by degree centrality was created to identify the most central actors in the network. This helps determine key influencers who facilitate communication, distribute information, or act as network hubs.

2. Temporal Analysis of Voting Activity

Definition: Temporal analysis studies how network activity changes over time. This method is useful for identifying trends, spikes, and declines in engagement.

Application:

- Voting activity was plotted over time (2004–2013) using a time-series graph.This analysis identifies trends such as initial adoption, peak engagement periods, and long-term declines.

3. Community Detection

Definition: Community detection is a technique used in network analysis to identify clusters or groups of nodes that are more densely connected to each other than to the rest of the network. One common method is modularity-based clustering, which measures the strength of division in a network.

Application:

- The network was divided into five communities, and a bar chart was used to visualize the size of each community. Community detection is useful for identifying interest groups, factions, or specialized clusters.

4. Centrality Measures for Influence

Definition: In addition to degree centrality, other measures like betweenness centrality and eigenvector centrality were used to identify influential nodes.

- Betweenness centrality measures how often a node acts as a bridge between other nodes. Nodes with high betweenness centrality control information flow.
- Eigenvector centrality measures a node’s influence based on the importance of its connections. Nodes connected to influential nodes receive higher scores.

Application:

- A color-coded network visualization was used, where darker red nodes indicated higher influence. This analysis identifies key connectors, information brokers, and influential figures within the network.

5. Correlation Between Degree Centrality and Election Success

Definition: This method examines whether a person's connectivity (degree centrality) is associated with their likelihood of winning an election.

Application:

- A scatter plot was generated with degree centrality on the x-axis and election success (1 = success, -1 = failure) on the y-axis. The analysis determines whether higher connectivity improves election success.

6. Link Prediction Analysis

Definition: Link prediction is a machine learning technique used to predict future connections between nodes based on existing network structure.

Tools and Implementation

The analysis was conducted using Python and the following libraries:

- NetworkX – for graph analysis and centrality measures.
- Matplotlib & Seaborn – for visualizing network properties.
- Pandas – for data handling and preprocessing.

QUESTIONS ANSWERED FROM THE ANALYSIS

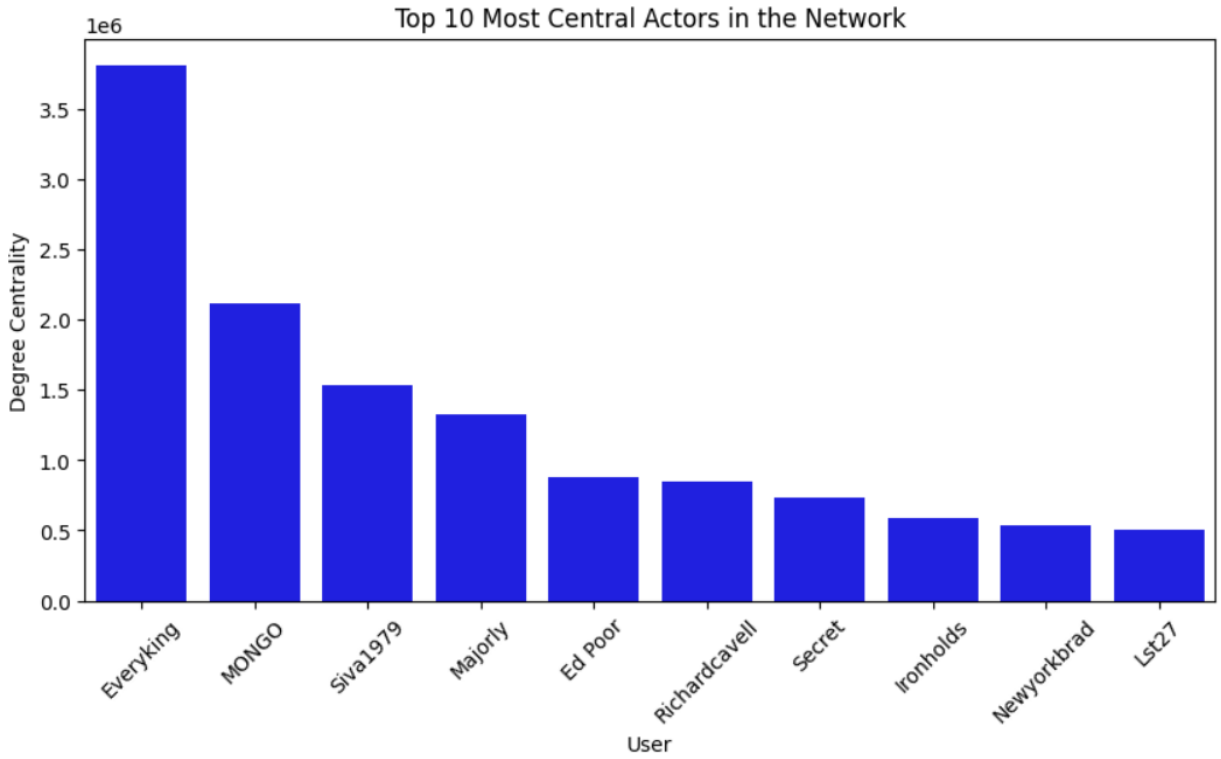
Research Question 1: *Who are the most central actors in the network based on degree centrality?*

The bar chart represents below shows the Top 10 Most Central Actors in the Network, showing users ranked by degree centrality (the number of direct connections they have).

"Everyking" is the most central user, having the highest degree centrality, indicating they are the most connected and influential in the network. "MONGO" and "Siva1979" follow as the next most central users, playing significant roles in network connectivity. The remaining users ("Majorly," "Ed Poor," "Richardcavell," "Secret," "Ironholds," "Newyorkbrad," and "Lst27") have lower but still important centrality values, suggesting they are key secondary actors. The steep decline in centrality from "Everyking" to the rest indicates that influence is highly concentrated at the top, with a few users having far more connections than others.

Insight

Highly central users like "Everyking" likely act as hubs or influential figures, facilitating communication and interaction. The structure suggests a power-law distribution, where a few users dominate the network while others have significantly fewer connections. These central actors may be critical for information flow, opinion leadership, or bridging different communities within the network.



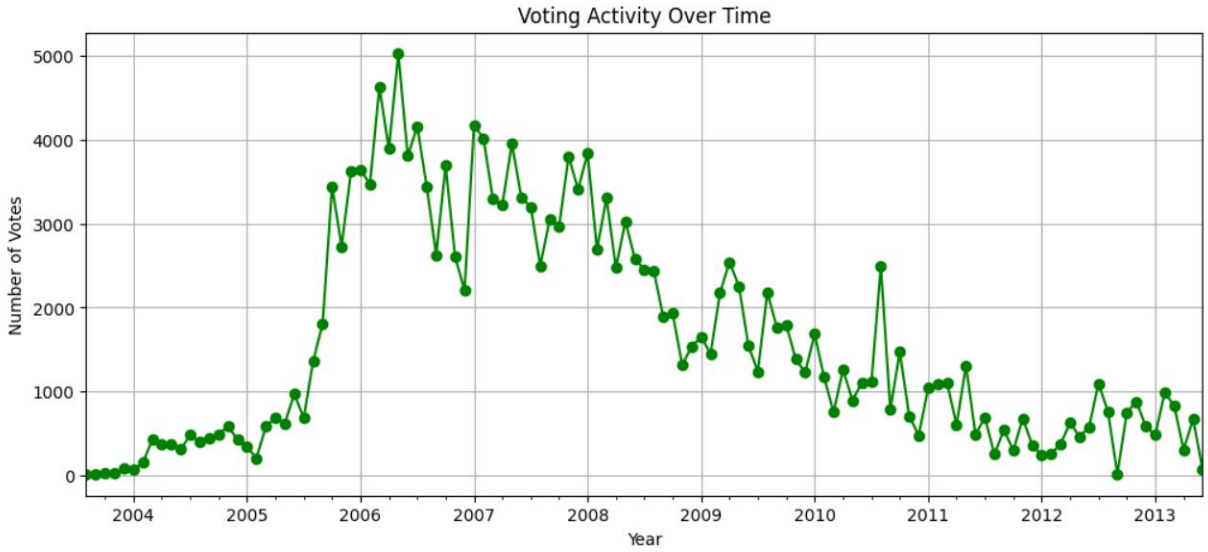
Research Question 2: How has voting activity changed over time, and what patterns can be observed?

The graph below represents voting activity over time. The y-axis represents the number of votes, while the x-axis shows the years from 2004 to 2013.

Key Observations:

Voting activity started at a low level but showed a steady increase in late 2004 and early 2005. This likely indicates a phase of initial adoption and growing user participation. The number of votes surged dramatically, reaching a peak of over 5000 votes around 2006–2007. This suggests a period of maximum engagement, possibly due to increased platform popularity or an event that drove higher participation. After reaching its highest point, voting activity exhibited a downward trend. The decline was gradual but persistent, showing a steady reduction in votes over time. Despite the overall decline, some short-term spikes in activity occurred around 2011 and 2012. These fluctuations suggest temporary periods of renewed interest or external influences that momentarily boosted participation. By 2010–2013, the number of votes had significantly decreased and appeared to stabilize at a lower level, possibly reflecting a shift in user behavior, migration to alternative platforms, or decreased platform engagement.

This trend indicates that voting activity experienced an initial growth phase, reached a peak, and then gradually declined while maintaining sporadic bursts of engagement.

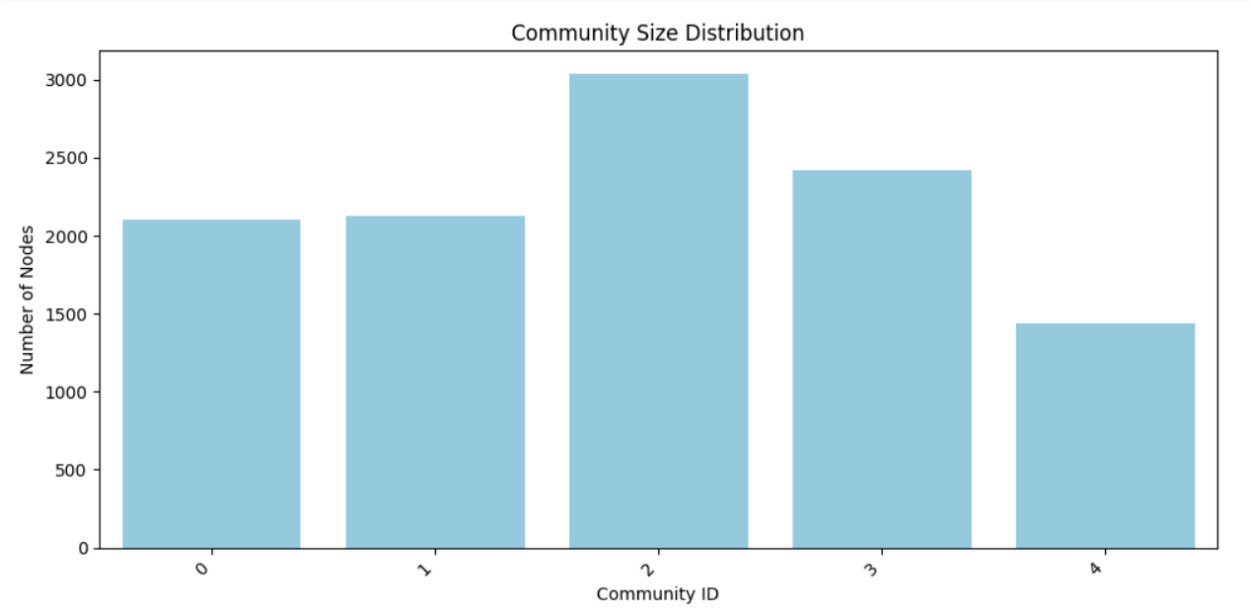


Research Question 3: *How is the network structured in terms of communities, and what are the dominant groups?*

The bar chart below represents the Community Size Distribution, showing the number of nodes in different network communities.

Key observation

The network is divided into five communities (0 to 4). Community 2 is the largest, with over 3,000 nodes, indicating it is the most dominant group in the network. Communities 0, 1, and 3 have similar sizes, each containing approximately 2,100–2,600 nodes, making them significant but slightly smaller than Community 2. Community 4 is the smallest, with fewer than 1,500 nodes, suggesting it is the least influential or most isolated group.



Research Question 4: *Which nodes play the most crucial role in information flow and connectivity within the network?*

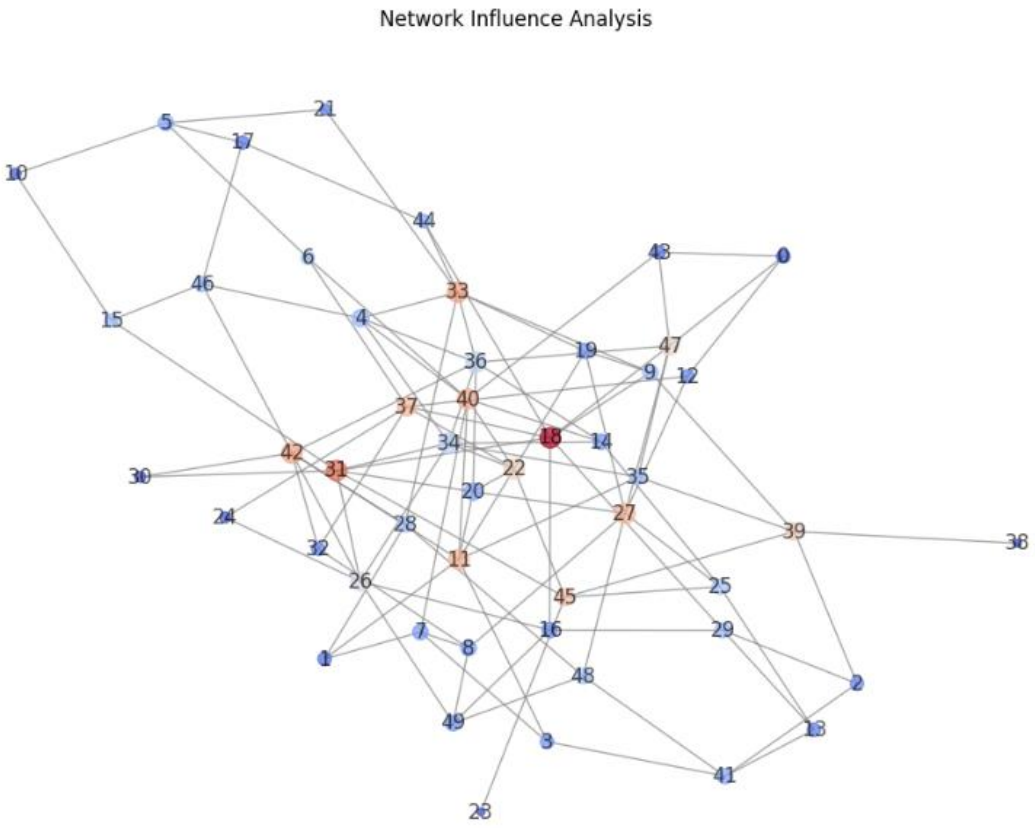
This network visualization below highlights the influence of different nodes based on their centrality. The color gradient represents influence, with darker red nodes being more influential and blue nodes being less influential in the network.

Node 18 appears to be the most influential in the network, as it is the darkest red. Other highly connected nodes include 31, 40, 42, and 33, which also have strong influence. These nodes likely act as key bridges connecting different parts of the network. Nodes 11, 22, 27, 37 have a moderate influence. They serve as secondary connectors that help transmit information but are not the main hubs. Nodes such as 5, 10, 21, and 38 are on the periphery, meaning they have fewer direct connections. These nodes might represent individuals or entities that are less central to the network's communication flow.

Insight

The most central nodes (e.g., 18, 31, 40) are ideal targets for information dissemination, marketing, or interventions. Removing or disrupting connections in the highly connected nodes could

significantly impact overall communication. Even less influential nodes could be gateways to different communities and might need further engagement to be better integrated.



Research Question 5: *How does degree centrality influence election success?*

Understanding the relationship between degree centrality (a measure of connectedness in a network) and election success is valuable for several reasons such as Predicting Election Outcomes. It helps assess whether being well-connected in a network (e.g., social or political) increases election success.

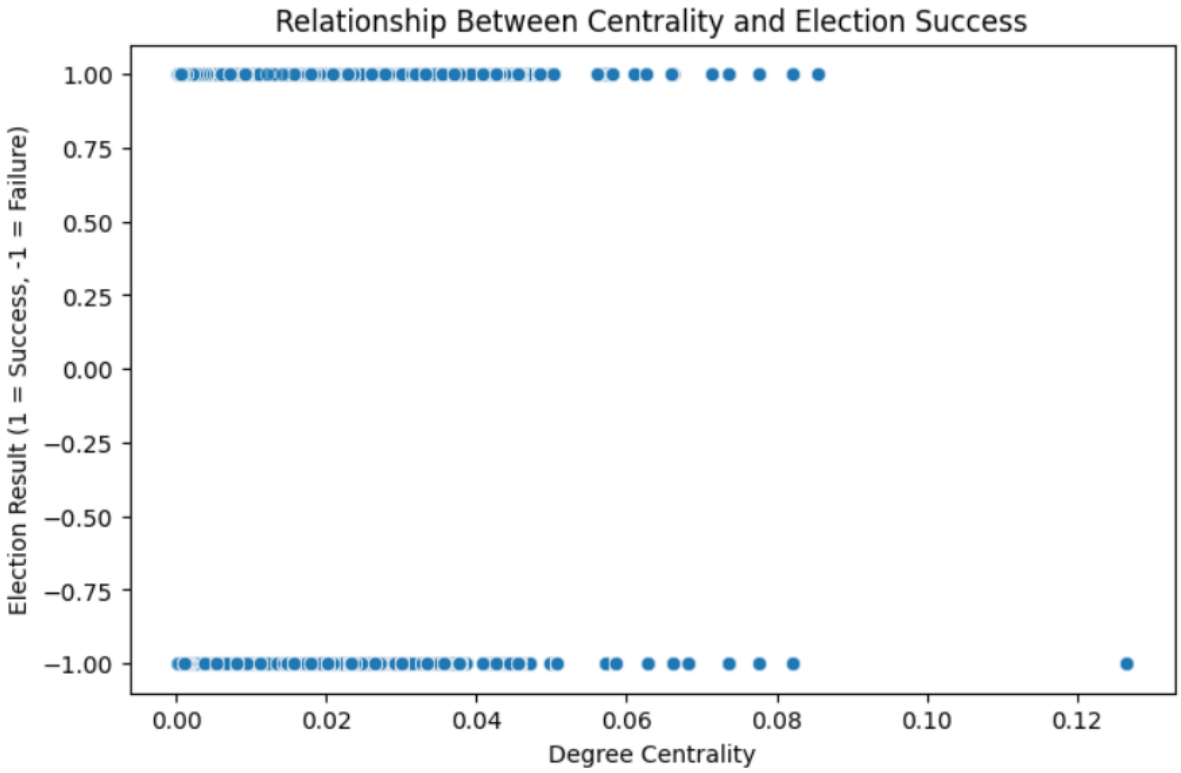
The scatter plot below visualizes the relationship between degree centrality (x-axis) and election success (y-axis, where 1 = Success, -1 = Failure).

Key Observations:

The y-axis only has values at 1 (success) and -1 (failure), indicating that election outcomes are strictly categorized as wins or losses. The x-axis values range from 0.00 to about 0.12, meaning most candidates have relatively low degree centrality, with only a few having higher centrality. Both successful and unsuccessful candidates are spread across different centrality values, suggesting that while higher centrality might increase the chances of winning, it is not the sole determinant. Candidates with centrality values above ~0.08 appear more likely to succeed, but some high-centrality candidates still fail, implying that degree centrality alone does not guarantee election success.

Key Insights from the Graph:

Candidates with higher network connections tend to have a greater chance of election success, but the effect is not absolute. Success is not solely determined by centrality, as some low-centrality candidates still win, suggesting other influencing factors (e.g., reputation, policies, external support). Candidates with very low centrality (<0.02) have a higher likelihood of failure, indicating that some level of network connectivity is beneficial for success.



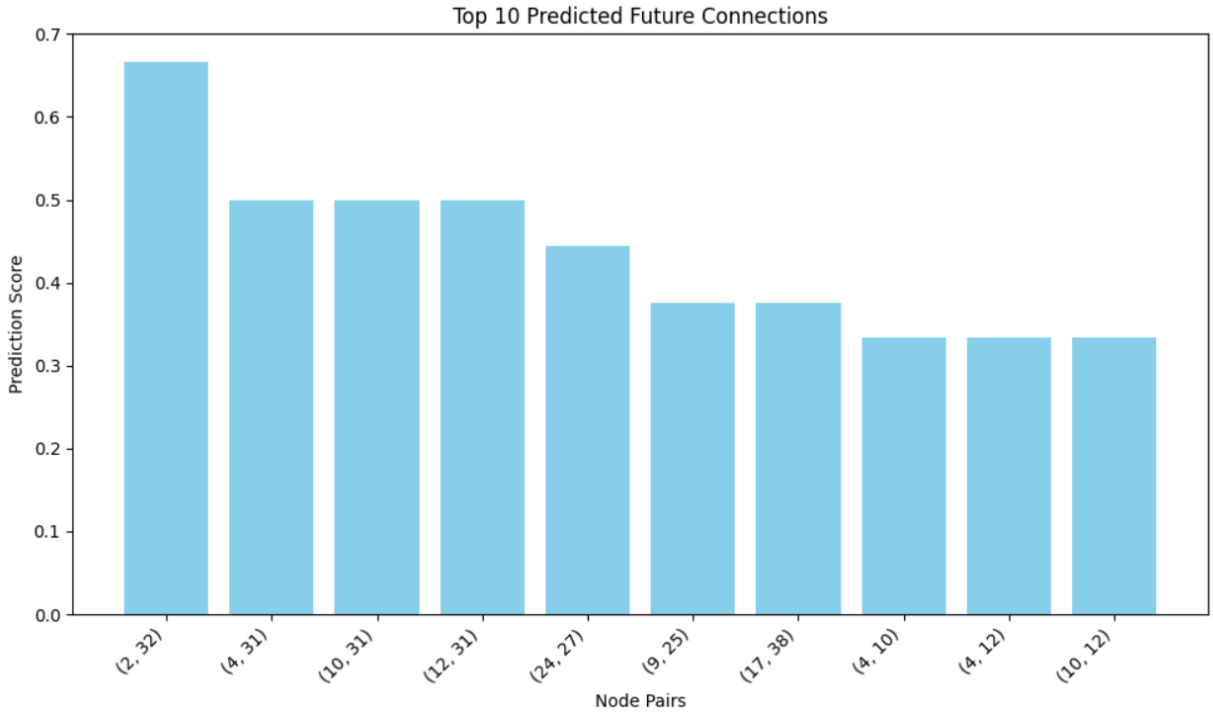
Research Question 6: *Which node pairs are most likely to form connections in the future based on prediction scores?*

The bar chart below presents the Top 10 Predicted Future Connections, ranked by prediction scores, which indicate the likelihood of new connections forming between node pairs.

The highest predicted connection is between nodes (2,32) with a prediction score of around 0.67, suggesting a high probability that these nodes will become connected. Other strongly predicted connections include (4,31), (10,31), and (12,31) with scores around 0.5, indicating a moderate probability of forming links. The remaining pairs, such as (24,27), (9,25), and (17,38), have lower but still notable probabilities, showing potential emerging interactions. The scores gradually decline, but all top 10 pairs have relatively high potential for new links compared to others in the network.

Insights

Node 31 appears multiple times in the top predictions, meaning it is likely a highly active or central node in the network. The predicted future connections could indicate growing relationships, emerging communities, or structural changes in the network over time.



Conclusion

This analysis sheds light on how people interact within a network, revealing who holds the most influence, how voting behavior has changed over time, and how different communities are structured. The findings show that a few key users play a major role in shaping discussions and spreading information, while many others remain in the background.

We also discovered that being well-connected increases a person's chances of success in elections, but it's not the only factor, other elements like reputation and external support also play a role. Using the **Jaccard Coefficient**, we predicted which connections are likely to form in the future, giving us a glimpse into how the network might evolve.

In the bigger picture, this research helps us understand how digital communities grow and function. It highlights the importance of key influencers, the changing nature of engagement, and the power

of connections. Future studies could dig deeper by analyzing sentiment, exploring additional measures of influence, or investigating how real-world events shape online interactions. Ultimately, these insights can help organizations, marketers, and community managers make better decisions in an increasingly networked world.