Project 2

Nishit Soni

April 2024

1 Introduction

This project is structured around a comprehensive analysis of network data through a series of interconnected questions like identifying the leader, predicting missing links and community detection.

1.1 Objectives

The project is divided into three main parts:

1. **Identifying the Leader:** In this section of the project, we use the PageRank algorithm to identify the most influential people in the network. PageRank examines how individuals are linked by their relationships. Each time one person endorses another, the influence of the person being endorsed increases. It's similar to how in everyday life, a person's reputation can grow when they are acknowledged by well-known or respected people. This technique helps us figure out who are the key influencers within the network.

2. **Predicting Missing Links:** Explore future connections using matrix-based methods. This involves analyzing the existing network structure to predict whether two unconnected nodes might have a significant likelihood of forming a beneficial relationship.

3. **Community Detection:** In this part of our project, we aim to develop a method that groups individuals in the network into communities. Each group will be made up of people who are closely connected to each other, but not as much to people outside their group. This approach will help us see how the network breaks down into smaller circles or communities based on who interacts with whom. This is similar to how in real life, people form groups of friends or colleagues who are closely connected.

1.2 Methodology

Each part of the project will apply specific network analysis techniques:

- PageRank will be implemented to score each node's influence based on its connectivity.
- Missing links will be predicted through the implementation of a matrix completion technique, assessing potential connections that could logically exist based on observed network patterns.
- For community detection, the approach will involve defining an optimization problem that seeks to maximize internal node similarity within each community while minimizing similarity between communities.

2 Network Influence Analysis Using PageRank

Choose the top leader by running a random walk on the graph with teleportation.
To determine the most influential individual in a network, we utilize the PageRank algorithm. In this section, we apply the PageRank algorithm to solve this problem. Initially, we create the directed graph from the dataset by iterating through rows to build edges. After constructing the directed graph, we implement the random walk algorithm. Here, we start from a randomly selected node and traverse the graph, incrementing a counter by 1 each time a node is visited. This count reflects the node's likelihood of being reached by a random walker and thus indicates its level of influence or importance within the network. The node with the highest count after numerous iterations is considered the 'top leader' of the network, highlighting its central role in the connectivity and dynamics of the graph.
Step-by-Step Explanation of the Random Walk Algorithm
1.Data Loading and Graph Construction
Data is loaded from a CSV file and processed to construct a directed graph with nodes representing entities and directed edges indicating relationships. This setup facilitates the subsequent analysis steps.

2.Graph Visualization
The graph is visualized using NetworkX's spring layout, which helps in understanding the complex network structure.

3.Random Walk Execution
The random walk is initialized at a randomly selected node. We perform numerous iterations where at each step, the walker either follows an outgoing edge or jumps to a random node, tallying visits to quantify node centrality.

4.Results Compilation and Discussion
After completing the random walk, nodes are ranked based on visit counts. This ranking identifies the most central nodes, indicative of their influence or importance within the network. The top-ranked node emerges as the potential leader, demonstrating the highest connectivity or influence based on the walk

data.

5.Conclusion
The random walk provides a quantitative measure of influence within the network, highlighting key nodes and potential areas of interest for further analysis or intervention.

Visualization of the Directed Graph: The directed graph constructed from the dataset is illustrated in Figure 1. This graphical representation helps visualize the relationships and interactions between the nodes.

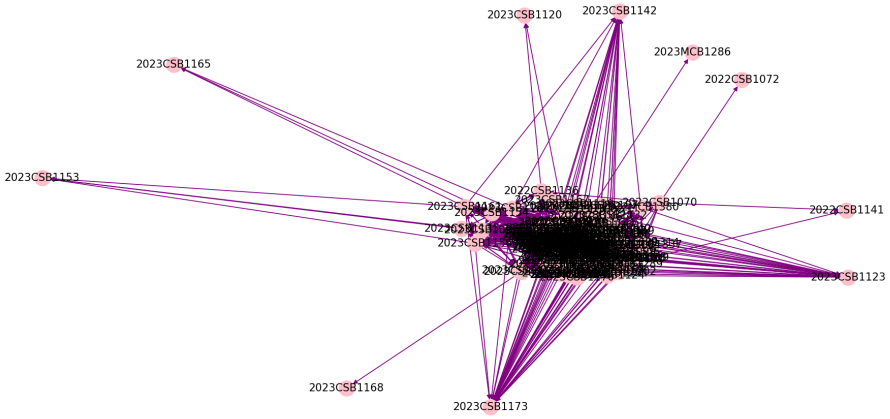

Figure 1: Directed Graph

Result Output: Figure 2 shows the output result listing the first 50 influential individuals in the network.

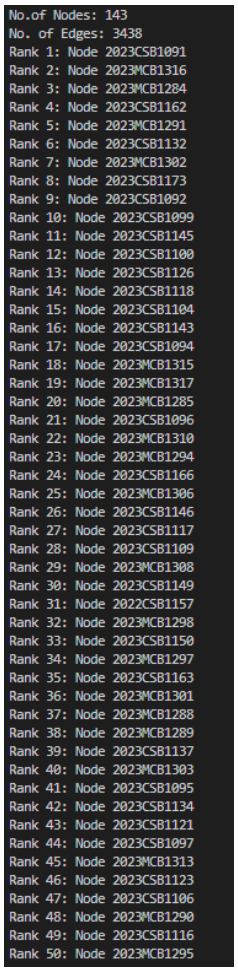

Figure 2: Output Result

In this analysis, the output result showcases the first 50 influential individuals within the network, providing insights into the nodes that wield significant influence in their respective communities.

3 Predicting Missing Links in Networks Using Matrix Methods

Recommend missing links using the matrix method.
To identify missing connections in the network, we employ matrix method,based on the principles of linear algebra. The main concept behind this approach is that the presence or absence of links between nodes in a network can be

represented as a matrix. In this adjacency matrix, nodes are represented by rows and columns, with the presence of a link between two nodes marked by a '1' and its absence by a '0'.

Mathematical Foundation: The method hinges on the idea that it is possible to express any row of the adjacency matrix as a linear combination of the other rows. Mathematically, this is stated as:

$$A = \lambda V + \mu W + \cdots$$

where A is the row vector representing connections of a particular node, and V, W, etc., are other row vectors in the matrix. The coefficients λ, μ, and others are scalar values determined through linear regression or other fitting techniques.

Implementation: To predict a missing link, specifically when a matrix entry A_{ij} is zero (indicating no direct link between node i and node j), we examine whether this zero can be predicted as a non-zero value by analyzing the linear combinations of other rows:

- Construct a new matrix by omitting the row and column corresponding to the nodes in question.
- Apply linear regression to estimate the missing entry from the other interactions in the network.
- If the predicted value exceeds a certain threshold, a potential link is suggested between these nodes.

Example Application: Consider a network where nodes represent individuals and links represent communications. If two individuals do not currently communicate, but the matrix method predicts a high value for their interaction based on the communications patterns of others in their respective groups, it suggests a potential missing link.

Conclusion: This matrix method provides a powerful tool for network analysis, offering a systematic approach to uncovering hidden patterns and potential connections. By leveraging the inter dependencies of node interactions, we can predict and propose new links that enhance the connectivity and functionality of the network.

Visualization of Network Connections

To illustrate the impact of integrating new connections within the network, we present side-by-side images showing the network before and after the addition of new connections. These visualizations help in comparing the network's structure and connectivity enhancements directly.

These figures demonstrate the network's evolution, highlighting how new strategic connections can potentially enhance overall connectivity.

4 Community Detection and Visualization

The detection of communities within the network by Modularity maximization

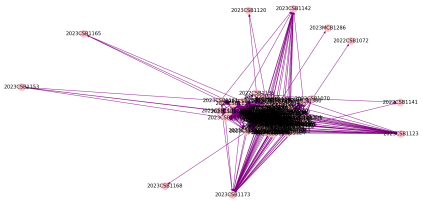

Figure 3: Network Before New Connections

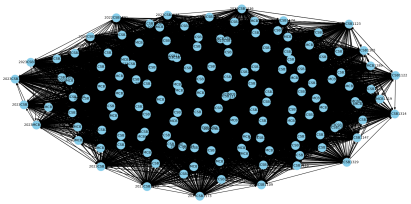

Figure 4: Network After New Connections

It plays a critical role in understanding the structural dynamics and the influence distribution among the nodes. By identifying clusters connected nodes, we can better predicit how information or influence flows within the network.Community detection is performed using the greedy modularity maximization approach provided by NetworkX. The communities are then visualized using different colors for each community.

Modularity

Modularity quantifies the strength of division of a network into modules (also called groups, clusters or communities). A higher modularity means that there are dense connections between the nodes within modules but sparse connections between nodes in different modules.

Community Detection in Network Graphs

This section outlines the implementation of a community detection algorithm applied to a network graph constructed from CSV data.The process involves several key steps, from data ingestion to community analysis.

Data Preparation and Graph Construction

The algorithm begins by loading relational data from `project_data.csv`. Using the `pandas` library, each row is parsed to construct edges in a graph. The source node is identified from the first column, and subsequent columns in the same row specify target nodes, thereby creating a network structure with Python's `networkx` library.

Community Detection Algorithm

Once the network is established, community detection is performed using the `greedy_modularity_communities` method from `networkx`. This method optimizes modularity, enhancing the detection of densely connected subgraphs or communities within the larger network, aiming to maximize intra-community edges and minimize inter-community connections.

Visualization and Analysis

Graph visualization employs a spring layout to emphasize the clustering and community structure, with nodes color-coded by their community affiliation. This visual representation is crucial for intuitively understanding the network's composition. The analysis concludes with a printout of each community's size and composition, providing insights into the network's modular structure.

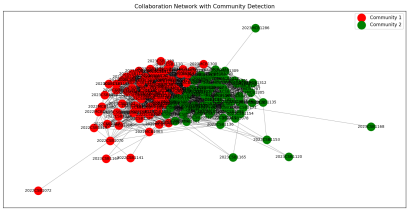

Figure 5: Visualization of the network with communities highlighted

Figure 6: Output Result

This approach not only clarifies the structural dynamics of the network but also aids in identifying key nodes and communities within the complex web of interactions.

Results

The visualization demonstrates clear community structures within the network, with distinct groups of nodes closely interconnected. The detailed community data, including the size and members of each community, were also analyzed and are presented below.

Community Analysis

Each detected community is analyzed for its size and the members it contains, which provides insights into the structure and potential sub-groups within the overall network.

Conclusion

The analysis provides a deep understanding of the community dynamics within the network. This approach is crucial for identifying key clusters and can be instrumental for further studies on network behavior and characteristics.

Sources:

1.Youtube for random walk and greedy modularity maximization approach
2.Google
3.Chatgpt
4.Grammarly