P3. Link Prediction in Letterboxd Ego Network and Movie Genre Preferences

Juan Fernandez Cruz
Data 7A
Universidad Politecnica de Yucatan
Ucu, Yucatan
Email: 2109061@upy.edu.mx

Didier Gamboa
Universidad Politecnica de Yucatan
Ucu, Mexico
Email: didier.gamboa@upy.edu.mx

***Abstract*—This paper presents a comprehensive analysis of a social ego network derived from Letterboxd, a social platform for movie enthusiasts. The study focuses on understanding the network's structure through basic characteristics such as node count, edge count, density, average degree, diameter, and average path length. The network's degree distribution is examined to highlight the skewness typical of social networks, where a few nodes hold a majority of the connections.**

Additionally, the paper explores link prediction techniques using both the Jaccard coefficient and the Adamic-Adar index to identify potential new connections within the network. The analysis demonstrates how these structural measures can suggest likely connections, with specific predictions provided for the user node 'mubius,' a streaming service account. Furthermore, the paper introduces an alternative approach to link prediction that incorporates movie genre preferences, arguing that such content-based predictions may offer more accurate and meaningful results in this context.

The findings of this study have practical implications for platforms like Letterboxd, where understanding and predicting user connections can enhance social interactions, facilitate personalized recommendations, and foster community building based on shared interests. By integrating both structural and preference-based data, the paper proposes a more holistic approach to network analysis that aligns with the goals of content-driven social platforms.

I. Introduction

In recent years, the analysis of social networks has become a powerful tool for understanding complex relational structures in various domains. This study focuses on an ego network derived from Letterboxd, a social platform centered around movie enthusiasts. The objective of this paper is to explore the characteristics of this network and to employ link prediction techniques to infer potential connections within the network, particularly by leveraging users' shared movie preferences.

A. Letterboxd as a Social Platform

Letterboxd is a popular online community where users can track, rate, and review films. Beyond its core function as a film diary, Letterboxd fosters a rich social environment, allowing users to follow others, comment on reviews, and share their cinematic tastes. The platform's social graph is built on these follower and following relationships, making it an ideal subject for social network analysis.

B. Ego Networks in Social Network Analysis

An ego network refers to the network centered around a single individual, known as the ego, and includes all the nodes directly connected to the ego, known as alters. The analysis of ego networks provides valuable insights into the structure and dynamics of personal relationships within a larger social network. By focusing on the ego and its immediate connections, researchers can study the local structure of the network, examine the degree distribution, assess clustering coefficients, and understand the roles and influence of the ego in the network.

C. Link Prediction in Social Networks

Link prediction is a fundamental task in social network analysis, aiming to predict the likelihood of future connections between nodes based on existing network structures and node attributes. Traditional methods like the Jaccard coefficient, which measures the similarity between nodes based on their shared neighbors, are commonly used for this purpose. However, the prediction accuracy can be enhanced by incorporating additional information, such as node attributes or shared interests, which can provide a more nuanced understanding of the factors driving connection formation.

D. Incorporating Movie Preferences in Link Prediction

In this study, we propose a novel approach to link prediction within the Letterboxd ego network by leveraging the movie preferences of users. By analyzing the genres of films liked by each user, we aim to predict potential links not only based on traditional network measures but also on the similarity of movie tastes. This approach acknowledges that shared interests are a significant factor in the formation of social ties and can lead to more accurate predictions of future connections.

This introduction sets the stage for a detailed analysis of the Letterboxd ego network, with a particular focus on how movie preferences can be used to enhance link prediction within the network.

II. Network Visualization

A. Visualization Overview

To visually represent the ego network from Letterboxd, we utilized the `networkx` library in Python, which provides

powerful tools for network analysis and visualization. Due to the complexity and the number of nodes involved, we decided to generate two types of visualizations: one with labels and one without labels.

B. Label and No-Label Representations

When dealing with a large number of nodes, visual clutter becomes a significant issue, especially when node labels are included. To address this, we created both a labeled and a no-label representation of the network. The labeled version offers detailed insight into individual nodes, while the no-label version provides a clearer overview of the network's structure without the distraction of overlapping text.

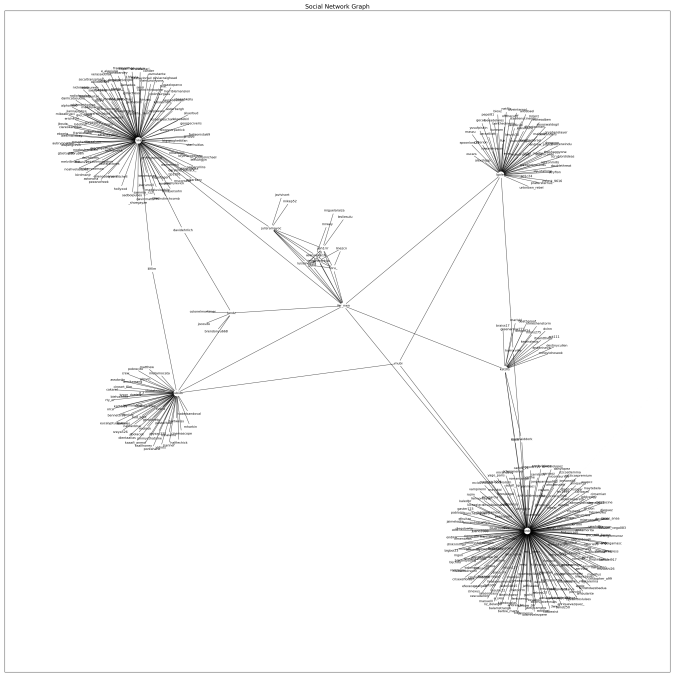

Fig. 1. Ego Network Visualization with Labels

C. Visualization of Depth-1 Network

In addition to the full ego network, we also visualized the network at a depth of 1, which includes only the ego and its immediate connections (alters). This more focused visualization will be crucial later in our analysis when we perform link prediction based on shared movie preferences. As these networks are smaller and less complex, the visualizations are straightforward but essential for understanding the structure relevant to the link prediction task.

These visualizations set the stage for the subsequent analysis, where we will explore the network's characteristics and apply link prediction techniques based on both structural properties and shared movie preferences.

III. Mapping Process

The mapping process of the Letterboxd ego network involved several key steps, including data collection, data preprocessing, and network construction. This section outlines the methodology used to transform raw data into a structured network suitable for analysis.

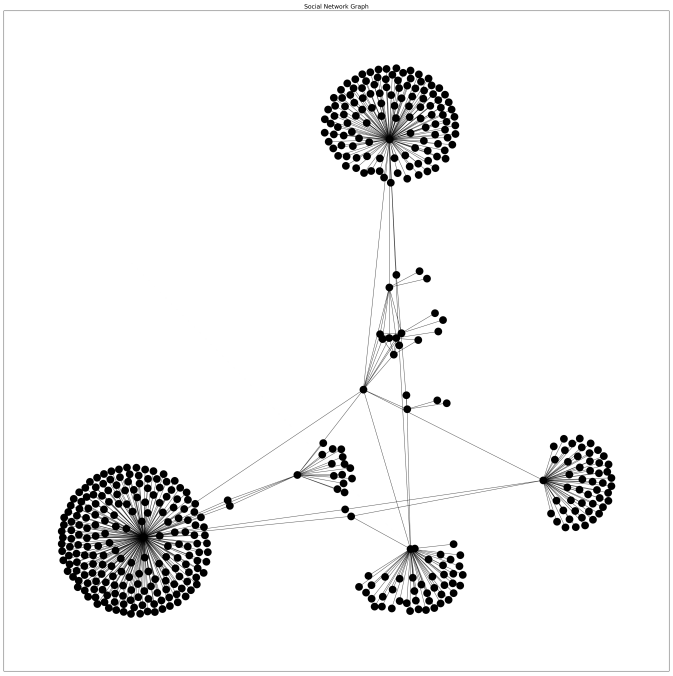

Fig. 2. Ego Network Visualization without Labels

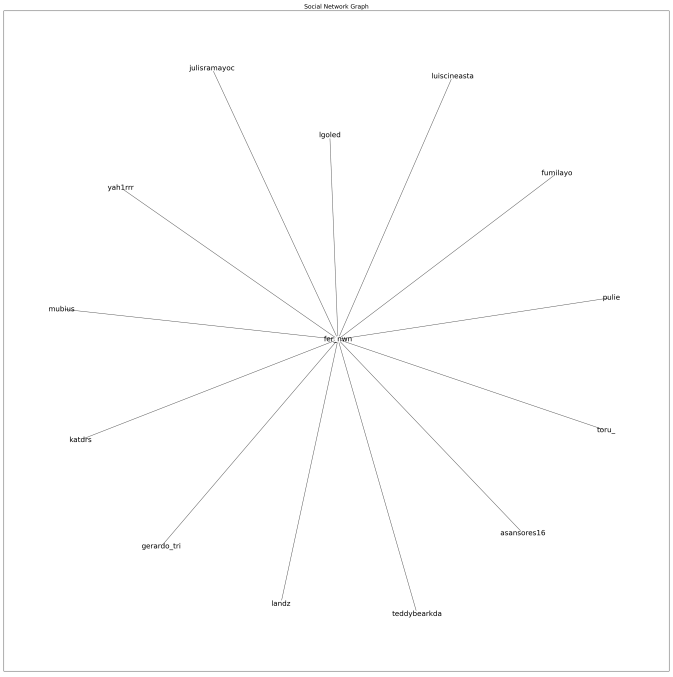

Fig. 3. Depth-1 Ego Network Visualization

A. Data Collection

The data for this study was collected using a custom-built Python scraper, designed to interact with the Letterboxd

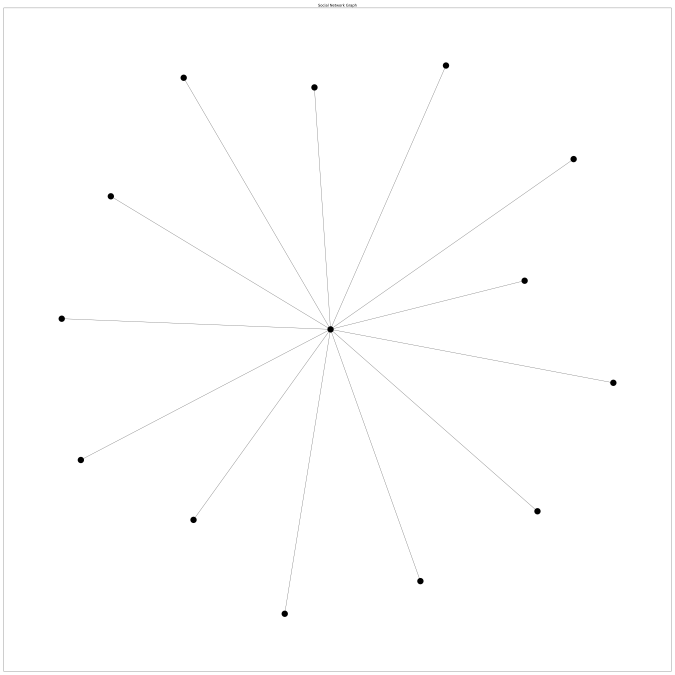

Fig. 4. Depth-1 Ego Network Visualization without Labels

website. The scraping process was divided into two main phases:

1) Basic Scraping: In the initial phase, the scraper requests the desired ego page on the Letterboxd website and navigates to the user's network, capturing both followers and followed users. The scraper then iteratively processes each follower and followed user to gather data for the level 2 depth ego network. This process was implemented using Python's `requests` library in conjunction with regular expressions to efficiently extract the necessary data.

2) Enhanced Scraping with Movie Preferences: In the second phase, the scraper was modified to include a function for gathering movie preferences. This enhanced scraper first collects a list of users from the level 1 network and then retrieves all the movies each user has liked. The scraper subsequently accesses the genre information for each liked movie and saves this data in a dictionary structure, where the key is the user, and the value is another dictionary containing the genre as the key and the count of occurrences as the value. This structured data allows for an in-depth analysis of genre preferences across the network.

B. Data Preprocessing

Before constructing the network, the raw data underwent preprocessing to ensure accuracy and consistency. This step involved several key tasks:

- **Cleaning and Filtering:** Duplicate entries and incomplete data were removed. Only users with complete genre preference data were retained, as this information is crucial for the link prediction and clustering analysis.
- **Normalization:** Movie genres were normalized to account for variations in genre naming conventions (e.g., "Sci-Fi" vs. "Science Fiction") to ensure consistency across the dataset.
- **Data Transformation:** Relationships between users were transformed into a graph structure, where nodes represent users, and edges represent follower/following relationships. The movie genre preferences were stored as node attributes for use in subsequent analyses.

C. Network Construction

The final step in the mapping process was the construction of the ego network using the preprocessed data. The network was built using the `networkx` library, allowing for the efficient creation of the graph structure and the assignment of node and edge attributes. The constructed network includes the ego, all its direct connections (alters), and the edges representing the relationships between these nodes.

D. Principal Component Analysis (PCA) of Movie Preferences

To gain insights into the movie preferences within the ego network, we applied Principal Component Analysis (PCA) to the genre data collected for each user. Using the `sklearn` library, we performed PCA to reduce the dimensionality of the genre data and visualize potential clusters related by the similarity of liked movies. This analysis provides a visual representation of how users in the network are grouped based on their genre preferences.

Fig. 5. PCA Visualization of Genre-Based Clusters

1) Correlation Matrix of Movie Genres: To further explore the relationships between different movie genres, we constructed a correlation matrix of the genres. This matrix highlights which genres tend to be liked together by the users in the network, providing insights into common patterns of genre preferences.

2) Heatmap of Genre Preferences: We also generated a heatmap to visualize the distribution of genre preferences among the users in the ego network. This heatmap allows us to observe the variability in movie tastes across different users,

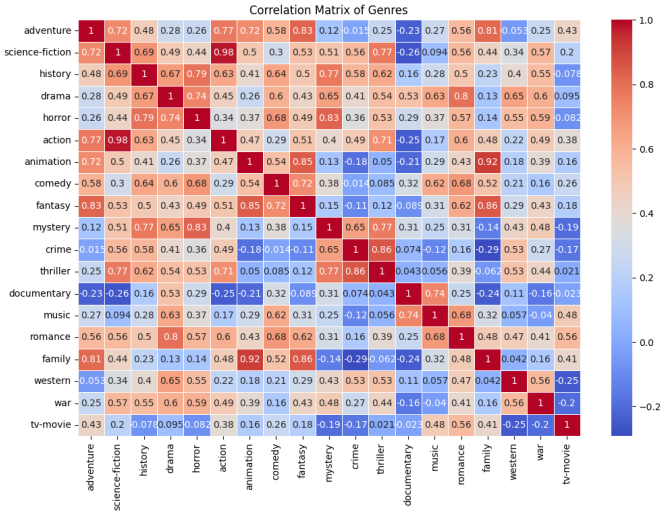

Fig. 6. Correlation Matrix of Movie Genres

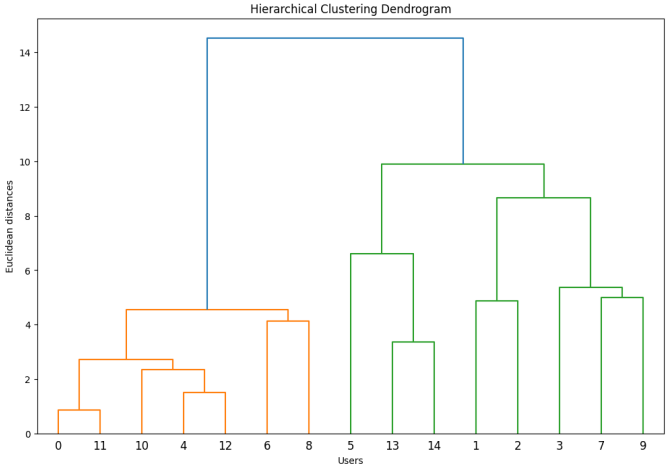

Fig. 8. Hierarchical Clustering Dendrogram of Users Based on Genre Preferences

highlighting the genres that are more or less popular within the network.

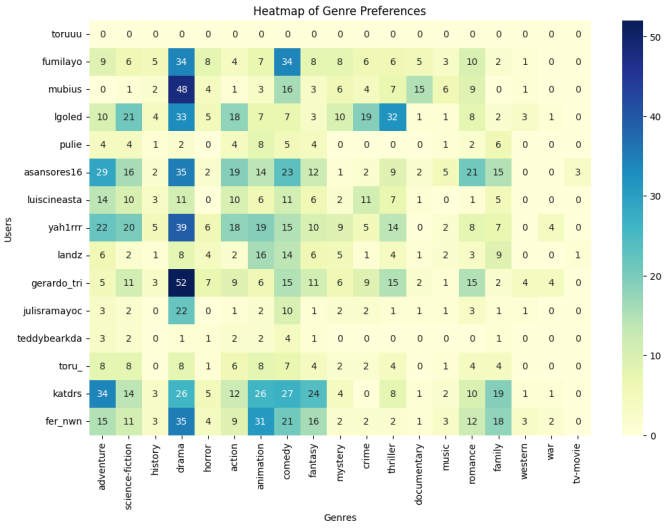

Fig. 7. Heatmap of User Genre Preferences

3) Hierarchical Clustering of Users Based on Genre Preferences: Finally, we applied hierarchical clustering to the genre preference data using Euclidean distances to assess the similarity between users. A dendrogram was generated to visualize the hierarchical relationships and potential clusters of users with similar movie tastes.

The mapping process, including the collection, preprocessing, and analysis of genre data, forms the foundation for the link prediction analysis that follows, leveraging both network structure and shared movie preferences.

IV. Characteristics of the Network

In this section, we analyze the basic characteristics of the Letterboxd ego network under study. These metrics provide a foundational understanding of the network's structure and dynamics.

A. Basic Characteristics

The basic characteristics of the network are as follows:

- **Number of Nodes:** 459
- **Number of Edges:** 974
- **Density:** 0.0046
- **Average Degree:** 4.24
- **Diameter:** 4
- **Average Path Length:** 3.30

1) Number of Nodes and Edges: The network consists of 459 nodes and 974 edges. The number of nodes represents the total number of users (the ego and its direct connections), while the number of edges indicates the relationships (following or follower connections) between these users. The network's relatively low number of edges compared to the number of nodes suggests a sparse network structure.

2) Density: The network density is calculated as 0.0046. Density measures the proportion of potential connections in the network that are actual connections. A density value close to zero, as seen here, typically indicates a sparse network where most nodes are not directly connected.

3) Average Degree: The average degree of the network is 4.24, which suggests that on average, each user is connected to a little over four other users. This value reflects the overall connectivity within the network but should be interpreted alongside other metrics such as the degree distribution for a more detailed understanding.

4) Diameter: The diameter of the network, which is the longest shortest path between any two nodes, is 4. This indicates that the maximum distance (in terms of connections) between any two users in the network is relatively small, suggesting a compact network where even distant nodes are relatively close.

5) Average Path Length: The average path length of the network is 3.30, which represents the average number of steps required to reach one user from another within the network. This measure, along with the diameter, indicates that the network is fairly navigable, with users being only a few connections away from each other on average.

B. Degree Distribution

The degree distribution is a fundamental metric in network analysis, showing how connections are distributed across the network. In this case, we analyzed both in-degree and out-degree distributions.

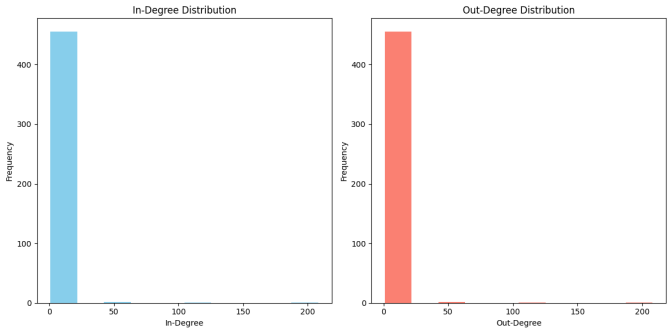

Fig. 9. In-Degree and Out-Degree Distribution of the Network

The degree distribution plots (Figure 9) show a highly skewed distribution, where most nodes have a very low degree, while a few nodes have a significantly higher degree. This type of distribution is characteristic of social networks, where a small number of users tend to be highly connected (often referred to as "hubs"), while the majority have only a few connections.

The in-degree distribution reflects the number of followers a user has, and the out-degree distribution shows how many other users a given user follows. The skewness observed here indicates that most users in this network have a low number of followers and follow only a few others, with only a handful of users having a large number of connections.

C. Interpretation of the Degree Distribution

The observed degree distribution is typical for social networks, which often follow a power-law distribution. This means that a few nodes (influential users) have a disproportionately high number of connections, while most nodes have relatively few connections. This structure is often indicative of a small-world phenomenon, where despite the large network size, the average path length remains relatively short, allowing for quick information dissemination and social interaction.

The skewness in the in-degree and out-degree distributions can also imply that certain users are more central or influential in the network, either by being followed by many others (high in-degree) or by following many others (high out-degree). These central users can play crucial roles in the dissemination of information and the overall connectivity of the network.

This analysis of the network's basic characteristics provides a foundation for understanding its structure and the dynamics of user interactions within it. Further analysis, including link prediction based on shared interests, will build on these insights to explore the potential for new connections within the network.

V. Link Prediction

Link prediction is a critical task in network analysis, aiming to identify potential future connections based on existing network structure and node attributes. In this study, we employed two widely-used similarity measures: the Jaccard coefficient and the Adamic-Adar index, to predict new links within the Letterboxd ego network.

A. Jaccard Coefficient and Adamic-Adar Index

The Jaccard coefficient is a measure of similarity between two sets, defined as the size of the intersection divided by the size of the union of the sets. In the context of network analysis, it is used to estimate the probability that two nodes will form a link based on their shared neighbors.

The Adamic-Adar index, on the other hand, is another measure that considers the shared neighbors between two nodes but gives more weight to less connected neighbors. This approach is particularly useful in networks where the presence of common connections is an indicator of a potential link but where highly connected nodes should have less influence.

B. Link Prediction Results for Node 'mubius'

To demonstrate the link prediction process, we applied both the Jaccard coefficient and the Adamic-Adar index to identify the top predicted connections for the node 'mubius,' which represents an account of a streaming service. The results for the top 10 Jaccard-based predictions are as follows:

- **(mubius, brendonyu668)** $\rightarrow$ 0.0208
- **(mubius, colonelmortimer)** $\rightarrow$ 0.0208
- **(mubius, jocosito)** $\rightarrow$ 0.0208
- **(davidehrlich, mubius)** $\rightarrow$ 0.0204
- **(lgoled, mubius)** $\rightarrow$ 0.0202
- **(luiscineasta, mubius)** $\rightarrow$ 0.0189
- **(mubius, toru_)** $\rightarrow$ 0.0189
- **(mubius, pulie)** $\rightarrow$ 0.0189
- **(teddybearkda, mubius)** $\rightarrow$ 0.0185
- **(asansores16, mubius)** $\rightarrow$ 0.0182

C. Comparison with Preference-Based Link Prediction

While structural measures like the Jaccard coefficient and Adamic-Adar index provide insights into potential connections based on network topology, they do not consider the content or context of interactions, such as shared interests or preferences. In the context of the Letterboxd network, movie genre preferences play a significant role in forming connections, as users are likely to connect with others who share similar tastes.

By incorporating movie preferences into the link prediction model, we can potentially achieve more accurate predictions, particularly in networks where shared interests are a strong

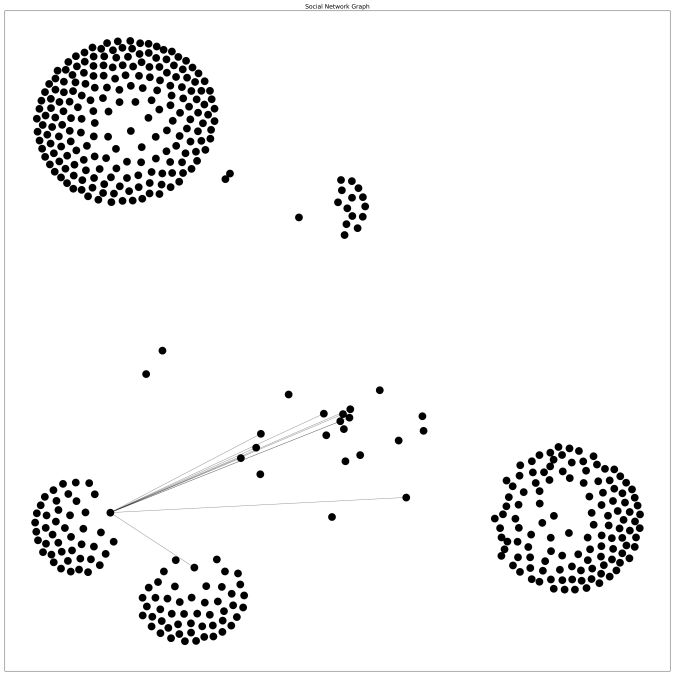

Fig. 10. Predicted Links for Node 'mubius' Based on Jaccard Coefficient

driver of social connections. For example, predicting links based on the similarity of liked movie genres might reveal connections that are more meaningful or likely to occur in practice, compared to purely structural predictions.

D. Application and Utility for the Enterprise

For an enterprise like the streaming service represented by the 'mubius' account, utilizing both structural and preference-based link prediction can provide a competitive advantage. Structural predictions help identify users who are structurally close in the network and may be influenced to connect, enhancing the platform's social dynamics.

On the other hand, preference-based predictions allow the enterprise to target users with similar tastes, facilitating personalized recommendations, targeted marketing, and the fostering of communities centered around shared interests. This approach not only improves user engagement but also enhances the quality of connections formed on the platform, leading to a more cohesive and active user base.