Social Circles: Community Analysis and Link Prediction Using Facebook100

By Khushi Patel and Narasimha Rohit Katta

CHAPTER 1: INTRODUCTION

1.1 Problem Statement

Social networks play a huge role in how we connect with people today. Understanding how connections are formed and predicting new ones can help in many ways, from recommending friends to suggesting content. This project focuses on predicting potential connections (links) in social networks by analyzing the Facebook100 dataset, which represents social interactions at a college level. We will use graph-based features and machine learning to predict links, allowing us to uncover hidden patterns in these networks.

1.2 Overview of the Data

The dataset used for this project is “Facebook100”, provided by the Stanford Network Analysis Project (SNAP). It represents the social networks of college students, where users are nodes, and friendships are edges connecting these nodes.

Here are some basic stats about the dataset:

- Number of users (nodes): 4,039
- Number of friendships (edges): 88,234
- Average number of friends per user: 43.69

1.3 Your Contribution:

Khushi Patel:

- Khushi handled Stage 1 of the project, which included:

- Importing libraries and loading the dataset.

- Check the data quality to ensure everything is clean and consistent.

- Normalizing features such as degree centrality and clustering coefficients.

- Visualizing the graph to better understand the structure and communities.

- Saving and downloading the processed graph to keep everything organized for future steps.

Narasimha Rohit Katta:

- Rohit worked on Stage 2, which involved:

- Cleaning the graph by removing self-loops, duplicate edges, and isolated nodes to make the graph more usable.

- Checking if the graph is connected and focusing on the largest connected component.

- Reloading the cleaned graph for continuity.

- Counting bridge edges (edges that connect different parts of the graph) and non-bridge edges, which helped us understand the critical points in the network.

CHAPTER 2: RELATED WORK

- Link prediction has been widely explored in network science. Many studies have used heuristics such as “Common Neighbors,” “Jaccard Coefficient,” and “Preferential Attachment” to predict new links in social networks.

- Recently, machine learning approaches have shown promising results, especially those that use graph neural networks (GNNs).
- Our approach blends traditional methods with machine learning models to predict links while keeping things computationally efficient.

CHAPTER 3: PRELIMINARY METHODOLOGY

Stage 1: Initial Dataset Preparation (Khushi Patel)

1. Import Libraries: We started by importing necessary libraries like “pandas,” “network,” and “matplotlib.”

2. Load Dataset: The dataset is an undirected graph using networkx.

3. Check Data Quality: We checked the dataset for missing values and inconsistencies to ensure its quality.

4. Normalize Features: We normalized key graph features like degree centrality and clustering coefficients.

5. Visualize Graph: We visualized the network to better understand the structure, relationships, and potential communities.

6. Save Processed Graph: The cleaned and processed graph was saved for further use in the project.

Stage 2: Graph Cleaning and Connectivity Analysis (Narasimha Rohit Katta)

1. Clean the Graph: We removed irrelevant elements like self-loops, duplicate edges, and isolated nodes that wouldn’t add value to the analysis.

2. Graph Connectivity: We checked the graph’s connectivity to make sure the graph was well connected. If not, we focused on the largest connected component.

3. Reload Cleaned Graph: Once the graph was cleaned, it was reloaded for the next steps.

4. Count Bridge Edges: We counted how many bridge edges (edges that are crucial to maintaining the network's connectivity) and non-bridge edges were in the graph. This helped us understand which parts of the graph were most critical to the overall structure.

CHAPTER 4: PRELIMINARY IMPLEMENTATION

4.1 Data Preprocessing

This phase was all about cleaning the data. We made sure the graph was free of unnecessary elements, normalized the key features, and visualized the network to spot patterns. These steps ensured that we had a high-quality dataset to work within the next stages of the project.

4.2 Preliminary Model Training and Validation

- We selected graph-based features (like common neighbors, Jaccard coefficient, and preferential attachment) to use for link prediction.

- Machine learning models such as "logistic regression" and "random forest" were chosen to make predictions about potential links.

- We decided to evaluate the performance of these models using metrics like "AUC-ROC" and "F1 score" to see how well they predict new connections.

CHAPTER 5 PRELIMINARY EXPERIMENTAL RESULTS AND DISCUSSION

5.1 Preliminary Results

The XGBoost model was employed for link prediction on the Facebook100 dataset, and the preliminary results are as follows:

- AUC-ROC Score: 0.9957

- F1 Score: 0.9690

These metrics indicate strong predictive performance, with the model effectively distinguishing between links and non-links in the network.

5.2 Classification Report

The classification report provides a detailed breakdown of the model's performance for each class:

- Precision: The precision for non-links (0) is 0.97, and for links (1) is 0.96, indicating that the model minimizes false positives.

- Recall: Both classes achieved a recall of 0.97, showcasing the model’s ability to identify most actual links correctly.

- Accuracy: The overall accuracy is 97%, indicating a well-trained model.

- Macro Average and Weighted Average: Both are 0.97, highlighting consistency across the dataset.

5.3 Feature Importance

XGBoost’s feature importance analysis reveals the relative contribution of different features:

1. Adamic-Adar (Importance: 0.927862): This metric is the most influential feature, accounting for over 92% of the total importance in the model.

2. Common Neighbors (Importance: 0.027922): This feature plays a supportive role in prediction but has less significance than Adamic-Adar.

3. Jaccard Coefficient (Importance: 0.027589): Contributes marginally to the model’s predictive power.

4. Preferential Attachment (Importance: 0.016628): Has the least influence but still adds value by accounting for degree-based probabilities of link formation.

5.4 Observations

1. Strong Model Performance: The high AUC-ROC and F1 scores affirm the reliability of the XGBoost model for the link prediction task.

2. Dominance of Adamic-Adar: The overwhelming importance of Adamic-Adar suggests that local graph properties strongly drive link formation in social networks.

3. Balanced Predictions: The balanced precision and recall values ensure that the model avoids overfitting to either links or non-links.

4. Scope for Refinement: While the results are promising, further tuning and testing with additional features may improve performance further

CHAPTER 6: CONCLUSION

At this point in the project, we've completed the initial steps of cleaning the dataset, visualizing the graph, and performing connectivity analysis. These early stages have helped us understand the network better and set up a solid foundation for link prediction.

Next steps include:

1. Feature engineering to create useful attributes for link prediction.

2. Training machine learning models and testing them.

3. Evaluating the models using metrics like AUC-ROC and F1 score to assess how well they predict new links.

We're excited to move forward and apply machine learning to uncover hidden connections in the network.

REFERENCES:

1. Leskovec, J., & Krevl, A. (2014). Facebook100 Dataset. Retrieved from [SNAP](http://snap.stanford.edu/data/ego-Facebook.html).

2. Newman, M. E. J. (2003). The structure and function of complex networks. ‘SIAM Review’, 45(2), 167–256.

3. Liben-Nowell, D., & Kleinberg, J. (2007). The link-prediction problem for social networks. ‘Journal of the American Society for Information Science and Technology’, 58(7), 1019–1031.