Social Circles: Community Analysis and Link Prediction Using Facebook100

Khushi Patel, Narasimha Rohit Katta

1. Project Title

Social Circles: Community Analysis and Link Prediction Using Facebook100

2. Project Members

Khushi Patel, Narasimha Rohit Katta

3. Project Dataset: Basic Dataset Statistics

The dataset used in this project is the Facebook100 dataset, sourced from the Stanford Network Analysis Project (SNAP). This dataset is available at: http://snap.stanford.edu/data/ego-Facebook.html. It contains anonymized user interactions and friendships within various college networks, represented as an undirected graph.

Basic Statistics:

- Total nodes (users): 4,039
- Total edges (connections): 88,234
- Average degree (average number of connections per user): 43.69

4. Kind of Method

We will use a supervised link prediction approach. This involves predicting potential friendships (edges) in the network based on existing structural properties.

5. Suggested Approach

1. **Data Preprocessing:** Clean and preprocess the network data, including removing isolated nodes and normalizing features such as degree centrality and clustering coefficients.
2. **Feature Engineering:** Generate features based on graph properties, such as common neighbors, Jaccard coefficient, and preferential attachment.
3. **Model Selection:** Train machine learning models (e.g., logistic regression, random forest) using the engineered features to predict the existence of links.
4. **Evaluation:** Use metrics like AUC-ROC and F1 score to evaluate the link prediction performance.