**Slide 2: What is Link Prediction?**

Link prediction is a process aimed at forecasting new or missing connections between nodes within a network. To illustrate this, let’s refer to the accompanying image where two blue nodes are highlighted. These nodes represent potential links, as they are currently not connected but share connections with other nodes in the network. The idea behind link prediction is to identify these possible relationships by analyzing existing patterns and connections within the network.

**Slide 3: Project Overview**

Our project consists of two main components: Research Objectives and Data Snapshot.

The primary goal of our research is to predict Potential friendships within a Facebook social network by utilizing the Facebook100 dataset. This dataset is part of the Stanford Network Analysis Project (SNAP) and comprises anonymized friendship data from several American universities. We will employ supervised link prediction techniques alongside machine learning models to analyze the likelihood of connections forming between any two nodes in this social network.

**Slide 4: Data Snapshot**

Let’s delve into the dataset we are working with. It includes a total of **4,039 nodes**, which represent unique users, and **88,234 edges**, indicating the friendships among these users. From this data, we can calculate the average degree, which measures the number of connections each user has. In our dataset, the average degree is approximately **43.69**. This figure suggests that the network is quite dense, meaning users have numerous connections with one another. If the average degree were significantly lower, the network would be less interconnected; conversely, a higher average would indicate a potentially complex structure with intricate connections.

**Slide 5: Network Characteristics**

Understanding network characteristics is key to analyzing how networks operate. These characteristics illustrate various properties of the network that describe its overall structure and behavior, including the total number of nodes, edges, density, and more.

We categorize these network characteristics into two main parts: **Network Statistics** and **Data Preparation**.

**Network Statistics** brings focus to three critical components:

1. **Isolated Nodes**: These users have no connections to anyone in the network, which indicates the presence of at least one friendship for every user in our dataset.

2. **Bridge Edges**: When removed, these unique connections would increase the number of disconnected components in the network, effectively dividing it into separate parts.

3. **Non-Bridge Edges**: In contrast, removing these edges would not affect the network's overall connectivity I.

In our dataset, we identified **75 bridge edges** and approximately **88,159 non-bridge edges**, which helps us understand the network’s connectivity robustness.

**Slide 6: Data Preparation**

For preparing our data, we divided it into two segments: one for training and the other for testing. The training set consists of **83,823 edges**, accounting for about **95%** of all connections. This substantial volume helps the model learn how links are formed within the network. The remaining **4,411 edges** serve as the test set, representing **5%** of the data. This smaller segment allows us to evaluate how effectively the model can predict unseen links. Importantly, all **4,039 nodes** in the network are included in the training dataset, ensuring that test edges connect nodes the model has already encountered.

**Slide 7: Feature Engineering**

Next, we discuss **feature engineering**, which is the process of creating valuable information from raw data to enhance machine learning models’ prediction capabilities. In the context of graph or network analysis, the features we derive are primarily based on the connections between nodes. The better our model understands these connections, the more accurately it can predict potential links.

To bolster our model’s ability to predict links, we engineered specific features using graph structure, including:

1. **Common Neighbors**: This feature counts the number of shared connections two nodes have. For example, if two individuals have many mutual friends, they are more likely to connect.

2. **Jaccard Coefficient**: This metric evaluates the similarity of two nodes’ neighborhoods by dividing the number of shared neighbors by the total number of unique neighbors. Essentially, it examines how much the social circles of the two nodes overlap.

3. **Adamic-Adar Index**: This feature places greater emphasis on shared neighbors that have fewer connections. It is particularly useful for identifying subtle relationships, as it highlights smaller, more meaningful connections.

4. **Preferential Attachment**: This feature predicts that nodes with a higher number of connections are more likely to create new links. It is calculated by multiplying the number of connections of both nodes being considered.

By incorporating these features, our model gains insights into the nature of link formations in the network, enhancing its predictive capabilities.