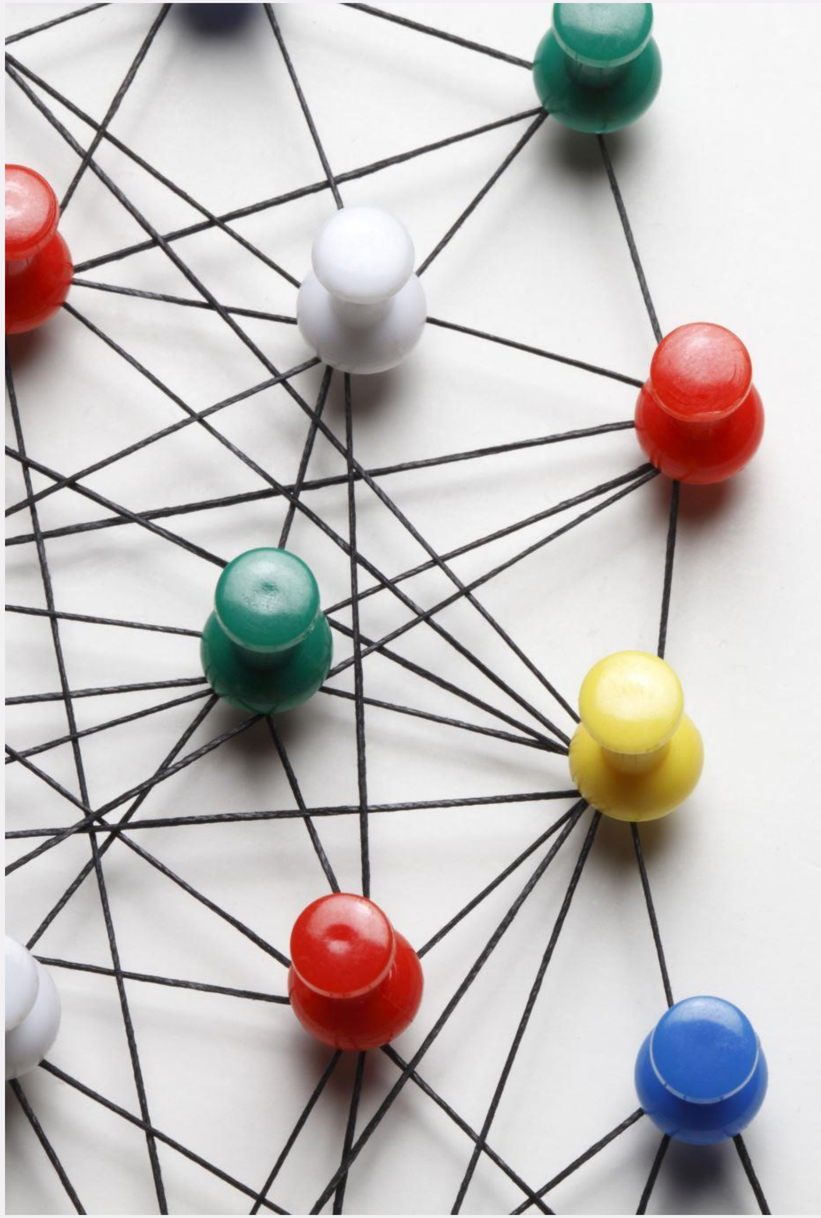
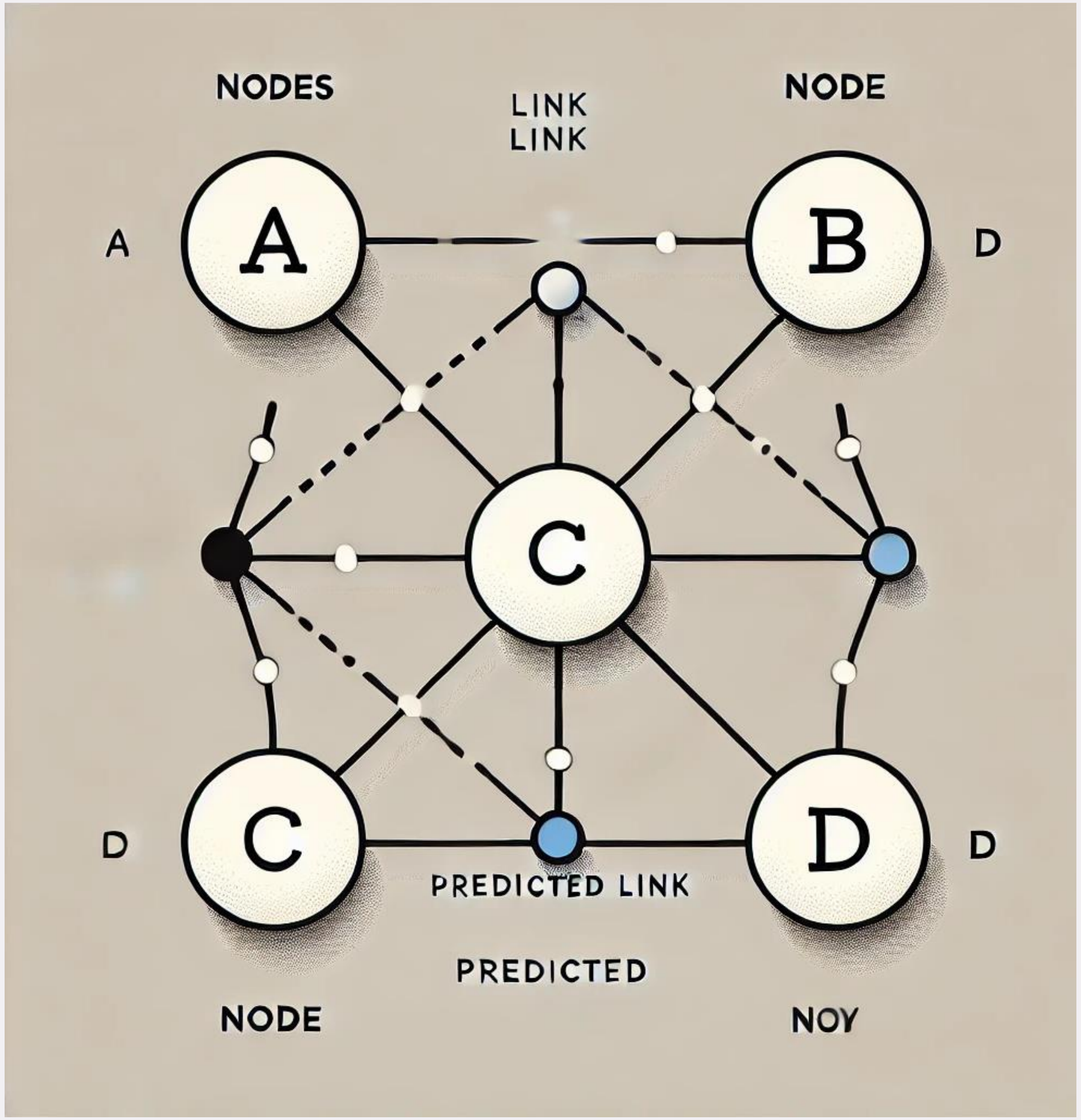


SOCIAL CIRCLES: COMMUNITY ANALYSIS AND LINK PREDICTION USING FACEBOOK100 DATASET

- by Khushi Patel and Narsimha Rohit Katta

What is Link Prediction?

- Link prediction is the process of forecasting (new or missing connections) between nodes in a network.
- It identifies "potential relationships" based on existing patterns in the network.



PROJECT OVERVIEW

1. Research Objective:

• Goal: The primary aim of this project is to predict potential friendships in a social network using the Facebook100 dataset.

• Dataset: The dataset is part of the Stanford Network Analysis Project (SNAP) and comprises anonymized Facebook friendship data from various American universities.

• Approach: A supervised link prediction approach was adopted, leveraging machine learning models to analyze and predict the likelihood of connections between nodes in the network.

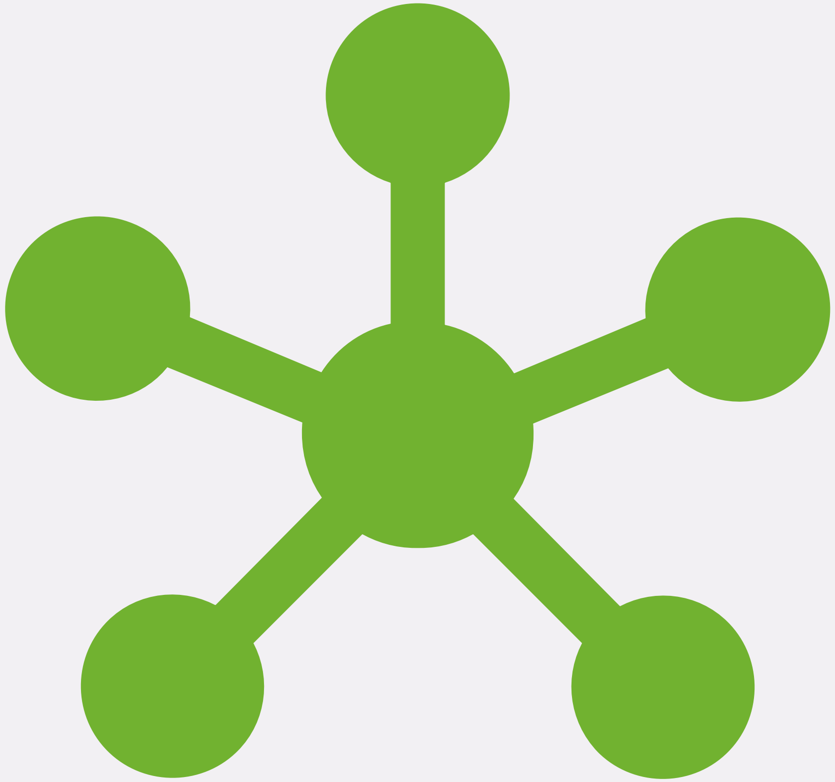
2. Dataset Snapshot:

- Total Nodes: 4,039 users (individuals in the network).
- Total Edges: 88,234 friendships (connections between users).
- Average Degree: 43.69 connections per user, indicating a relatively dense network.

NETWORK CHARACTERISTICS

1. Network Statistics:

- Isolated Nodes: There are no isolated nodes in the dataset, meaning every user is part of at least one friendship.
- Bridge Edges: 75 edges act as bridges, connecting distinct parts of the network. Removing these edges would fragment the graph.
- Non-Bridge Edges: 88,159 edges are not bridges and thus belong to the network's more robust regions.



2. Data Preparation:

- Training Set: 83,823 edges were used for training the model, representing approximately 95% of the total connections.
- Test Set: 4,411 edges (5% of the dataset) were reserved for testing the model's predictive accuracy.
- Nodes in Training Graph: The training graph encompasses all 4,039 nodes, ensuring that the test edges are drawn from a consistent node set.

FEATURE ENGINEERING

1. Common Neighbors

- Measures the number of shared connections between two nodes.
- A high count suggests a higher likelihood of a link forming.

2. Jaccard Coefficient

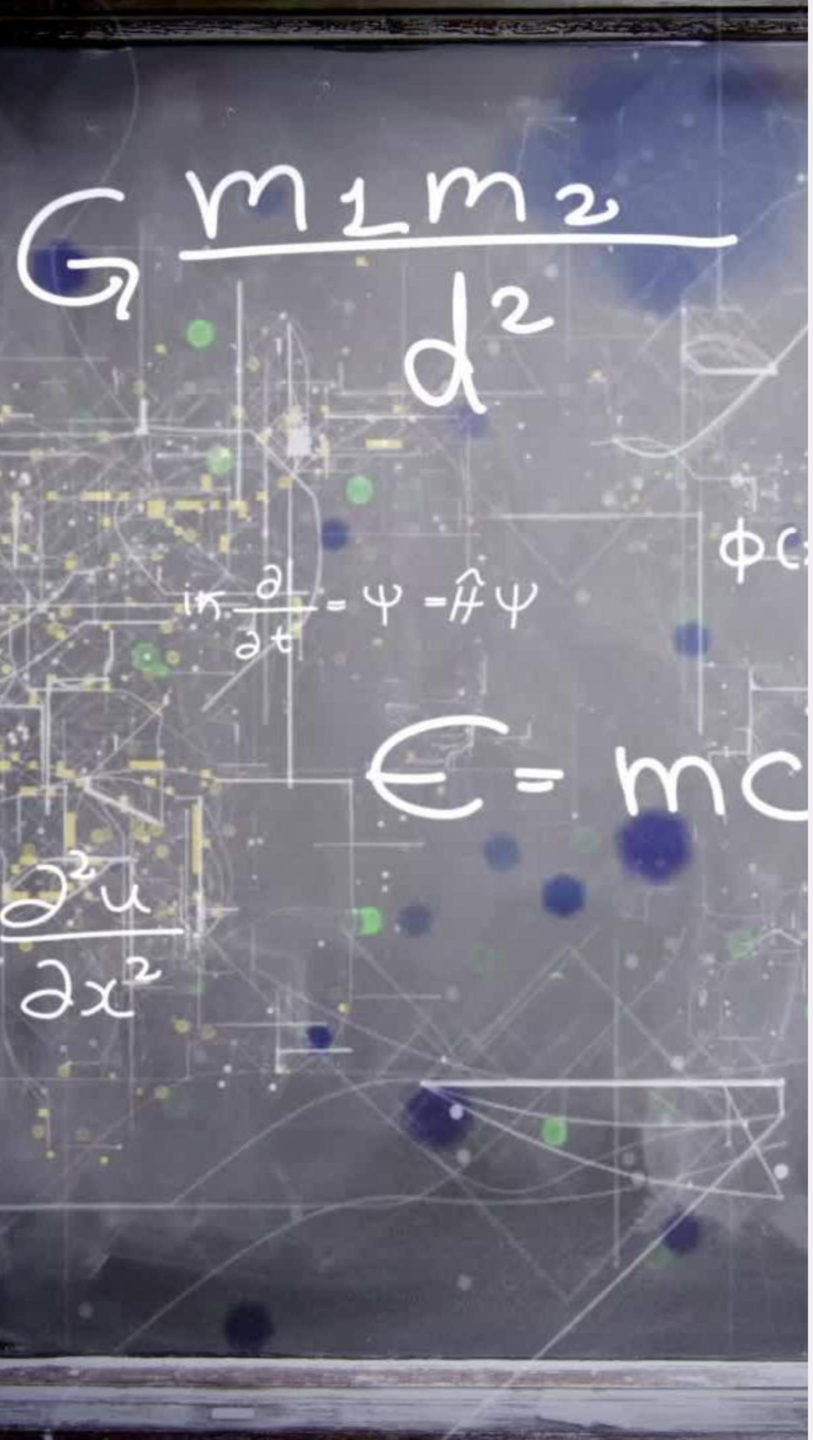
- Captures the similarity between the neighborhoods of two nodes.
- Defined as the size of the intersection divided by the size of the union of the two nodes' neighbors.

3. Adamic-Adar Index

- Weighs shared neighbors inversely by their degree, emphasizing less connected nodes.
- Helps capture nuanced relationship dynamics in the network.

4. Preferential Attachment

- Reflects the tendency of high-degree nodes to attract more links.
- Calculated as the product of the degrees of two nodes.


$G\frac{m_1m_2}{d^2}$
$i\hbar\frac{\partial}{\partial t}\Psi = \hat{H}\Psi$
$E = mc$
$\phi($
$\frac{\partial^2 u}{\partial x^2}$

METHODOLODY

- Three machine learning models were trained and evaluated to predict links based on the engineered features. The results are summarized below:

1. Logistic Regression
   - AUC-ROC: 0.9956
   - F1 Score: 0.9563
   - Precision: Ranges between 0.94 and 0.98 depending on the threshold.
   - Accuracy: Achieved an overall accuracy of 96%.
2. Random Forest
   - AUC-ROC: 0.9950
   - F1 Score: 0.9725
   - Precision: Consistently between 0.97 and 0.98.
   - Accuracy: 97%.
3. XGBoost
   - AUC-ROC: 0.9957 (highest among the models).
   - F1 Score: 0.9690
   - Precision: Ranges from 0.96 to 0.97.
   - Accuracy: 97%.

FEATURE IMPORTANCE ANALYSIS

1) XGBoost Feature Importance:

Adamic-Adar Index: 0.927862 (dominant feature).

Common Neighbors: 0.027922

Jaccard Coefficient: 0.027589

Preferential Attachment: 0.016628

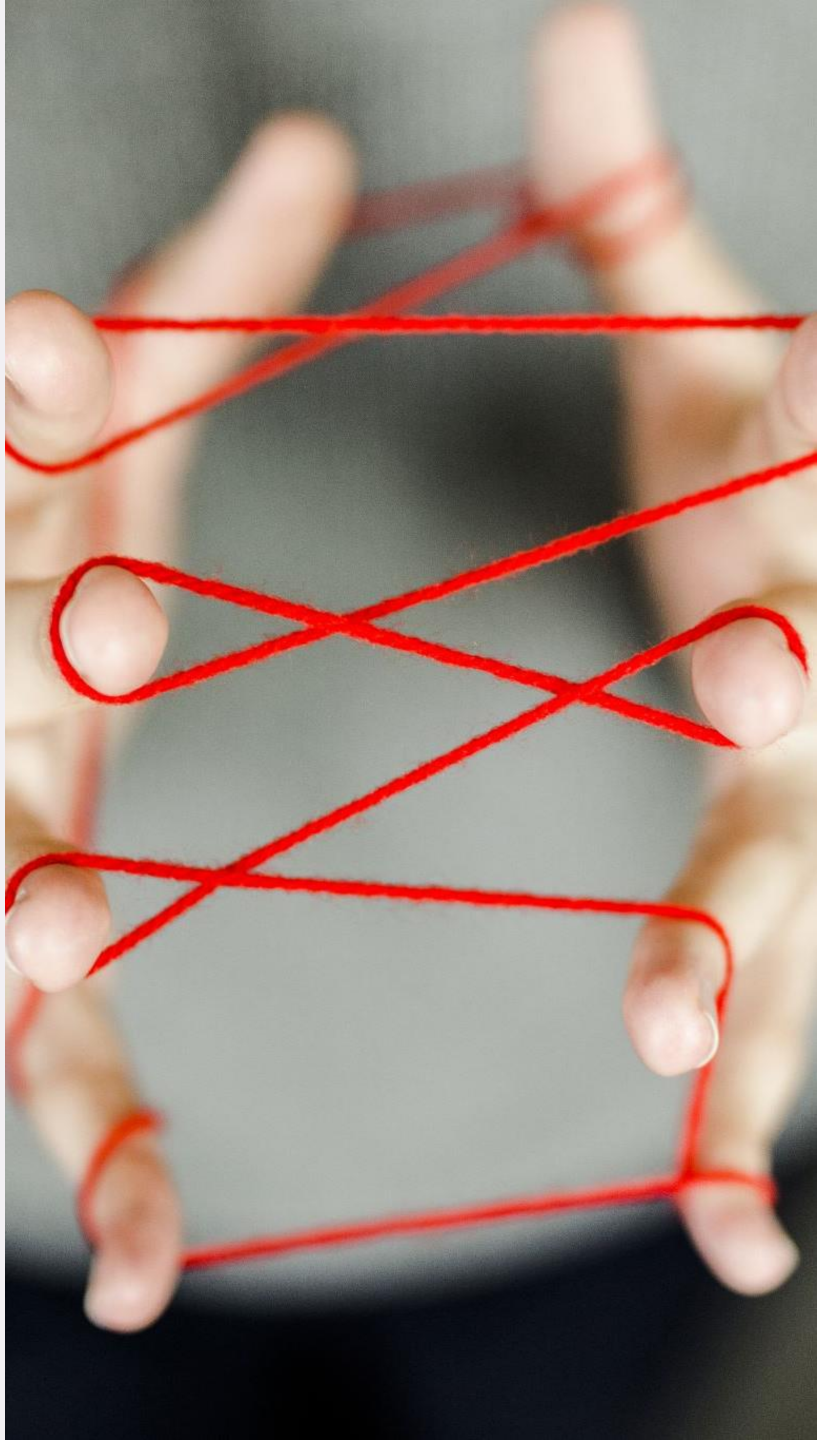
Logistic Regression Coefficients

- Adamic-Adar Index: 7.149487 (most significant feature).
- Jaccard Coefficient: 4.978354
- Common Neighbors: -0.265965 (negative correlation).
- Preferential Attachment: -0.270951

SAMPLE PREDICTIONS

- Random Node Pair Analysis
- Predictions for potential links are based on the models' probability outputs. Examples include:

1. Nodes (2957, 2784): High probability of a link forming (0.709228).
2. Nodes (3722, 3681): Moderate link probability (0.576029).
3. Nodes (1923, 2505): Moderate link probability (0.556137).
4. Most Random Pairs: Low link probabilities (<0.05), indicating unlikely connections.



NODE ATTRIBUTES INSIGHTS

➢ Key network metrics were computed for individual nodes to understand their role in the network:

a. Node 0:

- Degree Centrality: 0.0859
- Clustering Coefficient: 0.0420 (low clustering).

b. Node 2:

- Degree Centrality: 0.0025
- Clustering Coefficient: 0.8889 (high clustering).

➢ Observation:

- Nodes exhibit significant differences in their participation and influence within the network.

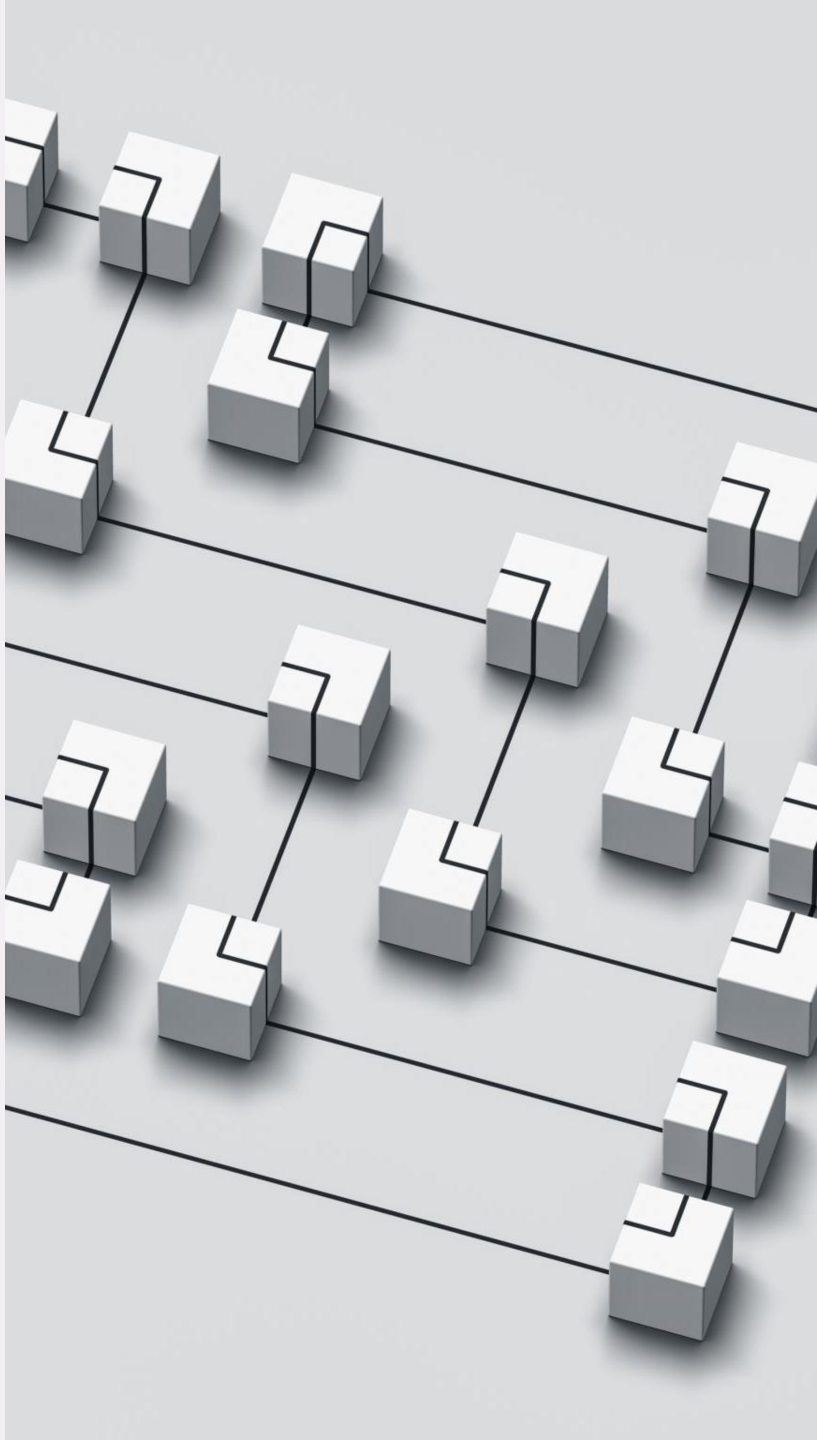
CHALLENGES AND LIMITATIONS

1. Technical Challenges:

- Designing features that effectively capture link prediction signals.
- Balancing model complexity with interpretability.
- Accurately modeling complex social network dynamics.

2. Data Limitations:

- The dataset is anonymized and specific to college networks, limiting generalizability to other contexts.
- Possible biases inherent in the original network structure.



KEY TAKEAWAYS

1. Project Achievements:

- Developed high-accuracy link prediction models with robust performance metrics.
- Demonstrated the importance of graph-based features in social network analysis.
- Provided insights into the mechanisms driving social connections.

2. Future Work:

- Experimenting with advanced methods such as Graph Neural Networks (GNNs).
- Applying the approach to a broader range of datasets to enhance generalizability.
- Refining prediction algorithms for better performance and scalability.

CONCLUSION

- This project highlights the effectiveness of machine learning models in predicting potential social links within networks. It advances our understanding of how friendships form and evolve, with implications for designing smarter social platforms and fostering community growth.

Thank You for Your Attention!

Feel free to ask questions or share feedback.