Community Detection and Analysis on DBLP v9 Dataset: Centrality, Network Metrics, and Predictive Modeling

Xiang Zheng ... **Author n**
21307110169 Address line

School of Data Science, Fudan University

Abstract

abstract

1 Introduction

Network analysis has become an essential methodology for investigating complex systems across diverse domains. The DBLP dataset, a comprehensive bibliographic repository of computer science publications, provides a rich foundation for exploring collaboration patterns and scholarly relationships. This project employs advanced graph analysis techniques to uncover the structural and functional properties of the DBLP v9 dataset.

The study is organized into five core components, each addressing a key aspect of network analysis:

- **Preprocessing**: The raw DBLP dataset was transformed into a structured network by performing data cleaning, filtering, and splitting. Key preprocessing steps included handling missing data, defining nodes and edges, and attributing properties to network elements to prepare the dataset for analysis.

- **Community Detection**: Cohesive subgroups within the network were identified using algorithms such as `Louvain`, `Label Propagation`, and `Multi-level`. These communities reveal collaborative dynamics, research specializations, and the structural foundations of scholarly interactions.

- **Centrality Analysis**: Degree centrality and PageRank centrality were employed to identify influential nodes, revealing hubs and authoritative figures within the network. Additional metrics, including community diameters and average citations per author, were analyzed to characterize structural and functional network properties. The degree distribution was assessed to evaluate the network's topology, with visual representations provided in the visualization section.

- **Link Prediction**: Using the Cora-ML dataset—a citation network of machine learning research papers—supervised learning techniques were applied to predict potential future connections. Structural features such as common neighbors and Jaccard similarity were used, with model performance evaluated through metrics such as AUC and AP. The analysis provides insights into the evolution of citation patterns and network connectivity.

- **Visualization System**: An intuitive visualization system was developed to facilitate the interpretation of results. This system presents community structures, central nodes, and network statistics through graphical representations, offering a comprehensive overview of the network's topology and dynamics.

By integrating these analytical components, this study aims to elucidate the structural characteristics, key actors, and potential evolution of the DBLP co-authorship and publication network. The methodologies and findings presented contribute to a deeper understanding of scholarly collaboration patterns and academic network dynamics.

2 Preprocessing

3 Community Mining

4 Centrality Measurement

5 Link Prediction

6 Visualization System

7 Conclusion

Acknowledgements