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**: The GLACE model was employed on the Cora-ML dataset, a citation network of machine learning research papers, to predict potential future connections between nodes. Structural features such as common neighbors and Jaccard similarity served as input, with the model's performance assessed using metrics like AUC and AP. This analysis offers insights into evolving citation patterns and the dynamics of scholarly connectivity.

- **Visualization System**: An intuitive visualization system was developed to facilitate the interpretation of results. Built using **Python** for the core implementation and **D3.js** for dynamic and interactive visualization, the system presents community structures, central nodes, and network statistics through graphical representations. This approach offers a comprehensive and visually accessible 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

The preprocessing phase plays a crucial role in transforming the raw DBLP v9 dataset into a structured format suitable for comprehensive analysis. This stage encompasses data cleaning, network construction, and feature engineering. The DBLP v9 dataset was selected due to its balance between

computational efficiency and the level of detail required to explore academic collaboration dynamics. Moreover, its extensive coverage of computer science literature ensures the robustness of the analysis.

2.1 Dataset Selection

The DBLP-Citation-network V9 dataset was selected based on several factors, as shown in Table 1. While earlier versions provide valuable data, V9 was preferred due to its balance between data richness and computational manageability, making it well-suited for the available resources.

DBLP-Citation-network V9 was chosen for the following reasons:

- **Data Completeness:** V9 includes 3,680,007 papers and 1,876,067 citation relationships, offering a comprehensive yet manageable dataset.
- **Timeliness:** Released on July 3, 2017, V9 reflects current citation trends up to that point.
- **Balanced Size:** Compared to later versions, V9 offers a good balance between data volume and computational feasibility.
- **Sufficient Coverage:** V9's size enables a wide-ranging analysis of academic collaborations and citation patterns.

Thus, DBLP-Citation-network V9 was selected for its balance of data richness and computational efficiency, making it ideal for this study.

2.2 Data Cleaning and Integration

The raw dataset contains metadata, including titles, authors, venues, and citations. The initial cleaning process involved grouping records by paper, addressing missing or inconsistent data, and resolving redundancies. Missing fields, such as titles, authors, and venues, were assigned default values to ensure consistency, preparing the dataset for subsequent analysis.

2.3 Network Construction

The dataset was structured as a bipartite graph, where authors are represented as nodes and co-authorships as edges. Attributes such as publication count and co-authorship frequency were assigned to the nodes and edges, respectively. Additionally, citation relationships were extracted to construct a directed citation network. This dual representation of collaboration and citation dynamics allows for a comprehensive analysis of academic interactions.

2.4 Feature Engineering

To enhance the analytical power of the dataset, several key features were engineered:

- **Co-authorship Mapping:** Authors were assigned unique identifiers, and co-authorship relationships were mapped, with edge weights representing the frequency of collaboration.
- **Citation Analysis:** Citation relationships were used to calculate in-degree (citations received) and out-degree (citations made) for each paper, providing insights into the influence and impact of academic works.
- **Venue Indexing:** Publication venues were indexed to ensure uniform representation across the dataset, enabling venue-specific analyses.

2.5 Dataset Filtering

To improve computational efficiency, papers with no citations and references were flagged as isolates and excluded from the analysis. Furthermore, thresholds were applied to prioritize significant relationships, ensuring that only meaningful data were retained for the analysis.

3 Community Mining

4 Centrality Measurement

5 Link Prediction

6 Visualization System

7 Conclusion

Acknowledgements

Data Set	# Papers	# Citation Relationships	Comment
Citation-network V1	629,814	>632,752	
Citation-network V2	1,397,240	>3,021,489	
DBLP-Citation-network V3	1,632,442	>2,327,450	
DBLP-Citation-network V4	1,511,035	2,084,019	Arnetminer [2011-01-08]
DBLP-Citation-network V5	1,572,277	2,084,019	Arnetminer [2011-02-21]
DBLP-Citation-network V6	2,084,055	2,244,018	Arnetminer [2013-09-29]
DBLP-Citation-network V7	2,244,021	4,354,534	Arnetminer [2014-05-25]
DBLP-Citation-network V8	3,272,991	8,466,859	Arnetminer [2016-07-14]
ACM-Citation-network V8	2,381,688	10,476,564	Arnetminer [2016-04-02]
ACM-Citation-network V9	2,385,022	9,671,893	Arnetminer [2017-01-20]
DBLP-Citation-network V9	3,680,007	1,876,067	Arnetminer [2017-07-03]
DBLP-Citation-network V10	3,079,007	25,166,994	
DBLP-Citation-network V11	4,107,340	36,624,464	OAG [2019-05-05]
DBLP-Citation-network V12	4,894,081	45,564,149	DBLP+Citation [2020-04-09]
DBLP-Citation-network V13	5,354,309	48,227,950	DBLP+Citation [2021-05-14]
DBLP-Citation-network V14	5,259,858	36,630,661	DBLP+Citation [2023-01-31]

Table 1: Summary of DBLP Citation Networks Versions