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

Xiang Zheng
21307110169

...

Author n
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 is essential for transforming the raw DBLP V9 dataset into a structured format suitable for in-depth analysis. This stage includes data cleaning, network construction, and feature engineering. The DBLP V9 dataset was selected due to its balanced computational efficiency and the required level of detail to explore academic collaboration dynamics. Furthermore, its extensive coverage of computer science literature ensures the robustness of subsequent analyses.

2.1 Dataset Selection

The DBLP-Citation-network V9 dataset was chosen for its optimal balance between data richness and computational feasibility, as outlined in Table 1. While earlier versions of the DBLP dataset offer valuable data, V9 strikes a favorable balance between dataset size and manageability, making it well-suited to the available computational resources.

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

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

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

2.2 Data Cleaning and Integration

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

2.4 Feature Engineering

Several key features were engineered to enhance the dataset's analytical power:

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

2.5 Dataset Filtering

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

2.6 Overview of the Dataset

Following preprocessing, a series of exploratory analyses were conducted using Jupyter Notebooks, pandas, and matplotlib to understand the fundamental characteristics and structure of the dataset. The visualizations presented in Figure 1, 2, 3, and 4 summarize key aspects of academic collaboration and citation dynamics, providing insights into author networks, citation patterns, reference behaviors, and co-authorship structures.

- **Distribution of the Number of Authors per Paper:** This figure presents the distribution of authorship in the DBLP V9 dataset. The majority of academic papers have few authors, with a smaller proportion featuring larger author teams. This reflects the diverse nature of academic collaborations, from single-author studies to multi-author research projects.
- **Citation Distribution of Papers:** This visualization reveals the typical skew in citation

Dataset	# 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 Network Versions

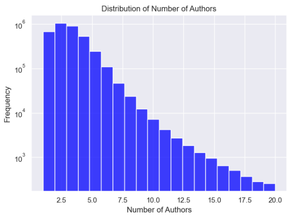

Figure 1: Number of Authors per Paper

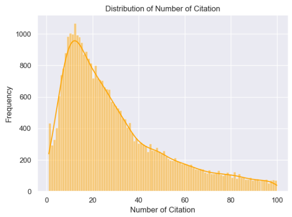

Figure 2: Citation Distribution of Papers

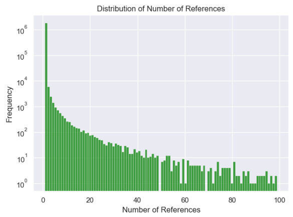

Figure 3: Reference Distribution of Papers

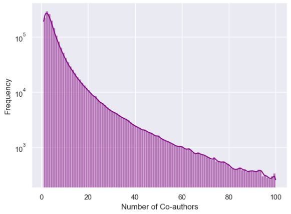

Figure 4: Number of Co-authors per Author

patterns, where a small subset of papers receives the majority of citations, while the majority of papers receive fewer or no citations. This highlights the concentration of academic influence in a limited number of highly cited works.

- **Reference Distribution of Papers:** This figure illustrates the reference distribution, showing that papers with higher citation counts tend to reference more works. This suggests a correlation between the volume of citations and the number of references cited.

- **Distribution of the Number of Co-authors per Author:** This visualization shows the distribution of co-authors per author. A small subset of authors collaborates with many individuals, while most authors collaborate with fewer co-authors. This reflects the presence of central, highly collaborative figures within the academic community, alongside more isolated researchers with fewer collaborative connections.

These visualizations provide valuable insights into the collaborative and citation dynamics within the DBLP V9 dataset. They highlight the structure of academic work, the distribution of citations, and the nature of author collaboration networks, forming a foundation for further exploration of academic collaboration and citation behaviors.

3 Community Mining

4 Centrality Measurement

5 Link Prediction

6 Visualization System

7 Conclusion

Acknowledgements