A statue of a person

Description automatically generated

STAMATOULA-GERASIMOULA MESOLORA

Registration number: f2822308

Social Network Analysis

July 2024

|  |
| --- |
| **MSc Business Analytics** |
| From Raw Data to Temporal Graph Structure Exploration |

Contents

[Introduction 3](#_Toc172847785)

[Μethodology 4](#_Toc172847786)

[Data Preprocessing and Graph Construction 5](#_Toc172847787)

[Graph Analysis 5](#_Toc172847788)

[Average Degree Over Time 5](#_Toc172847789)

[Important Nodes 6](#_Toc172847790)

[Community Detection 7](#_Toc172847791)

[Figures 10](#_Toc172847792)

[Tables 10](#_Toc172847793)

# Introduction

The aim of this project is to analyse the structure and evolution of temporal graphs obtained from raw co-authored data. Using the DBLP dataset, a weighted undirected graphs is constructed to represent co-authorship relationships over the last five years. The project involves several steps, including preprocessing and filtering the data using Unix commands, creating coauthoring graphs using Python and analyzing various graph metrics and community structures using R. Then the graphs are evaluated over time, we examine metrics such as number of vertices, edges, graph diameter and average degree. Additionally, the top-10 authors based on degree and PageRank centrality measures have been identified and three community detection algorithms –“Fast Greedy”, “Infomap” and “Louvain” – have been employed to help us understand the community dynamics. This report documents the methodologies and findings from each stage of the analysis, providing insights into the temporal evolution of co-authorship networks in the selected conferences.

# Μethodology

The analysis starts with data pre-processing, where the raw data from DBLP are filtered to include records from specific conferences (CIKM, KDD, ICWSM, WWW, IEEE BigData) over the last five years. Unix commands are used to manipulate the CSV file, identifying and extracting records by year and conference fields. After that a filtered dataset is created, containing only the necessary co-authored records. In the next step, are constructed weighted undirected co-authorship graphs, one for each of the last five years, extracted from the filtered dataset using Python. These graphs are represented in CSV format, with edges indicating co-authorship and weights representing the number of co-authored publications. This involved creating pairs of authors from each paper and calculating their co-authorship weights.

Subsequently, graph analysis performed using R and the igraph library computed metrics such as the number of vertices (authors), edges (co-authorship links), graph diameter and mean degree. These metrics provided information about the evolution of the co-authoring networks. By calculating the rank and PageRank for each author and extracting the top 10 authors per year we identified influential authors and tracked their importance over time. Community detection is conducted using “Fast Greedy”, “Infomap” and “Louvain” algorithms, identifying clusters of frequently collaborating authors. By comparing these results and visualizing the communities, were revealed significant groupings, tracking the evolution of a selected author’s collaborative network over the five years.

# Data Preprocessing and Graph Construction

The pre-processing of the raw data includes filtering records related to specific conferences ("CIKM", "KDD", "ICWSM", "WWW" and "IEEE BigData") and within the last five years, e.g. years 2016 - 2021. This was achieved by using Unix commands to operate on the original CSV file. In more detail, we calculated the maximum year present in the dataset and set a threshold to filter entries older than five years, while using regex to filter the dataset, keeping only the required conferences. The resulting dataset contains only the necessary co-authored records for further analysis.

With the filtered dataset, we utilized Python to construct weighted undirected co-authorship graphs for each of the last five years. Each graph was represented in a CSV format, where edges between authors indicated co-authorship and the weights represented the number of joint publications. From each paper’s authors we generated pairs and updated their co-authorship weights accordingly. This step involved iterating over the dataset, creating key-value pairs for each unique author combination and calculating their respective weights based on the frequency of their collaboration.

# Graph Analysis

Using R and the igraph library, we imported the generated co-authorship graphs and performed various analyses to uncover temporal trends. Key metrics, as the number of vertices (authors), number of edges (co-authorship links), graph diameter and average degree were calculated and provided insights into the structural changes in the co-authorship networks over the five-year period. The results were visualized in order to allow a clear understanding of the temporal evolution of these metrics.

## Average Degree Over Time

In Figure 1 are depected the charts of each metric that we analize. In more detail, the number of peaks (unique authors) increases steadily over the five-year period, suggesting that an increasing number of authors attend the selected conferences each year. Similarly, the number of peaks (co-authors) also increases steadily, reflecting increasing collaboration between authors over the years. The diameter of the graph shows greater variation compared to other metrics, with a notable peak in 2018 and subsequent decline, indicating varying levels of the longest shortest path in the graph. Finally, the average grade shows a general increase over the years, with a notable jump in 2019, implying that, on average, authors are collaborating with more co-authors each year. Altogether, while the number of peaks and edges shows a steady upward trend, the diameter and average degree show more variability, indicating dynamic changes in the structure and connectivity of the co-authoring network over the five-year period.

A graph of different stages of growth

Description automatically generated with medium confidence

Figure 1 - Graph Measures

## Important Nodes

To identify influential authors in the co-authoring networks, we calculated the degree and PageRank for each author. By extracting the top 10 authors based on these centrality measures for each year, we examined the potential of authors' influence over time.

Tables 1 & 2 below show that authors Philip S. Yu, Jiawei Han and Hui Xiong consistently appear in the top 10 author lists for both degree and PageRank. PageRank provides additional information about the quality of connections, highlighting authors who are not only productive collaborators but also connected to other influential authors. The presence of consistent top authors indicates that certain authors or research groups are pivotal in the key conferences under study, driving important parts of research production and collaboration. Variations in the lists suggest dynamic changes in the network, with new authors occasionally appearing as influential collaborators. In sum, these insights emphasize the key role of some authors in the research community and highlight the dynamic nature of the co-authoring network.



Table 1- Top 10 Authors by Degree



Table 2 - Top 1 Authors by PageRank

## Community Detection

Finally, community detection is performed through three different algorithms: Fast Greedy, Infomap and Louvain. These methods were applied to each year's co-authorship graph to identify clusters of authors who collaborated regularly. The results of these algorithms were compared to assess their performance and consistency. In addition, a random author was selected to track its community membership over the five years. Finally, we visualized medium-sized communities to create meaningful representations of co-authoring networks. This visualization was enhanced by filtering out very small and very large communities to focus on more important groupings.

* The Fast Greedy algorithm and the Infomap algorithm have a relatively high similarity, approximately 0,76.
* The Fast Greedy algorithm and the Louvain algorithm have a lower similarity, approximately 0,42.
* The Infomap algorithm and the Louvain algorithm have a moderate similarity, approximately 0,64.

This suggests that the Fast Greedy (Figure 4) and Infomap (Figure 3) algorithms detect similar community structures, while the Louvain algorithm (Figure 2) detects rather different structures.

After that, we tested the participation of a random user "Yong Li 0008" in the communities detected by each algorithm. This shows that this user is included in different communities by each algorithm:

* Community 2 in the Fast Greedy algorithm
* Community 10 in the Infomap algorithm
* Community 7 in the Louvain algorithm

This indicates that different algorithms can produce different community assignments, even for the same graph and the same user. This emphasizes the importance of choosing the right algorithm based on the specific attributes of the graph and the research question.

A group of dots and lines

Description automatically generated

Figure 2 - Communities According to 'Louvain' Clustering

A group of colorful dots

Description automatically generated

Figure 3 - Communities According to 'Infomap' Clustering

A group of dots and lines

Description automatically generated

Figure 4 - Communities According to 'Fast Greedy' Clustering

Figures

[Figure 1 - Graph Measures 6](#_Toc172847769)

[Figure 2 - Communities According to 'Louvain' Clustering 8](#_Toc172847770)

[Figure 3 - Communities According to 'Infomap' Clustering 8](#_Toc172847771)

[Figure 4 - Communities According to 'Fast Greedy' Clustering 9](#_Toc172847772)

Tables

[Table 1- Top 10 Authors by Degree 6](#_Toc172847773)

[Table 2 - Top 1 Authors by PageRank 7](#_Toc172847774)