Project 2: Analysing the DBLP dataset

Fall 2022

Course: DASC 5300/CSE 5300

Team 30 members:

1. Joseph, Nelson (1002050500)

2. Manavalan, Jeeva (1002067528)

Introduction

The data in this project was analysed using a graph data structure. The DBLP data manages authors, their publications, conferences, citations, etc in computer science discipline. Analysis was performed on sample data initially, then on 21000 rows of data in accordance with the parameters provided, and ultimately on the whole dataset. The Python Networkx library was used for the analysis. For data analysis, three graphs—the **Known-Author-Graph**, the **Paper-Citation-Graph**, and the **Author-Venue-Graph** were created. The network characteristics of a sample of each type of graph were analysed, and the findings were manually verified to ensure accuracy.

Overall Status

- There are 362865 papers in the data without any references. They wrote the paper without referring to anyone. with a total of 11.7% of the papers are without any reference.
- There are 506699 Null values in the venue Column that was empty before. We can understand that there are 506699 papers in this 3079007 papers that are not published in any Venue from the initial pre-processing
- The most papers were given in 2016, with a total of 4207, when 21,000 rows of data were analysed.

File Description

Jupyter Notebook	• DASC5300_Proj2_Fall22_team_<30>.ipynb
JSON	• dblp-ref-0. Json, dblp-ref-1. Json, dblp-ref-2. Json, dblp-ref-3. Json
Python	• MLN.py, MLN_IO.py, network_summary.py
Other required	• meta-info-for-v10.docx
Folder	• known_author_graph_sample, known_author_graph_21k, paper_edge_list_sample, Paper_citation_21k, Author_venue_graph_sample, Author venue graph 21k

Division of Labour

TASK	MEMBER
Pre-processing and graphs creation (including sample)	Pre-Processing is done together
Analysis 0: graph characteristics analysis	Analysis by Nelson Manual verification by Jeeva
Analysis 1: Analysis on sample data and interpretation	Nelson
Analysis 2. Finding groups of authors	Jeeva
Analysis 3a: Top 5 to 10 authors with most papers	Nelson
Analysis 3b: Finding 5 to top 10 most cited papers	Nelson
Report including analysis and how you have verified results	Jeeva and Nelson together

Pre-Processing

- Imported the necessary libraries and researched about network graphs
- Only 2 columns have integer type. Except for 'n_citation' and 'year,' all are objects, and we removed the abstract column as it was not necessary for analysis.
- Checked for Null values and identified that 11.7% of the papers are without any reference and 506699 papers in this 3079007 rows of data that are not published in any Conference.

Visualization Nomenclature

1. **Author Node** - White Node with Dark orange Node edges
2. **Paper referred Node** - White Node with Light blue Node edges
3. **Paper being referred Node** - Gold Node with white Node edges
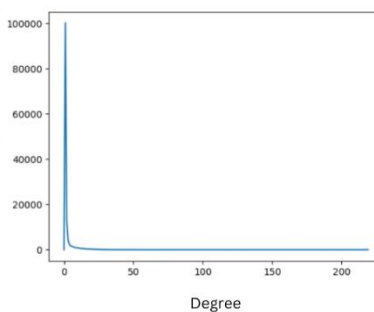4. **Venue Node** - Lime Green Node

Analysis 0: Graph Characteristics Analysis

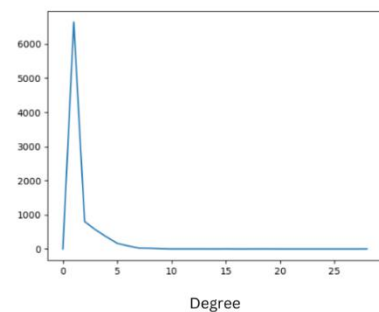| GRAPH CHARACTERISTICS	KNOWN-AUTHOR GRAPH	PAPER-CITATION GRAPH	AUTHOR-VENUE-GRAPH
Number of nodes	53740	132374	8681
Number of edges	96592	153533	6637
Density	6.689349873130404e-05	1.7523854067997532e-05	0.00017616204041136503
Number of Connected Components	11261	3418	2058
Connected Components (their characteristics)	📄 output .txt	📄 outpu 1t.txt	📄 output2.txt
Diameter	-1	-1	-1
Minimum degree	1	1	1
Maximum degree	68	219	28
Average degree	3.5947897283215484	2.319685134543037	1.5290865107706486
Std dev of degree	3.0615492203658454	4.519652538304593	1.2372346178242324

Degree distribution of Known Author Graph
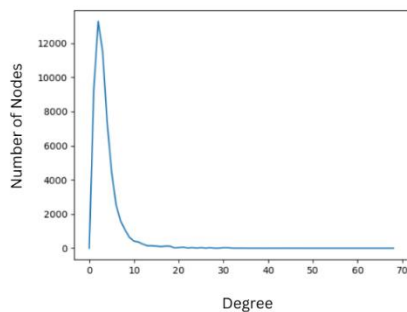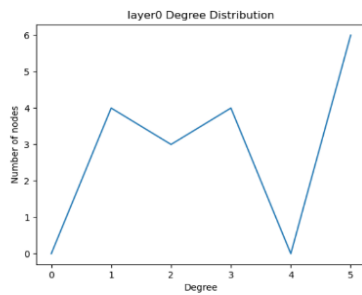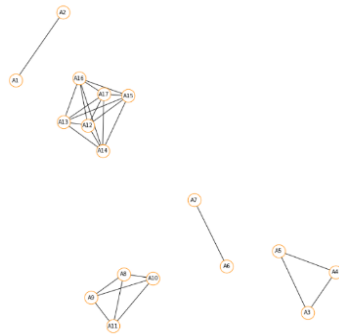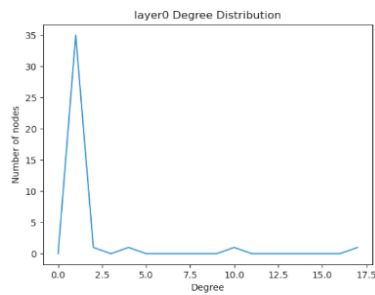

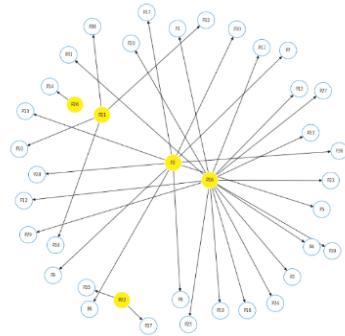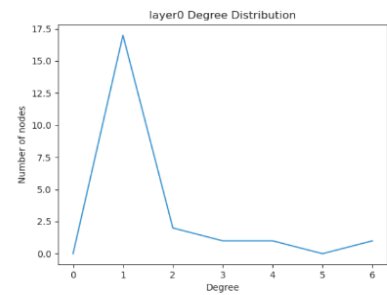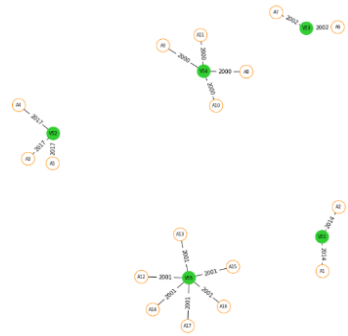Degree distribution of Paper Citation Graph

Degree distribution of Venue Author Graph

Analysis 1: Analysis of Sample Data and Interpretation

GRAPH CHARACTERISTICS	KNOWN-AUTHOR GRAPH		PAPER-CITATION GRAPH		AUTHOR-VENUE-GRAPH	
	Values	Manual values	Values	Manual values	Values	Manual values
Number of nodes	17	17	39	39	22,	22
Number of edges	26	26	34	34	17	17
Density	0.1911764705 8823528,	0.1911	0.045883940 62078273	0.045883940 62	0.073593073 5930736,	0.073593073 59
Number of Connected Components	5	5	5	5	5	5
Connected Components (their characteristics)	[6,4,3,2,2]	[2,2,3,4,6]	[18,11,5,3,2]	[2,3,5,11,18]	[7, 5, 4, 3, 3]	[3,3,4,5,7]
Diameter	1	1	1	1	1	1
Minimum degree	1	1	1	1	1	1
Maximum degree	5	5	17	17	6	6
Average degree	3.0588235294 117645	3.05	1.743589743 5897436	1.743589743 59	1.545454545 4545454	1.545454545 45
Std dev of degree	1.6382379343 098379	1.588	2.926448797 451027	2.888761725 67	1.262170191 8020378	1.229337106 65

The table above displays the results from the network summary file. Data is manually checked using the appropriate formula. Each graph's diameter is indicated as 1, however the dimension of an unconnected graph cannot be determined.

Known-Author-Graph   **Paper -Citation Graph**   **Author Venue Graph**

We determined network's structures and discriminated between various network types using the degree distribution plot. When compared to another undirected graph, the directed paper citation graph has a higher probability distribution for these degrees. Each graph's degree is represented by the X axis, while its number of nodes is represented by the Y axis.

Analysis 2. Finding groups of authors

- The Number of Cliques of size 3: 286
- The Number of Cliques of size 4: 715
- The Number of Cliques of size 5: 1287
- The Number of Cliques of size 6: 1716

These figures show how many writers collaborated to create a paper.

Analysis 3a: Top 5 papers that are cited the most.

'P87953', 'P120388', 'P114955', 'P87674', 'P72189' are the papers that are being sited the most. We already have the reference id of the paper. So, we can identify the Title and unique id from the Formatted paper id.

These are the papers mostly related to science and technology.

Analysis 3b: Top 5 Authors with most no of Papers.

AUTHORS	AUTHOR INFORMATION	SOURCE LINK
Toshio Fukuda	Toshio Fukuda has 756 publications, a purple ribbon medal, and a Humboldt Prize holder.	[Hyperlink]
Wei Wang	Wei Wang started his career in 2010 and he is still in paper publications.	[Hyperlink]
Hideto Ide	Researcher in biomedical field.	[Hyperlink]
Lei Wang	Lei Wang is a researcher in biomedical	[Hyperlink]
C.-C. Jay Kuo	C.-C. Jay Kou is a google scholar from MIT Citations:42471, 14460 h-index:93, 59 i10-index:630, 259	[Hyperlink]

Problems encountered

- We had trouble with self-loops while visualizing the known-author graph and subsequently discovered that the network includes a function for it, but we were able to fix the issue by altering the new dictionary.

- Creating unique paper identifiers proved difficult. The paper is referred to as both the paper id and the reference id, therefore in order to maintain consistency in formatting across all papers. In order to prevent the same paper from receiving the same name more than once, we consolidated the papers in paper id with reference id, removed the unique ones from paper references, and made formatting changes.

- Removing nan values from keys so they can be considered when plotting the graph. Since such papers were not published at any conferences, NaN values were not considered. When creating the dictionary for the paper citation graph, this was a difficulty. When the paper id and paper reference columns were combined. NaN ended up being the key for several paper references. To clean it, we combined the dictionaries. **Copy ()** function with another technique.

Conclusion

According to the overall analysis of the 21000 rows of data, most publications were published in 2016. Up until 2005, the number of papers published year grew linearly, but after that point it grew exponentially. The Known-Author-Graph, the Paper-Citation-Graph, and the Author-Venue-network Graph's properties were examined. The outcomes were manually checked. Each graph's degree distribution was plotted using Matplotlib and the networkx summary file. The authors who published the most papers over the period were identified. The In-degree centrality on paper citation graph is used to determine the top 5 papers that are most frequently cited. Most publications have been published in the fields of science and technology, with the biomedical discipline coming in second.