Report On:

**Project II: DBLP Data Analysis Using Graph Characteristics**

**Team 18**

**Fall 2022**



Report By:

Nikhil Das Karavatt

1002085391

Table of Contents

[Overview: 3](#_Toc118411859)

[File Descriptions: 3](#_Toc118411860)

[Division of Labour: 4](#_Toc118411861)

[Problems Encountered: 4](#_Toc118411862)

[Analysis: 4](#_Toc118411863)

[1. Analysis 0 4](#_Toc118411864)

[2. Analysis 1 7](#_Toc118411865)

[3. Analysis 2 9](#_Toc118411866)

[4. Analysis 3.2 9](#_Toc118411867)

# Overview:

We are given a large DBLP data set, with authors, their publications, conferences, citations etc. for each computer science field. During this project, we were able to explore various domains, applications, and challenges, related to Python and its libraries, especially when it comes to the field of graph analysis and visualization. After uploading the DBLP data to google drive, we read the data in the collab from the google drive by making it into chunks. Then we concatenated the chunks into a new DataFrame and took a random sample of 5 from that DataFrame. We pre-processed the extracted sample for each graph as per the requirement by removing NaN values and empty sets. After cleaning the data, we performed graph analysis.

**Graph 1:** We created an undirected graph connecting each pair of authors which are related to each other. We mapped unique author names with distinct naming convention (Ai where i =1, 2, …, n) to understand the graph in a better way. We created an author node for each author and connected it to another author of same paper using undirected edge and made sure that authors were not self-looped. This graph is termed as **known author graphs**.

**Graph 2:** We created a directed graph in which we created a relation of paper id with paper references that were referenced in the paper. Furthermore, each paper with a unique paper ID may be cited by other papers as well thereby making it as a reference for that paper. This graph is termed as **paper citation graph.**

**Graph 3:** We created an undirected graph for authors who have published a paper in conference. In this graph we connected each author of the paper to the venue where they represented the paper in conference in which the paper was published. This graph is termed as **author venue graph.**

# File Descriptions:

There are in total 7 files in which first four are the main files which has helped me with data visualization while other three was just for reference purpose. All files are in .csv format

|  |  |
| --- | --- |
| File Name | File Description |
| dblp\_v10.zip | This is an input file which contains 4 .Json files + 1 meta data file which we uploaded in google drive for analysis |
| author\_edge\_list\_5.txt | This file provides the edge list of known author graph when the sample is 5 |
| author\_edge\_list\_output\_5.txt | This file provides output of known author graph characteristics when the sample is 5 |
| author\_edge\_list\_20.txt | This file provides the edge list of known author graph when the sample is 20000 |
| author\_edge\_list\_output\_20.txt | This file provides output of known author graph characteristics when the sample is 20000 |
| reference\_id\_edge\_list\_5.txt | This file provides the edge list of paper citation graph when the sample is 5 |
| reference\_id\_edge\_list\_output\_5.txt | This file provides output of known paper citation graph characteristics when the sample is 5 |
| reference\_id\_edge\_list\_20.txt | This file provides the edge list of paper citation graph when the sample is 20000 |
| reference\_id\_edge\_list\_output\_20.txt | This file provides output of known paper citation graph characteristics when the sample is 20000 |
| author\_venue\_edge\_list\_5.txt | This file provides the edge list of author venue graph when the sample is 5 |
| author\_venue\_edge\_list\_output\_5.txt | This file provides output of author venue graph characteristics when the sample is 5 |
| author\_venue\_edge\_list\_20.txt | This file provides the edge list of author venue graph when the sample is 20000 |
| author\_venue\_edge\_list\_output\_20.txt | This file provides output of author venue graph characteristics when the sample is 20000 |
| DASC5300\_Proj2\_Fall22\_team\_18 | This is .ipynb file which contains the code which I used to analyse the input data using graph characteristics. |

# Division of Labour:

The project was done solely by me and it took me around 60 hours of work to finish it.

# Problems Encountered:

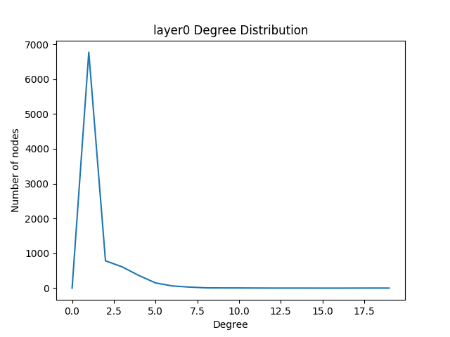
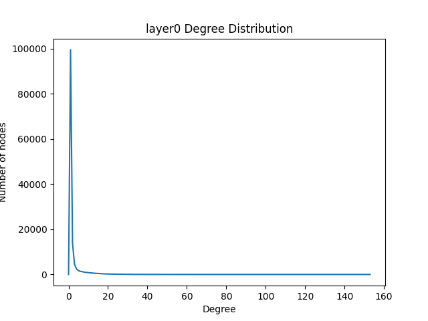
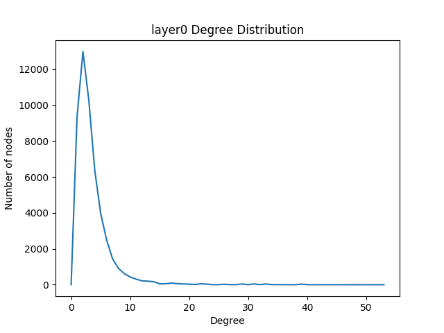
1. While making the edges for the first graph, we were getting an issue of self-node, that is (A1, A1) for a graph edge for which we fixed by fixing the code for loop while creating the dictionary and also we found a predefined function in Networkx which does the same, g.remove\_edges\_from(nx.selfloop\_edges(g))
2. The graph layout was overlapping. Hence making it difficult to read the graph. We used various Networkx functions and parameters to fix this and make it as readable as possible.
3. Cleaning the data with NaN values and empty values wasn’t an easy task, as using dropna will unnecessarily drop rows which may affect the graph analysis so to fix this we used each for function for each graph as per that graph’s requirement. There by, avoiding unwanted drop of rows.

# Analysis:

## Analysis 0

I have taken sample as 20,000 for find out below graph characteristics:

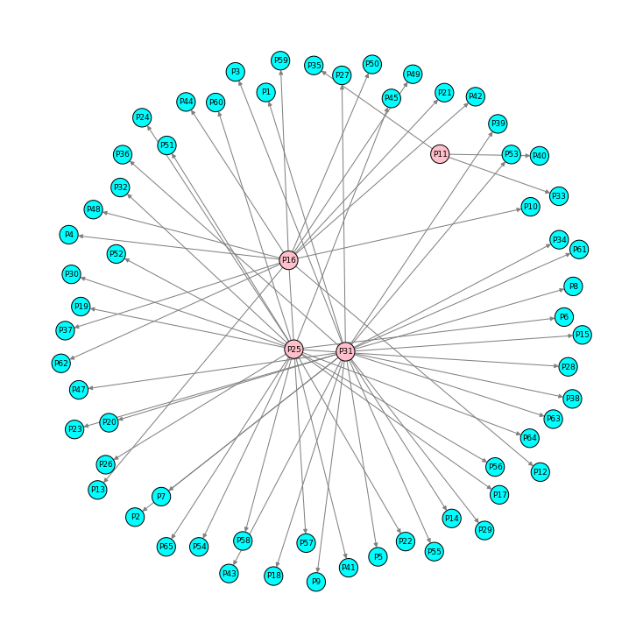
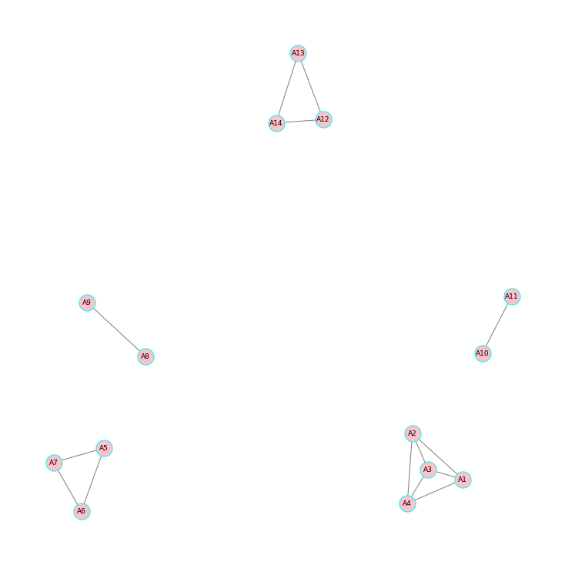
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Graph Characteristics | Known Author Graph | Paper Citation Graph | Author Venue Graph | Remarks | |
| Number of Nodes | 50436 | 133329 | 8805 | We can see that for Paper Citation Graph has the greatest number of nodes created while Author Venue Graph has the least. This implies that the count of paper references is the maximum while the count of authors and venue is minimum. |
| Number of Edges | 90153 | 155457 | 6646 | We can see that for Paper Citation Graph has the greatest number of nodes created while Author Venue Graph has the least. This implies that the connectivity between nodes or inter relation between the nodes is highest between references while lowest in authors and venue. |
| Density | 7.088225372874833e-05 | 1.7490180515350128e-05 | 0.00017146715356527065 | The Density refers to how closely the nodes are connected to each other. This shows that author- venue are close connected while authors and the references are sparsely connected. |
| Number of Connected Components | 11064 | 3613 | 2168 | From this we can see that Known Author Graph has more connected components while Author Venue Graph has the least. |
| Diameter | -1 | -1 | -1 | The diameter is same for all which is 1. |
| Minimum degree | 1 | 1 | 1 | The minimum degree is same for all. The node with least number of connections which is 1 for all. |
| Maximum degree | 53 | 153 | 19 | We can see that Paper Citation Graph has a node which is connected to 153 other nodes while Author Venue Graph has a node which is connected to 19 other nodes. |
| Average degree | 3.574946466809422 | 2.331930787750602 | 1.509596819988643 | On average Authors are more connected to each other, while author-venue are least connected. |
| Std dev of degree | 3.322134229275585 | 4.36288917175383 | 1.1878549742627864 | The standard deviation is maximum for Paper Citation Graph and least for author-venue. |



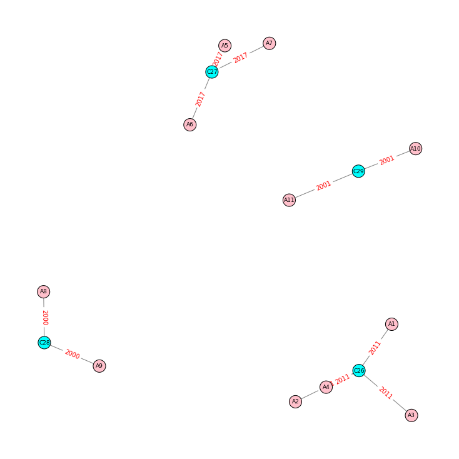
Known Author Graph Known Paper Citation graph Author Venue Graph

## Analysis 1

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Graph Characteristics | Known Author Graph | | Paper Citation Graph | | Author Venue Graph | |
| Calculation | Coded | Manual | Coded | Manual | Coded | Manual |
| Number of Nodes | 14 | 14 | 64 | 64 | 15 | 15 |
| Number of Edges | 14 | 14 | 60 | 60 | 11 | 11 |
| Density | 0.15384615384615385 | 0.15384615384615385 | 0.02976190476190476 | 0.02976190476190476 | 0.10476190476190476 | 0.10476190476190476 |
| Number of Connected Components | 5 | 5 | 4 | 4 | 4 | 4 |
| Diameter | -1 | -1 | -1 | -1 | -1 | -1 |
| Minimum degree | 1 | 1 | 1 | 1 | 1 | 1 |
| Maximum degree | 3 | 3 | 24 | 24 | 4 | 4 |
| Average degree | 2 | 2 | 1.875 | 1.875 | 1.4666666666666666 | 1.4666666666666666 |
| Std dev of degree | 0.7844645405527362 | 0.7844645405527362 | 3.938011745530542 | 3.938011745530542 | 0.9154754164341269 | 0.9154754164341269 |



Known Author Graph Paper Citation Graph



Author Venue Graph

Ground Truth:

Graph 1: Author-Author Graph (Undirected Graph)

Number of Nodes (n)= 14

Number of Edges (e)= 14

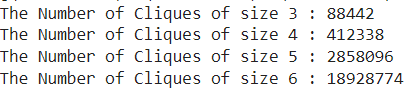
Max(E)= n(n-1)/2= 14(14-1)/2= 91

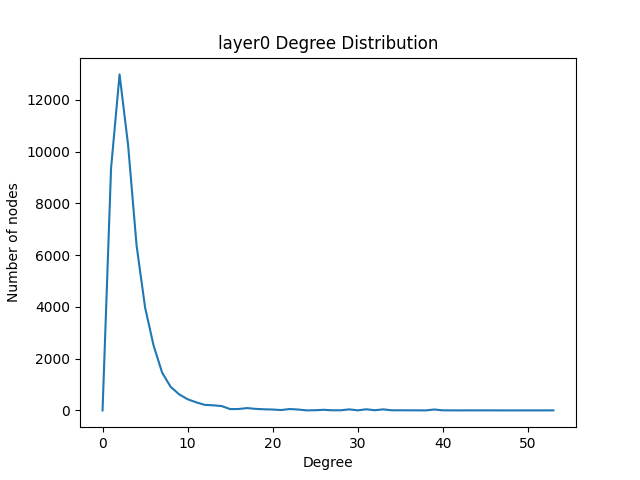
Density= e/Max(E) = 14/91= 0.153846153

Average Degree= Sum [Degree (Node)] / Total Number of Nodes

Average Degree = (3+3+3+3+2+2+2+1+1+1+1+2+2+2)/14= 2

## Analysis 2





By using Known Author Graph, we found out the most number of cliques for different sizes and we found the clique with size 6 has the greatest number of cliques which means that 6 authors are connected the most.

## Analysis 3.2

Top 10 most cited papers:

2983ea61-599e-40d3-8ee2-59ce563244c9

30edd3ae-e4a8-46e5-9a34-a16b4eb4d9fd

4aefe226-85d2-4005-99bb-1392f063fce2

597aa03c-4ad8-4821-aa19-248040cfcd59

6a993fe1-34a0-4e41-b124-9ff35db019d9

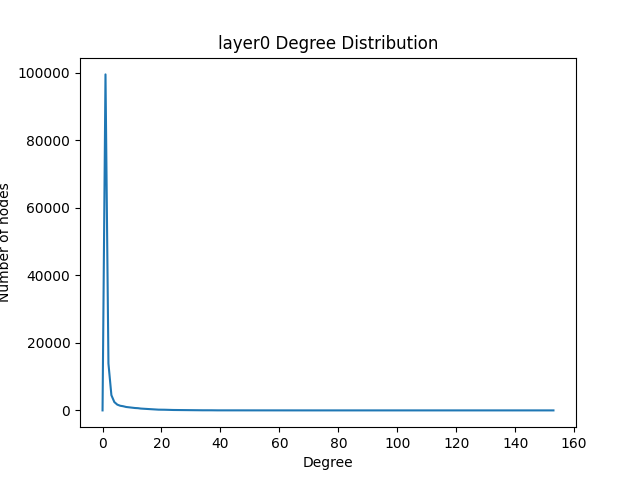
810b77dc-0362-4814-b515-16257a529539

881ed4ce-ed88-4c86-9dd4-83e715e779d3

902abff0-9109-4281-ae1e-20d3dfd0b3ed

990162b7-c3e4-4de9-ad46-4b1637861a54

9a525732-6489-48e3-bd5b-3276d09dee75



By using the code and the nodes of Paper Citation Graph, we were able to find the paper references which was cited the most, this is basically found using the nodes which has highest degree of centrality and we sorted this in a negative way or descending order.