DASC-5300

Foundation of Computing

Project 2: DBLP Data Analysis using Graph Characteristics

UNIVERSITY OF
TEXAS
ARLINGTON

Submitted by: Team 52

Hevin Patel - - 1002036919

Sakshi Nischal - - 1002085832

Overview:

In this project we were given a large DBLP data set, with authors, their publications, conferences, citations etc. for each computer science field. The data given was in json format. We were able to explore various domains, applications, and challenges, related to Python and its libraries. First and foremost, we read the data from the google drive into chunks. Then we concatenated the chunks into a new DataFrame and took a random sample from that DataFrame. We pre-processed the extracted sample and used that data to perform graph analysis.

Graph 1: We created an undirected graph connecting each pair of authors of authors in pair. We mapped unique author names with distinct naming convention 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. Authors were not duplicated.

Graph 2: Then we created a paper citation graph in which one paper id was connected to the other paper id that were referenced in the paper. Furthermore, this paper may be cited by other papers. This was a directed graph from the paper citing other papers. Paper Labels were not duplicated.

Graph 3: Finally, we created an undirected graph for authors who have published a paper in conference. The graph connected each author of the paper to the conference in which the paper was published.

File Description:

File Name	File Description
DASC5300_Proj2_Fall22_Team_<52>.ipynb	Code for graph analysis
graph_edgelist	Input edge list file for network summary package
graph_ntw_summary	Output network summary using network summary package
graph_ntw_summary.txt_layer0_deg_dist	Output graph of number of nodes and degree distribution
Manual Verification for Graph Calculations	Pdf for handwritten calculations

Division of Labor:

We equally contributed to the successful completion of the project.

Hevin Patel (1002036919) contributed to the pre-processing of the data for the graph analysis. Also, he made Author-Author Graph and performed analysis for the same. He also performed Analysis for finding the number of cliques of different sizes. He did manual calculations for the various graph characteristics. Also, he significantly helped in making the report.

Sakshi Nischal (1002085832) contributed to the pre-processing of the data for the graph analysis. She worked on Paper Citation graph and Author-Venue graph and performed the analysis for the same. She performed the analysis for finding top 10 most cited papers in the two years. Also, she significantly helped in making the report.

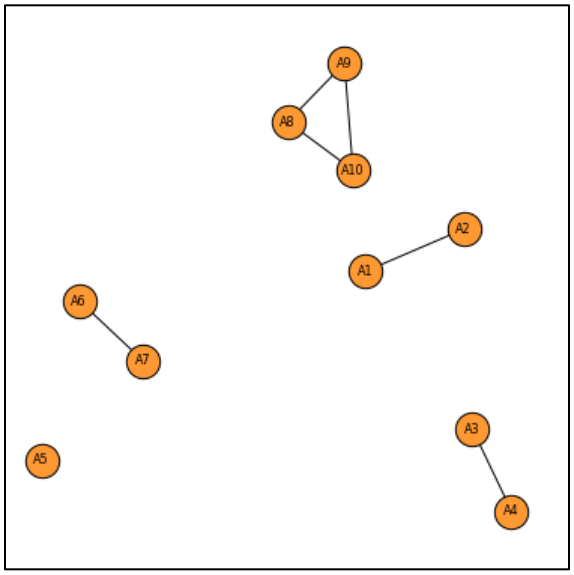
Problems Encountered:

1.) While creating the first graph, we were getting self-loops for the author nodes. To overcome this problem, we ensured that no key of the author dictionary is having itself as a value. Also, we can resolve the issue of self-loops by using a pre-defined function from networkX package.

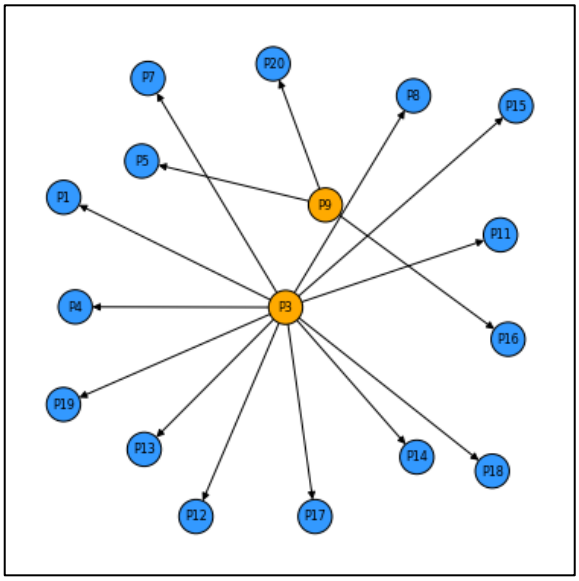
2.) Moreover, we encountered that there were many authors who published multiple papers. To handle those duplicate author values, we used a for loop to iterate through the entire column and just add those values to the new dictionary that are not seen already.

3.) Also, we saw that the nodes and the edges of the graphs were overlapping. We resolved it by using spring_layout() function of the NetworkX package. Here, we adjusted value for k (Optimal distance between nodes) to make the nodes bit apart from each other.

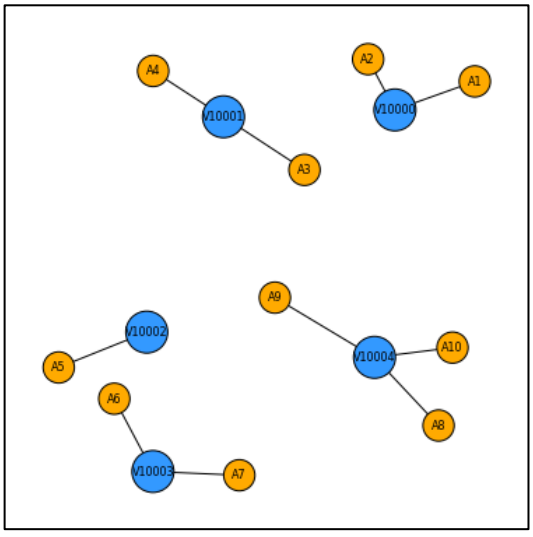
Graph Snapshots for Sample of Five Rows



Author-Author Graph



Paper Citation Graph



Author-Venue Graph

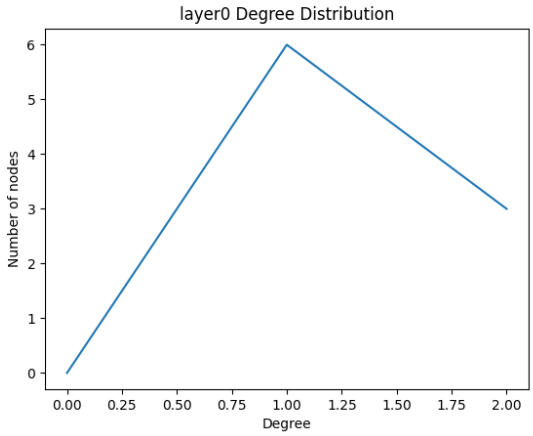
Analysis:

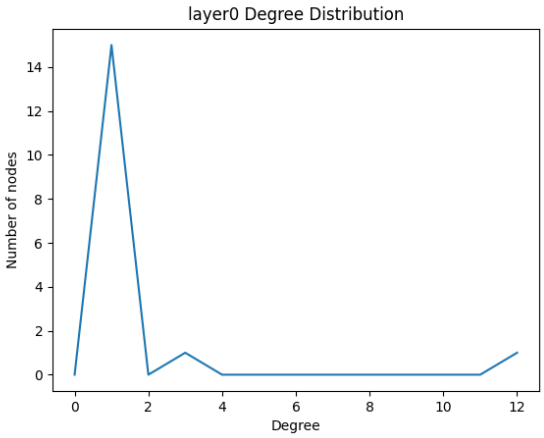
Analysis 0: Graph Summary using Network Characteristics Package

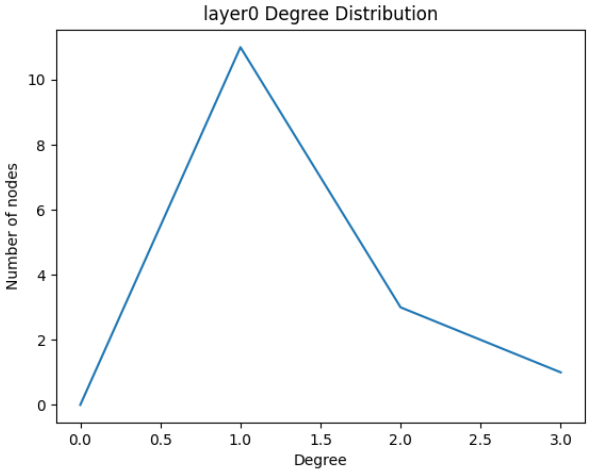
For Sample=5

Layer 0	Graph-1 (Author-Author graph)	Graph-2 (Paper Citation Graph)	Graph-3 (Author-Venue Graph)
Number of nodes	9	17	15
Number of edges	6	15	10
Number of connected components	4	2	5
Density	0.17	0.11	0.09
Diameter	-1	-1	-1
Min Degree	1	1	1
Max Degree	2	12	3
Average Degree	1.33	1.76	1.33
Standard Deviation Degree	0.5	2.69	0.61

Table 0.1







For Sample= 25000

Layer 0	Graph-1 (Author-Author graph)	Graph-2 (Paper Citation Graph)	Graph-3 (Author-Venue Graph)
Number of nodes	62500	152586	8786
Number of edges	114355	182002	6753
Number of connected components	12437	3710	2054
Density	5.8550696811148977e-05	1.563434264520827e-05	0.00017498216298734688
Diameter	-1	-1	-1
Min Degree	1	1	1
Max Degree	80	219	19
Average Degree	3.66	2.38	1.54
Standard Deviation Degree	3.21	4.56	1.22

Table 0.2

From Table 0.1 and Table 0.2, it is observed that the Paper Citation Graph has the maximum number of nodes created. However, the Author Venue Graph has the least. This implies that the connectivity between nodes is highest between references while lowest in authors and venue.

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.

From this we can see that Known Author Graph has more connected components while Author Venue Graph has the least.

The diameter is same for all which is 1.

The minimum degree is same for all. The node with least number of connections which is 1 for all.

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.

On average Authors are more connected to each other, while author-venue are least connected.

The standard deviation is maximum for Paper Citation Graph and least for author-venue

Analysis 1: Manual Calculations for Verification

For Sample=5

Graph 1: Author-Author Graph (Undirected Graph)

Number of Nodes (n)= 9

Number of Edges (e)= 6

Max(E)= n(n-1)/2= 9(9-1)/2= 36

Density= e/Max(E) = 6/36= 0.167

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

Average Degree = (1+1+1+1+0+1+1+2+2+2)/10= 1.2

Graph 2: Paper Citation Graph (Directed Graph)

Number of Nodes (n)= 17

Number of Edges (e)= 15

Max(E)= n(n-1)/2= 17(17-1)/2= 136

Density= e/Max(E) = 15/136= 0.11 (For undirected graph)

Density=e/n(n-1) = 15/272= 0.05 (For directed graph)

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

Average Degree = (1+1+1+1+1+1+1+1+1+1+1+1+1+1+12+3)/17= 1.76

Graph 3: Author-Venue Graph (Undirected Graph)

Number of Nodes (n)= 15

Number of Edges (e)= 10

Max(E)= n(n-1)/2= 15(15-1)/2= 105

Density= e/Max(E) = 10/105= 0.09

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

Average Degree = (1+1+1+1+1+1+1+1+1+1+1+2+2+1+2+3)/15= 1.33

Analysis 2: Maximal groups of authors who are all mutually connected

Number of cliques of size 3: 5

Number of cliques of size 4: 498

Number of cliques of size 5: 792

Clique means that every element of the graph is connected to each other. Looking at the number of cliques of different sizes, it can be assumed that the implies that most papers were written by the group of 5 authors. However, very less papers were published by the group of 3 authors.

Analysis 3a: Top 10 papers that are cited most from these two years

b944f77f-113b-4a02-ae5e-d4a124b8fd5b

c1b6b493-01ef-420f-be44-7bacfe34e846

6a6b9aa6-683f-4c7c-b06e-9c3018d10fd3

f6bd8b64-684d-429a-aab5-8ff3a2c23cd6

8026f56a-a93e-4933-8ead-c9aa9e3f0498

dd83785a-dd19-41e3-9b25-ebabbd48d336

9d912297-e52f-4ab6-add4-633e0f263933

f56b877b-4060-4754-b303-e8140968544c

50dd56db-151d-4d62-8576-65f0ef6f381b

d3e00e7e-1c64-4d7a-b2b2-1ad98ba4c706

The above mentioned are the paper ids of the top 10 papers that are cited most from the two years. It can be concluded that these papers were referred in most of the papers.