**DASC 5300 Foundation of Computing**

**Project-2 DBLP Data Analysis Using Graph Characteristics**

**Overall Status:**

Initially we have thoroughly gone through the description of the project and tried to understand the requirement for the given DBLP Data. After the clear analysis of the project, Then, figured out which data is needed to plot the graphs and found the maximum cliques and centrality of graphs. Generated graph characteristics with given Network summary, MLN\_IO, MLN files.

**File Descriptions:**

Out of the 4 json files we created new input.txt file which will be holding the data of the nodes and created the output.txt file which will holds the graph characteristics data. Out of these there were also three output png files which was used for the degree vs number of nodes.

**Division of Labor:**

We divided the equal amount of work initially and started the project work. Divided the complete project into two equal parts for both coding and report

1. Concatenating the Data frames and creating the graph Data structure.
2. Generating the Graph Characteristics
3. Report

For the completion of this project, it took around 10 days.

**Problems encountered and handling:**

1. Handling large data was problem as data is crashed very often. By using the chunk method in Pandas and upgrading the resources was helpful.
2. When we tried to concatenate all the data frames we could not do it because of ram issues
3. For the Graph characteristics, the conversion of string to int was difficult.

**Analysis 0:**



* From the above table we can see that the number of nodes is at the highest place for Analysis 3a - paper citation and the lowest is analysis 3b – conference (Author - Venue)
* Since the number of nodes are high the number of edges is also higher for the 3a - paper citation and lower for the analysis 3b – conference (Author - Venue). Similarly for the density analysis too.
* But when we compare the average degree, the known authors have the highest value (3.287536084) and the lowest for the conference (1.520685901)
* We analyzed this pattern because the number of venues is less than number of authors, as generally authors will more.
* Connected components are deleted from this table as the data is large.

Chart, line chart, histogram

Description automatically generatedChart

Description automatically generatedChart, line chart

Description automatically generated

* For the Analysis 2 – known Authors and Analysis 3b – Author and Venue nodes are evenly distributed when compared with the Analysis 3a – Paper Citation graph.
* Among the three analysis the highest degree and the number of nodes is presented in the Analysis 3a – Paper Citation graph.
* The lowest degree and the number of nodes is presented in 3b – Author and Venue.

**Analysis 1:**

* For the manual calculation verifications, we verified with the sample size of 2, 4, 8 and calculated number of cliques. Mentioning the graphs below for the sample size 2 for which we got 6 cliques for the size 3 and 0 for size 4,5.

Shape, polygon

Description automatically generated

* Degree centrality = n/N-1 where n = number of connected nodes and N = Total number of nodes.
  + For the sample size of 2 the degree centrality is

= 1/17-1 = 0.0625

A picture containing sky, outdoor, day

Description automatically generated

From the code for the verification:

Graphical user interface, text, application

Description automatically generated

**Analysis 2:**

Chart, scatter chart

Description automatically generated

* Graph generated which will connect pair of Authors in the paper, where each author is considered as node and will connect to another author.
* It is an undirected graph and removed the self-looping edges
* By using NetworkX and matplotlib packages for plotting the graph
* As part of preprocessing, we have dropped Nan values from Authors column .
* We took 500 samples and plotted the graph for clear visualization
* As part of analysis the maximum group of authors, mutually connected i.e., Cliques

Number of cliques for size 3-407

Number of cliques for size 4-165

Number of cliques for size 5-47

**Analysis 3 a:**

A picture containing pie chart

Description automatically generated

* Graph which is generated is Paper Citation graph where each paper has the unique ID and mapped it to reference.
* This is directed graph from the citing the paper
* We have dropped Nan values from ID and references.
* We took 40 samples and plotted the graph because if the data is large visualization of graph is not clear.

**Analysis 3 b:**

Chart, scatter chart

Description automatically generated

* Constructed a graph with the author whose papers are published in Conference.
* Considered Author and Venue from the Dataframe and plotted the data.
* This is undirected graph from each author of a paper to Venue where the paper is published.
* We have dropped Nan values from Venue and authors
* We took 40 samples and plotted the graph for clear visualization
* For the maximum number of authors who have published most papers in dataset is 0.001389

**Top 5 Authors who have published most papers:**

Table

Description automatically generated

* For all top 5 authors we can observed the similar centrality which has 0.001389.

**Conclusion:**

We analyzed how authors are being socialized with each other for pushing the paper with the citations. We handled large data set and relation between Author and paper citations with the help of Graph Data structures and Graph characteristics.