DASC-5300-001-Foundation of Computing

Project-2: DBLP Data Analysis Using Graph Characteristics

Overall Status:

After carefully reading the project-2 description file and the problem statement, we have developed a general understanding of the project. In order to keep on schedule, we have established the initial expectations for each team member's contribution to the project and allocated timestamps for each task. We generated the necessary graphs, carried out all the analyses, and the resulting report stands as our project.

File Descriptions:

We have created files named graph1_pairs.csv, graph2_pairs.csv, graph3_pairs.csv to insert the data containing pre-processed data of Known-Authors, Paper Citations, publish papers in conference and generate graphs respectively. Generated networkx_output.txt, networkx_output_layer0.png, networkx_output_layer2.png, networkx_output_layer2.png from the NetworkX package to store the graph characteristics.

Division of Labor:

Based on the difficulty and amount of time needed, we ensured that the tasks were split fairly. The project's coding and analysis portions were both divided into manageable tasks, and each work was given a time limit.

1. **Rohith Kumar Nayakanti (1002024866)** – Performed the initial analysis together, generated all the 3 graphs. Worked on Analysis 0 and Analysis 2.
2. **Sasank Kinnera (1001874178)** – Worked on Analysis 1 and Analysis 3a after performing the initial analysis together. Graphs and data were analyzed, and a final report was prepared.

It took about 36 hours to complete the initial understanding and coding phase. Over 12 hours were spent on the report's approach and analysis.

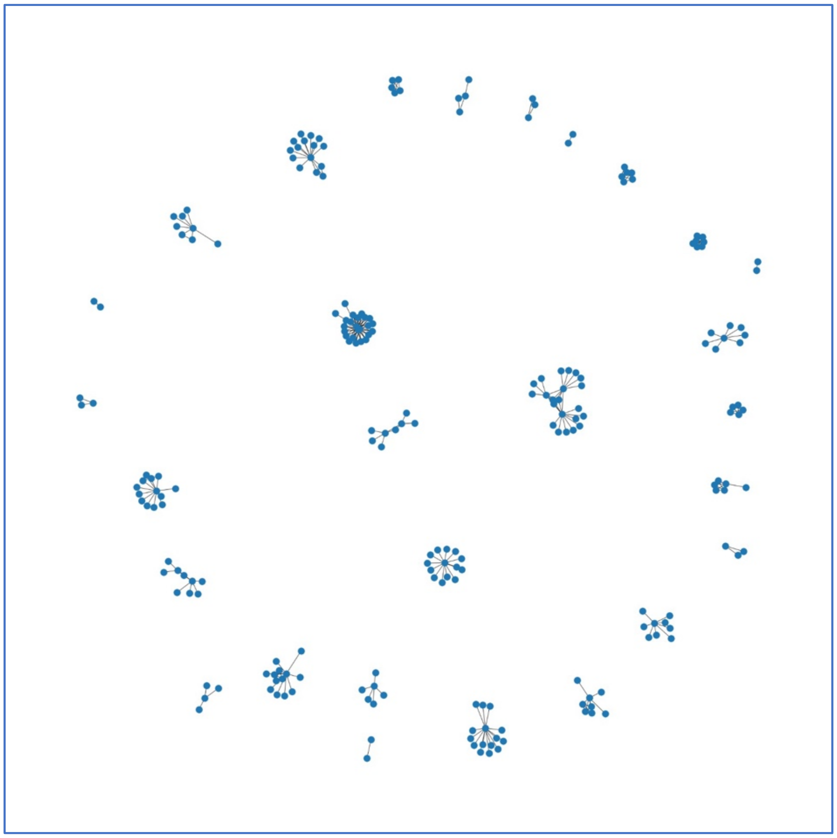
Problems encountered and handling:

1. Since this was the first time dealing with graphs and sizable data sets, it required some time to fully comprehend the project. We were able to understand and finish the project, nonetheless, thanks to the professor's helper slides and the support of the TAs.
2. Due to the large size of the data set, we encountered RAM availability issues while processing the graph features for the full data set. We improved the code and fixed the RAM availability problem by getting a little forward in the pre-processing.

3. During the generation of graph of authors who have published a paper in conference, we have faced an issue in fetching the venue and author details in a dictionary. The duplicate venue details have been ignored which is resulted in a smaller number of records. Then, we have come up with an idea of writing data onto a file and it worked well.

I. Generation of Known-Authors Graph:

- The graph is generated from the data frame “sample_df”, which is having 20k records with seed value as Date of Birth.
- The graph is generated with the randomly selected sample of 100 unique records over the sample data set.
- The generated graph is an undirected graph.

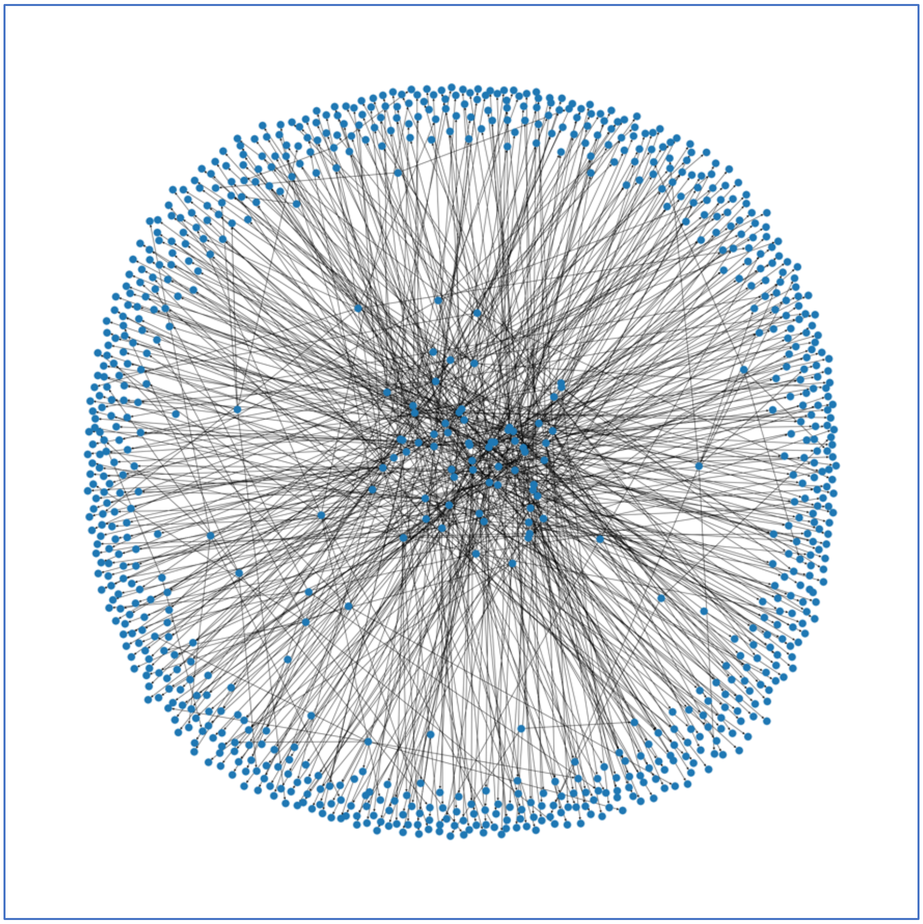


Observations:

1. It can be observed that the greater number of common authors are placed in the center of the graph.
2. The authors with less connections are situated around and away from the center of the graph.

II. Generation of Paper Citation Graph:

- The graph is generated with the randomly selected sample data frame of 20k records with seed value as Date of Birth.
- The graph is generated with the randomly selected sample of 100 unique records over the sample data set.
- We have plotted the graph using the unique values of 'id', 'references.
- Constructed a paper citation graph, from the papers citing another paper internally.
- We have generated a directed graph.

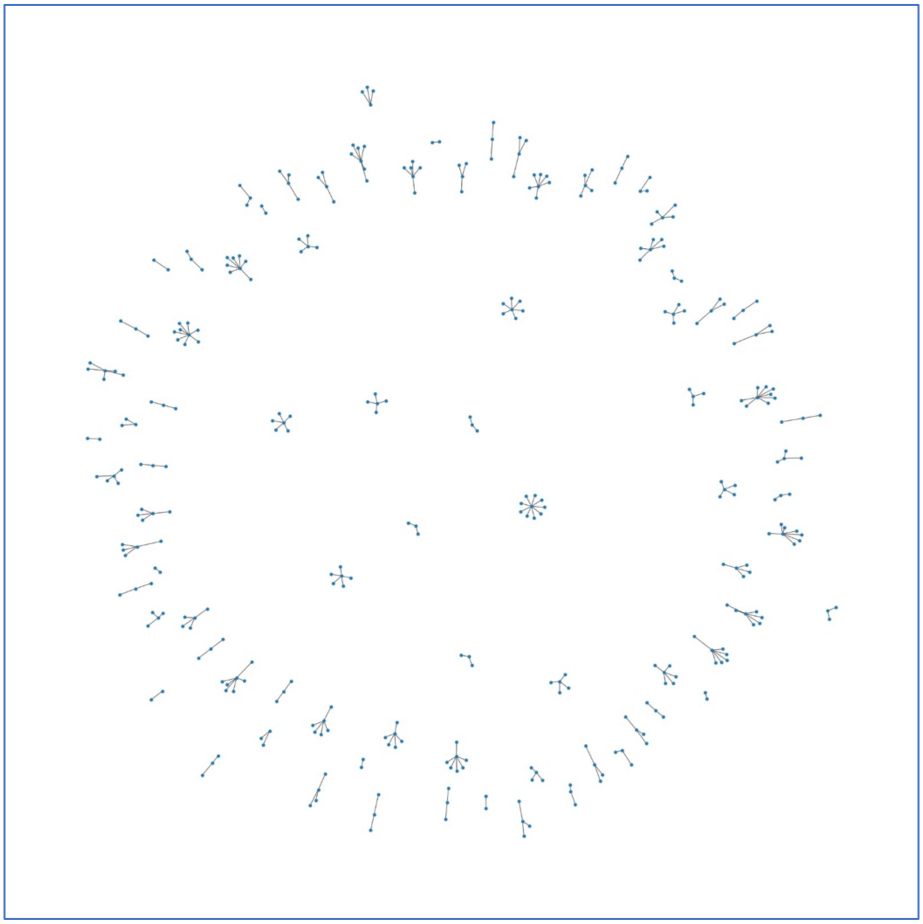


Observations:

1. It can be observed that almost every paper had been cited in one or more papers.
2. The graph has been generated in a circular pattern.
3. We can observe that the main papers which have most citations are situated in the center of the graph and all the edges are popping outwards to other papers.

III. The Graph of Authors – Published a Paper in a Conference:

- We have run the data on the sample of 20k records using seed value as Date of Birth and the graph is generated with the randomly selected sample of 100 unique records.
- We have plotted the graph using the unique values of 'venue', 'authors'
- The generated graph is an undirected graph.



Observations:

1. It can be observed that almost every venue had been utilized to publish one or more than one paper in a conference.
2. The graph has been generated in a circular pattern.
3. In the above graph, couple of venues have hosted multiple conferences and are situated towards the center of the graph.

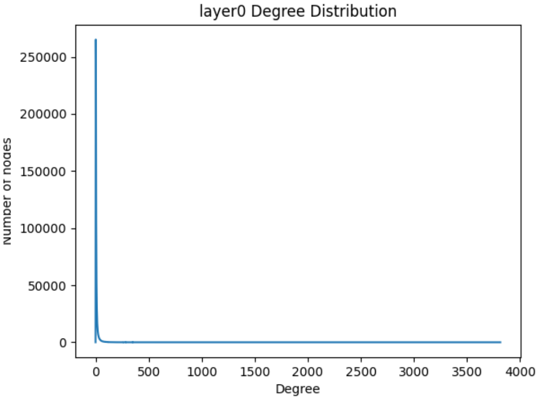
Graph Analysis

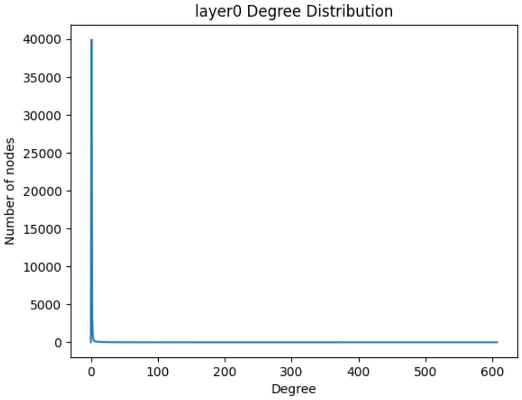
Analysis 0: Generating Graph Characteristics

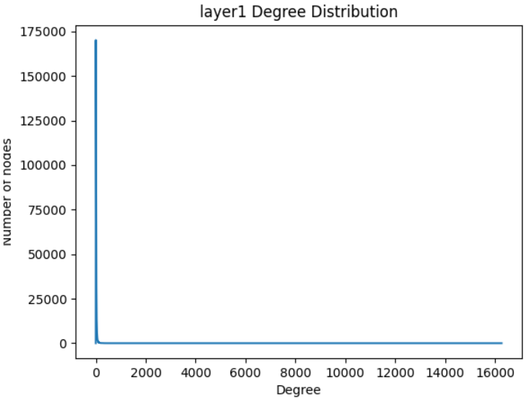
- In this analysis we run the network characteristics package on all graphs to derive the network characteristics.
- The graph is between Number of Nodes and Degree.
- Have generated Graph Characteristics for “known-authors”, “paper citation” and “published paper in a conference” graphs.
- From layer0 Degree Distribution graphs, it can be observed that over 250k authors have less than 30 known authors.
- From layer1 Degree Distribution graphs, it can be observed that over 150k papers have less than 20 citations.
- From layer2 Degree Distribution generated by 20k records, it can be observed that over 100k venues have hosted around 15 conferences.

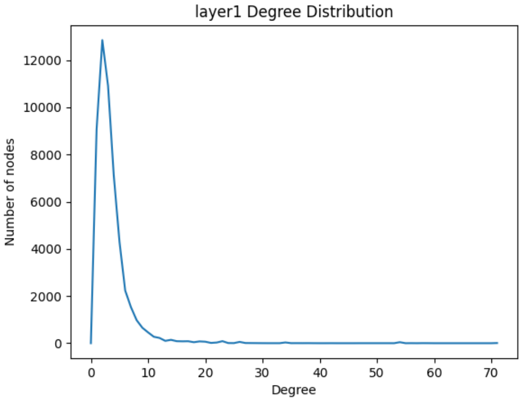
Here is the Degree distribution for the entire data set.

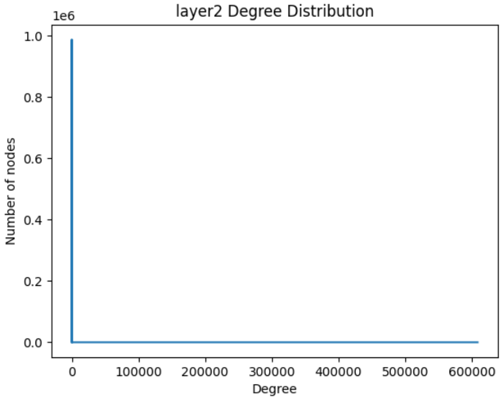
Here is the Degree distribution for the sample of 20k records using seed value.

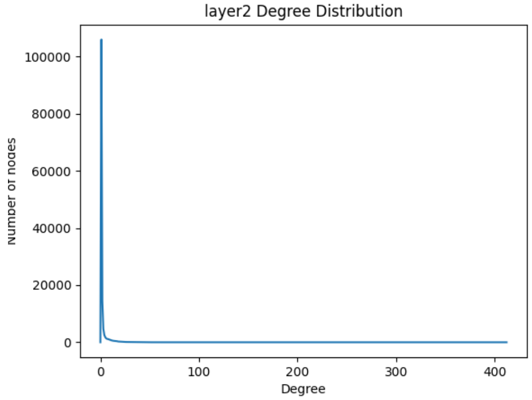












Here is the tabular network characteristics report of all three graphs created over the entire data set:

	known-authors	paper citation	published paper in a conference graph
number_of_nodes	1729923	2725533	1740240
number_of_edges	8546419	25115566	5386112
density	5.711640730447979e-06	6.7619232234567766e-06	3.5570283905028744e-06
number_of_connected_comp	36133	7820	215
diameter	-1	-1	-1
min_degree	1	1	1
max_degree	3815	16260	608721
avg_degree	9.880692955698029	18.429838127074593	6.190079529260332
std_dev_degree	23.93176662276831	50.0027202694069	481.1150181584455

Here is the tabular network characteristics report of all three graphs created over the Sample of 20k records in data set:

	known-authors	paper citation	published paper in a conference graph
number_of_nodes	45917	51601	140991
number_of_edges	48734	93342	164895
density	4.623006476441811e-05	7.011311751214484e-05	1.6590435304212102e-05
number_of_connected_comp	786	11100	3490
diameter	-1	-1	-1
min_degree	1	1	1
max_degree	607	71	412
avg_degree	2.122699653723022	3.617836863626674	2.339085473540864
std_dev_degree	9.62316674643089	3.5315788653222313	4.667364318098201

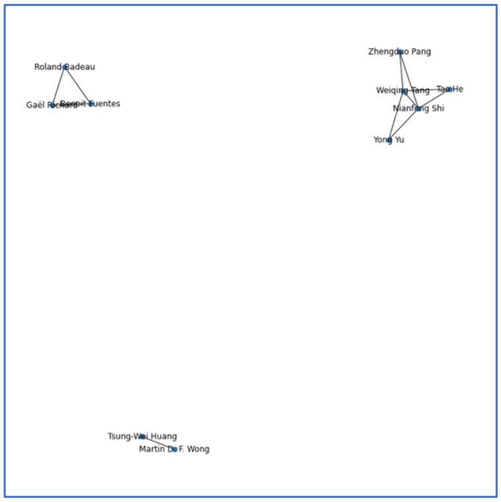
Comments/Remarks: Upon observing the table generated on 20k records, below are the comments

- The std_dev_degree is relatively high for known-authors graph, which make us understand that the number of known authors is being varied with the deviation of 9.62316674643089.
- The density is higher for Paper Citation graph, which means the density representing the no of paths per given area are higher in the Paper Citation graph.
- The min_degree having value 1is similar with all three graphs. However, the max_degree is highest in known-authors graph.

Analysis 1: Showing the Ground Truth

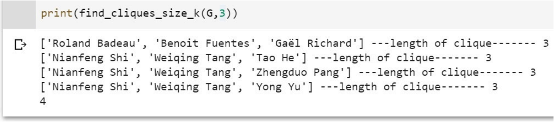
We have performed the analysis on Sample data and included the manually verified values.

a. **For Analysis 2,** generated a sample graph with 8 records and clique size of 3 to double check the data manually. Here is the graph generated.



i. As we can observe in the graph, it has 10 nodes.

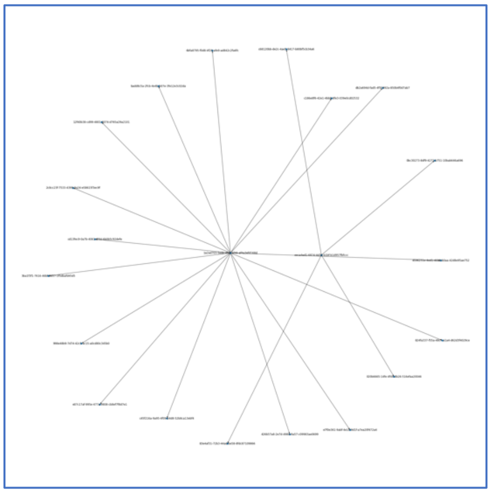
ii. It can be manually observed that there are 4 cliques of clique size 3, which is same as the value provided by code.



```
print(find_cliques_size_k(G,3))
```

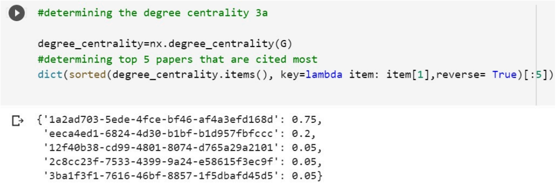
```
['Roland Badeau', 'Benoit Fuentes', 'Gaël Richard'] ---length of clique------- 3
['Nianfeng Shi', 'Weiqing Tang', 'Tao He'] ---length of clique------- 3
['Nianfeng Shi', 'Weiqing Tang', 'Zhengduo Pang'] ---length of clique------- 3
['Nianfeng Shi', 'Weiqing Tang', 'Yong Yu'] ---length of clique------- 3
4
```

b. **For Analysis 3a,** we have generated sample graph for 2 records along with paper citations.



i. As we can observe in the graph, node1 has 15 citations and node2 has 4 citations.

ii. The top 2 values of degree centrality have been manually observed from the below table, which the values are same as the values provided by code.



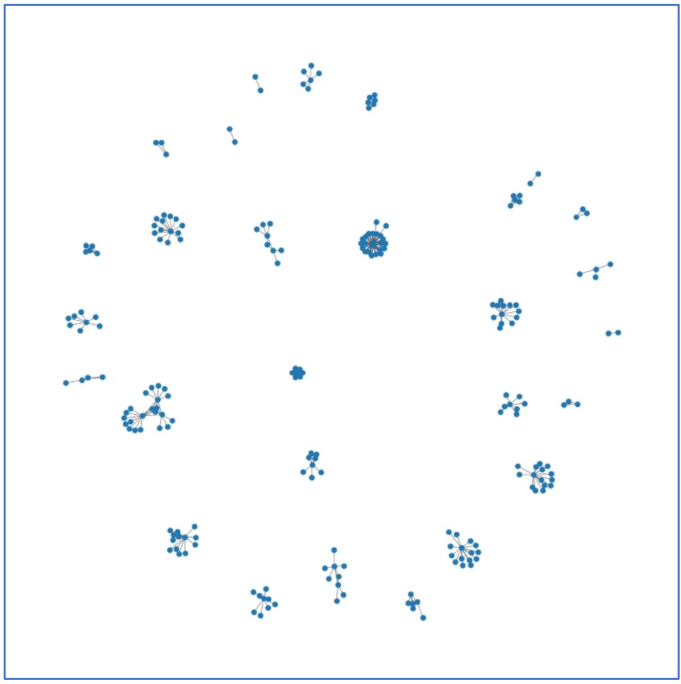
```
#determining the degree centrality 3a

degree_centrality=nx.degree_centrality(G)
#determining top 5 papers that are cited most
dict(sorted(degree_centrality.items(), key=lambda item: item[1],reverse= True)[:5])
```

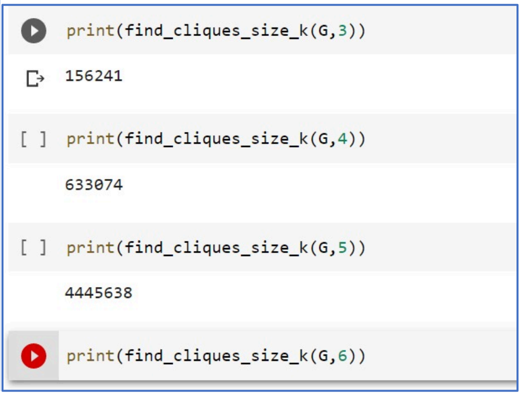
```
{'1a2ad703-5ede-4fce-bf46-af4a3efd168d': 0.75,
 'eeca4ed1-6824-4d30-b1bf-b1d957fbfccc': 0.2,
 '12f40b38-cd99-4801-8074-d765a29a2101': 0.05,
 '2c8cc23f-7533-4399-9a24-e58615f3ec9f': 0.05,
 '3ba1f3f1-7616-46bf-8857-1f5dbafd45d5': 0.05}
```

Degree Centrality		
Node	**Score**	**Standardized Score**
eeca4ed1-6824-4d30-b1bf-b1d957fbfccc	15	15/20 = 0.75
1a2ad703-5ede-4fce-bf46-af4a3efd168d	4	4/20 = 0.2

Analysis 2: Finding maximal group of authors who are mutually connected.

- We have run the data on the sample of 20k records using seed value as Date of Birth.
- Considered random 100 records out of 20k records and generated the graph.





```
print(find_cliques_size_k(G,3))
156241
```

```
[ ] print(find_cliques_size_k(G,4))
633074
```

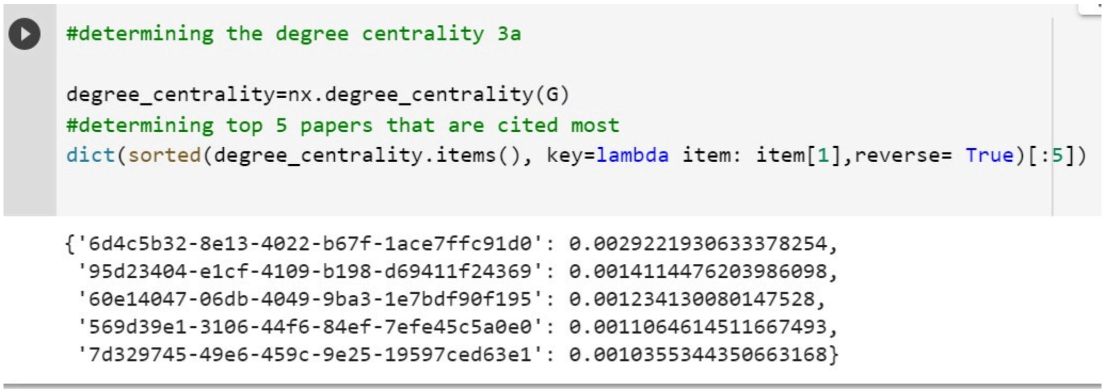
```
[ ] print(find_cliques_size_k(G,5))
4445638
```

```
print(find_cliques_size_k(G,6))
```

- From the above graph, the number of authors who are mutually connected is in increasing fashion from top to bottom.
- Calculated the count of hits for Clique size 3, 4, 5 but for the Clique size 6, the RAM in going out of memory on Colab.
- Out of the result, it can be observed that the count is higher for the clique size 5 and least is with the clique size 3.
- Therefore, for the clique size 5, the authors are forming more connections with the co-authors in a circular loop as compared to the other clique sizes

Analysis 3a: Finding top 5 papers that are cited most from the paper citation graph

- We have generated the degree centrality in decreasing order by considering 20k sample records using seed value as Date of Birth.
- We have picked the top 5 records having higher degree centrality from the rest and posted the results as below:



```
#determining the degree centrality 3a

degree_centrality=nx.degree_centrality(G)
#determining top 5 papers that are cited most
dict(sorted(degree_centrality.items(), key=lambda item: item[1],reverse= True)[:5])
```

```
{'6d4c5b32-8e13-4022-b67f-1ace7ffc91d0': 0.0029221930633378254,
 '95d23404-e1cf-4109-b198-d69411f24369': 0.0014114476203986098,
 '60e14047-06db-4049-9ba3-1e7bdf90f195': 0.001234130080147528,
 '569d39e1-3106-44f6-84ef-7efe45c5a0e0': 0.0011064614511667493,
 '7d329745-49e6-459c-9e25-19597ced63e1': 0.0010355344350663168}
```

- We have manually checked the values by taking a sample of two records and double checked our work in finding top 5 papers that are cited most from the paper citation graph. The detailed explanation on manual check process has been provided in the section Analysis 1(a).
- We went ahead and checked the count of citations of top 5 papers

Id	Title	Citations
6d4c5b32-8e13-4022-b67f-1ace7ffc91d0	Using formal specifications to support testing	412
95d23404-e1cf-4109-b198-d69411f24369	Expertise Retrieval	198
60e14047-06db-4049-9ba3-1e7bdf90f195	Dynamic object process graphs	174
569d39e1-3106-44f6-84ef-7efe45c5a0e0	Bibliography on cyclostationarity	156
7d329745-49e6-459c-9e25-19597ced63e1	Parametric optimization of storage systems	146

Conclusion:

From the given data, we have preprocessed and stored the data in respective files. We have considered 20k records of data from the created files which were huge and performed multiple analysis stated in the project description.