**DBLP Data Analysis Using Graph Characteristics**

**Overall Description:**

Extracted the JSON files from the given dblp\_v10.zip file. And read the 4 different JSON files separately using chunks. Later concatenated this separately read JSON files into a data frame. Done preprocessing by only considering one year which is 1998 year. Tried to build the graph on that year’s data and graph characteristics were also done on three different graphs.

**Known Author Graph:**

In the known author graph we extract two columns from the dataset. That is the author and title columns. Based on these two columns, we constructed the known author graph which is an undirected graph.

Chart

Description automatically generated

The problem we faced was we can’t make it with dictionary as given in helper slide. We construct it by putting manual values.Its because in dataframe there are 42k values and frame does not support this huge values and we cannot plot graph for huge values.

**Paper Citation Graph:**

For the second graph we extracted two columns that is id and references from the dataset, and this graph is a directed graph.

Diagram

Description automatically generated with low confidence

**Author Venue Graph:**

For this graph, we extracted author and venue columns from the given dataset, after extracting these columns, we convert them into a list and construct an undirected graph.

A picture containing radar chart

Description automatically generated

**DIVISION OF LABOR:**

Data preprocessing and reporting were done by Akshata Agine and building the graphs by done by Bharath. It nearly took us a whole month to complete the project and report.

**Problem Encounter & how you handle them:**

Since the data set was way too large for reading the JSON files using pandas made colab crashed several times. So, we used chunks to read data sets and merged them by concatenating them. Another problem we faced was we can’t make it with a dictionary. We construct it by putting manual values. It's because in the data frame it had 60kvalues and building the graph on these values has given output that is both clustered and overlapped.

**Analysis 0:**

**Table-1:**

|  |  |  |
| --- | --- | --- |
| Known Author  Graph | Citation Graph | Author Venue  Graph |
| Characteristics:  MinDegree = 1  MaxDegree = 2  AvgDegree = 1.333  StdDeviationDegree = 0.5773  Degree = [1,2,1] | Characteristics:  MinDegree = 1  MaxDegree = 3  AvgDegree = 2  StdDeviationDegree = 0.666  Degree = [1, 2, 2, 3, 2, 2, 1, 2, 3, 2] | Characteristics:  MinDegree = 1  MaxDegree = 4  AvgDegree = 1.5  StdDeviationDegree = 1  Degree = [1, 4, 1, 1, 1, 1, 2, 1, 1, 3, 1, 1] |

**Analysis 1:**

Instead of taking sample data by using a random state, we analyzed a particular year's data.

**Analysis 2:**

In analysis 2, we construct the undirected graph, and we calculate the authors' frequency that shows how many authors are mutually connected. We use 3 authors for 1 paper. We construct the undirected only 3 graphs for 3 papers according to the given information.

**AUTHORS MUTUALLY CONNECTED TO EACH OTHER:**

Authors have a relation with themself because there were only 5 to 6 authors, and we use different ids.

**Analysis 3(a):**

We work only on 1 year’s dataset extracted from the whole dataset.

**Most Cited Papers:**

1. BOUNDS FOR THE RANGE OF A BIVARIATE POLYNOMIAL
2. Reasoning about Set-Oriented Methods in Object
3. Some problems in an approximation of a set of point
4. Parallel algorithms for deblurring of MR images
5. Intensity Calibration for Stereo Images Based

**Analysis 3(b):**

**Author’s Name:**

1. 'Makoto J. Hirayama',
2. 'Taro Sugahara',
3. 'Zhiyong Peng',
4. 'Junichi Yamazaki'
5. ,'Giovanna Guerrini',
6. 'Isabella Merlo'
7. ,'P. K. A. Menon'

**Conclusion:**

In this dataset, we can use different data frame columns and then perform analysis. First of all, we extract the dataset from the source and the preprocess and convert it into a data frame. After all, we have done all tasks given in this project2. In mentioned task, we create three different kinds of directed and undirected graphs. And then after creating a graph we check the author’s mutual connection. How many authors are mutually connected and publications of papers and many more.