



VIT[®]
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

School of Computer Science and Engineering
(WINTER 2022-2023)

Student Name: Ketan Kolte

Reg No: 22MCB0016

Email: ketansanjay.kolte2022@vitstudent.ac.in

Mobile: 8329324714

Faculty Name: DURGESH KUMAR

**Subject Name with code: SOCIAL NETWORK ANALYTICS LAB –
MCSE618P**

Date : 14/06/2023

Analysis of Bibliographic Graph and Visualization of Heterogeneous Network

ABSTRACT

This report presents an analysis of a bibliographic graph representing a heterogeneous network. The aim is to plot the bibliographic graph or its subgraph and visualize it for further examination. The graph consists of nodes and edges, with each node representing a topic or paper in the network. The analysis involves classification of nodes and edges, and clustering using a preferred method. The results provide insights into the structure and relationships within the bibliographic graph.

INTRODUCTION

The availability of vast amounts of scholarly data has led to the emergence of bibliographic graphs as a means to explore the connections and relationships between research topics or papers. Heterogeneous networks, which incorporate different types of nodes and edges, are particularly useful in representing such bibliographic graphs. This report aims to analyze a bibliographic graph by plotting it or its subgraph and visualizing it to gain a deeper understanding of the network.

DATASET DESCRIPTION

The "cwurData.csv" dataset consists of following columns :

1. world_rank: Represents the university's rank in the world according to the Center for World University Rankings.
2. institution: The name of the university.

3. country: The country where the university is located.
4. national_rank: The ranking of the university within its own country.
5. quality_of_education: This column represents the quality of education offered by the university based on a range of factors such as the number of alumni who have won major international awards, prizes or medals relative to the university's size, faculty members who have won major international awards, prizes, or medals, etc.
6. alumni_employment: This column represents the number of alumni who have held CEO positions at the world's top companies compared to the university's size.
7. quality_of_faculty: The quality of the faculty members working at the university based on various factors such as the number of faculty members who have won major international awards, prizes or medals, publications, and citations.
8. publications: The number of research publications produced by the university.
9. influence: This column represents the influence of the research produced by the university based on the number of research papers appearing in highly influential journals.
10. citations: The number of times the university's research has been cited by others.
11. broad_impact: This column represents the university's overall performance based on several indicators.
12. patents: The number of international patent applications filed by the university.
13. score: The total score of the university based on all the factors mentioned above.
14. year: The year for which the rankings were published.

METHODOLOGY

Part A : Plot bibliographic graph(or its subgraph) for a heterogeneous network and visualize it

- 1. Dataset Preparation:** In this step, the dataset is prepared for analysis. This typically involves loading the dataset using a library like pandas and performing data cleaning operations. For example, removing rows with NaN or null values, removing duplicates, and resetting the index. The cleaned dataset may also be saved to a new file for future use.
- 2. Data Extraction:** In this step, relevant data is extracted from the dataset for constructing the graph. This could involve selecting specific columns or features that are necessary for creating the graph. The extracted data is typically stored in a suitable data structure to be used in the subsequent steps.
- 3. Graph Construction:** The graph construction step involves creating a graph data structure that represents the relationships or connections between different entities. In the context of the provided code, a directed graph is constructed using the NetworkX library. Nodes in the graph represent topics or papers, while edges represent the connections or relationships between them (e.g., citations). The relevant data extracted in the previous step is used to populate the graph with nodes and edges.
- 4. Graph Visualization:** Once the graph is constructed, it can be visualized to gain insights into its structure and relationships. Graph visualization tools, such as matplotlib or network visualization libraries like Gephi or Cytoscape, can be used for this purpose. The visualization typically involves defining a layout for the graph, setting up the plot with appropriate dimensions, and drawing the graph with node and edge attributes. The resulting visualization provides a visual

representation of the graph, allowing for easier interpretation and analysis.

Part B : Node and edge classification Number of topics or papers (similarity in bibliographic graph like topic modeling(just an eg do not do) Use any classifier or clustering(preferable)

- 1. Data Preparation:** Data preparation involves cleaning and preparing the dataset for further analysis. This step typically includes removing rows with missing or null values, handling duplicates if necessary, and ensuring the dataset is in a suitable format. Data cleaning ensures the dataset is of high quality and ready for subsequent analysis steps.
- 2. Node Feature Extraction:** Node feature extraction involves extracting relevant features or attributes from the graph's nodes. In the provided code, the number of topics or papers is extracted as a node feature. This step calculates the desired feature for each node in the graph and organizes them into a suitable data structure for further analysis.
- 3. Text Processing and Vectorization:** Text processing and vectorization are performed on the textual data associated with the nodes (e.g., abstracts or document titles). Text processing techniques, such as removing stop words or performing stemming, can be applied to clean the text data. Vectorization, often using techniques like TF-IDF (Term Frequency-Inverse Document Frequency), converts the text into numerical representations that machine learning algorithms can process.
- 4. Dimensionality Reduction:** Dimensionality reduction techniques are employed to reduce the dimensionality of the data while preserving relevant information. In the provided code, Truncated SVD (Singular Value Decomposition) is used for dimensionality reduction. It transforms

the high-dimensional data into a lower-dimensional representation while capturing as much information as possible.

5. **Clustering:** Clustering algorithms are applied to group similar data points together based on their features or attributes. In the given code, K-means clustering is used. K-means is an unsupervised learning algorithm that partitions data into a predefined number of clusters based on their similarity. The algorithm iteratively assigns data points to clusters to minimize the within-cluster sum of squares.
6. **Cluster Visualization:** Cluster visualization aims to visually represent the clusters obtained from the clustering step. In the provided code, scatter plots are used to visualize the clusters. Each data point is plotted based on its reduced dimensional representation, and different clusters are distinguished by using different colors or markers. This visualization allows for an intuitive understanding of the clustering results and patterns in the data.
7. **Variable Distribution Evaluation:** Variable distribution evaluation involves analyzing the distribution of selected variables in the dataset. This step helps in understanding the frequency or occurrence of different values within each variable. The count plot is a commonly used visualization tool for this purpose. It displays the count or frequency of each unique value in a categorical variable. This evaluation helps to identify patterns, imbalances, or anomalies in the variable distributions.

IMPLEMENTATION

Part A :

1. Dataset Preparation:

- Load the `cwurData.csv` dataset using `pandas`.
- Remove rows with `NaN` or null values.
- Remove duplicates from the dataset.
- Reset the index of the cleaned dataset.
- Save the cleaned dataset to a new file.

2. Data Extraction:

- Extract relevant columns for bibliographic connections, such as 'country' and 'citations'.
- Create a set of unique nodes by combining the 'country' and 'citations' columns.

3. Graph Construction:

- Create an empty directed graph using the `NetworkX` library.
- Iterate over the rows of the extracted data.
- For each row, extract the paper title and citation.
- Add the paper title and citation as nodes to the graph.
- Add an edge between the paper title and citation.

4. Graph Visualization:

- Define the layout for visualizing the graph using the spring layout algorithm.
- Set up the figure and plot using `matplotlib`.
- Draw the `networkx` graph with labels, node size, font size, and edge color.
- Set a title for the graph visualization.
- Turn off the axis display. Display the graph visualization.

Part B :

1. Data Preparation:

- Load the dataset using pandas.
- Remove rows with NaN or null values.
- Select relevant columns for clustering and evaluation. Define the columns to evaluate health on.

2. Node Feature Extraction:

- Extract the node features, such as the number of topics or papers, from the graph.
- Remove rows with NaN or null values from the dataset.

3. Text Processing and Vectorization:

- Prepare the data for clustering by selecting the relevant feature(s) (e.g., abstracts) from the dataset.
- Apply text processing techniques, such as removing stop words, using the TfidfVectorizer from sklearn.
- Vectorize the abstracts using TF-IDF to represent them as numerical features.

4. Dimensionality Reduction:

- Reduce the dimensionality of the TF-IDF matrix using Truncated SVD.
- Perform dimensionality reduction to two components to enable visualization.

5. Clustering:

- Apply K-means clustering on the reduced data using the KMeans algorithm from sklearn.
- Set the desired number of clusters. Fit the clustering model to the reduced data and obtain cluster labels.

6. Cluster Visualization:

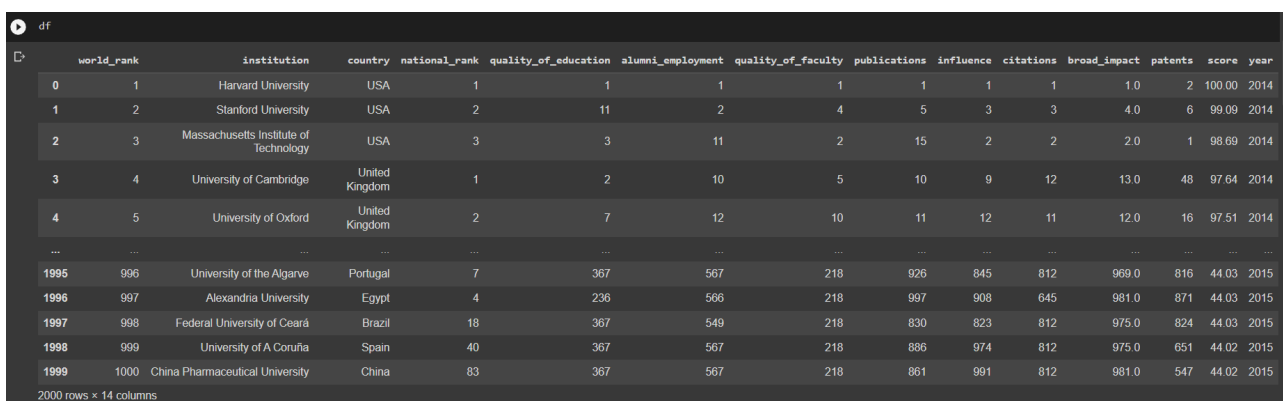
- Plot the clusters using a scatter plot, with the reduced data points as coordinates and the cluster labels as colors.
- Set appropriate labels for the x-axis, y-axis, and the title of the plot.
- Display the scatter plot to visualize the clustering results.

7. Variable Distribution Evaluation:

- Evaluate the distribution of selected variables using count plots.
- Create a subplot grid with the number of subplots equal to the number of selected variables.
- For each variable, plot a count plot using the seaborn library to visualize the distribution.
- Set appropriate labels for the x-axis, y-axis, and the title of each subplot.
- Display the subplot grid to visualize the distribution of variables.

RESULTS

The analysis of the bibliographic graph have the following results:



	world_rank	institution	country	national_rank	quality_of_education	alumni_employment	quality_of_faculty	publications	influence	citations	broad_impact	patents	score	year
0	1	Harvard University	USA	1	1	1	1	1	1	1	1.0	2	100.00	2014
1	2	Stanford University	USA	2	11	2	4	5	3	3	4.0	6	99.09	2014
2	3	Massachusetts Institute of Technology	USA	3	3	11	2	15	2	2	2.0	1	98.69	2014
3	4	University of Cambridge	United Kingdom	1	2	10	5	10	9	12	13.0	48	97.64	2014
4	5	University of Oxford	United Kingdom	2	7	12	10	11	12	11	12.0	16	97.51	2014
...
1995	996	University of the Algarve	Portugal	7	367	567	218	926	845	812	969.0	816	44.03	2015
1996	997	Alexandria University	Egypt	4	236	566	218	997	908	645	981.0	871	44.03	2015
1997	998	Federal University of Cear�	Brazil	18	367	549	218	830	823	812	975.0	824	44.03	2015
1998	999	University of A Coru�a	Spain	40	367	567	218	886	974	812	975.0	651	44.02	2015
1999	1000	China Pharmaceutical University	China	83	367	567	218	861	991	812	981.0	547	44.02	2015

2000 rows x 14 columns

Fig.1. Shows the columns present in the dataset

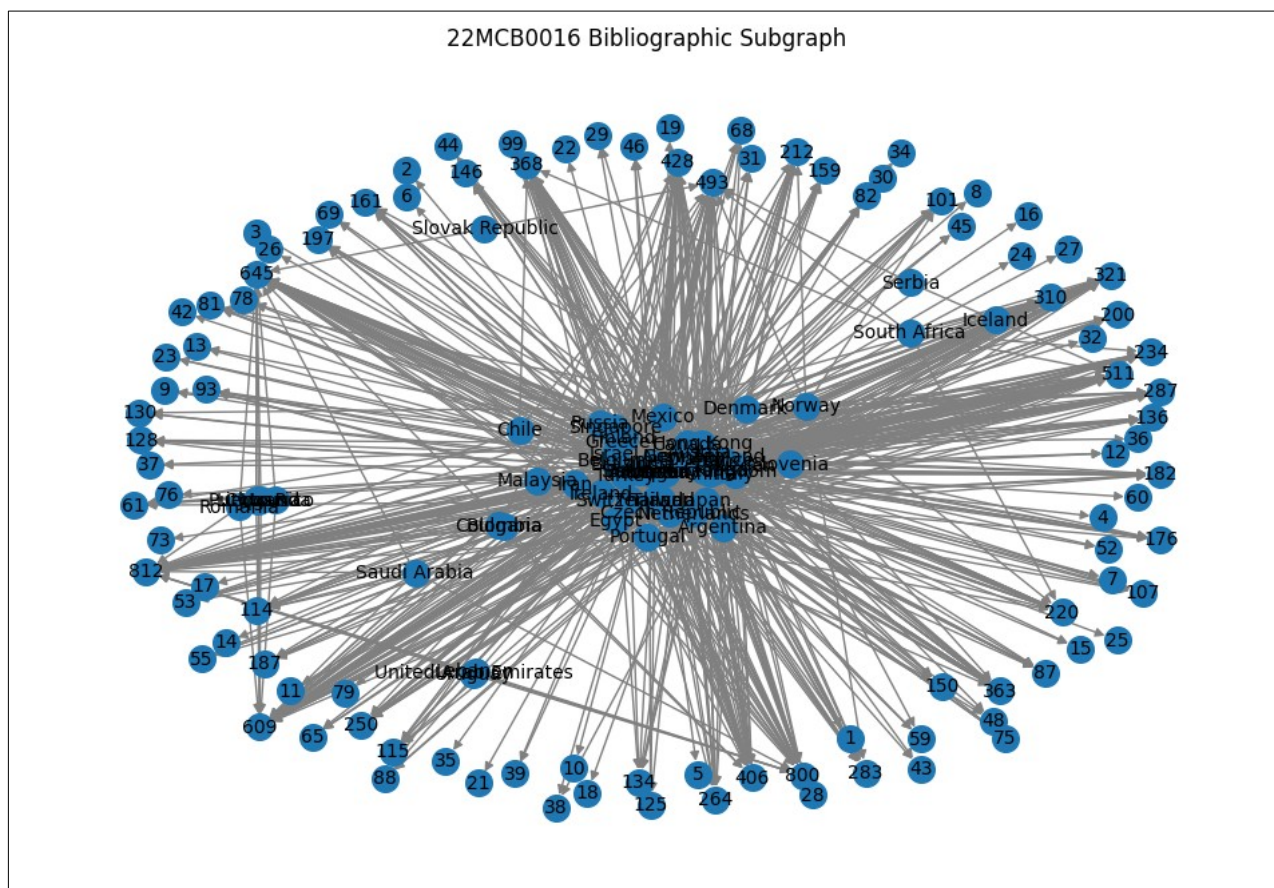


Fig.2. Shows the resulting bibliographic subgraph having nodes and edges based on the paper titles and their corresponding citations

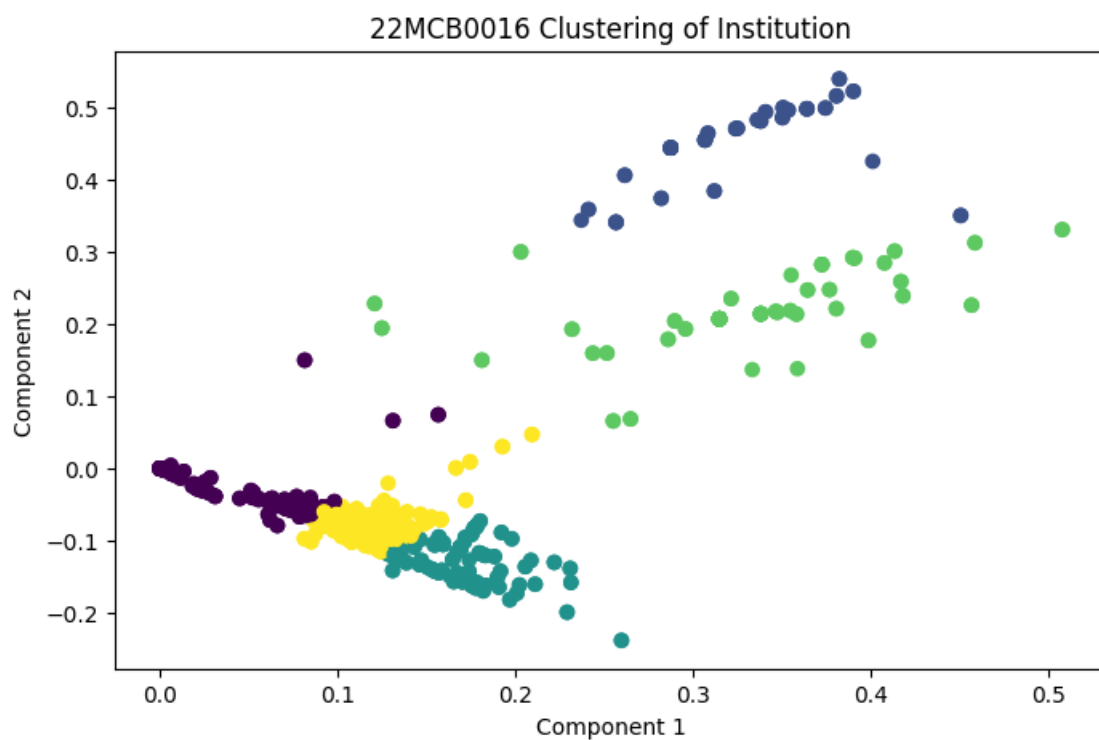


Fig.3. Shows K-means clustering on the Science Direct papers using TF-IDF vectorization and Truncated SVD dimensionality reduction, and visualizes the resulting clusters in a scatter plot.

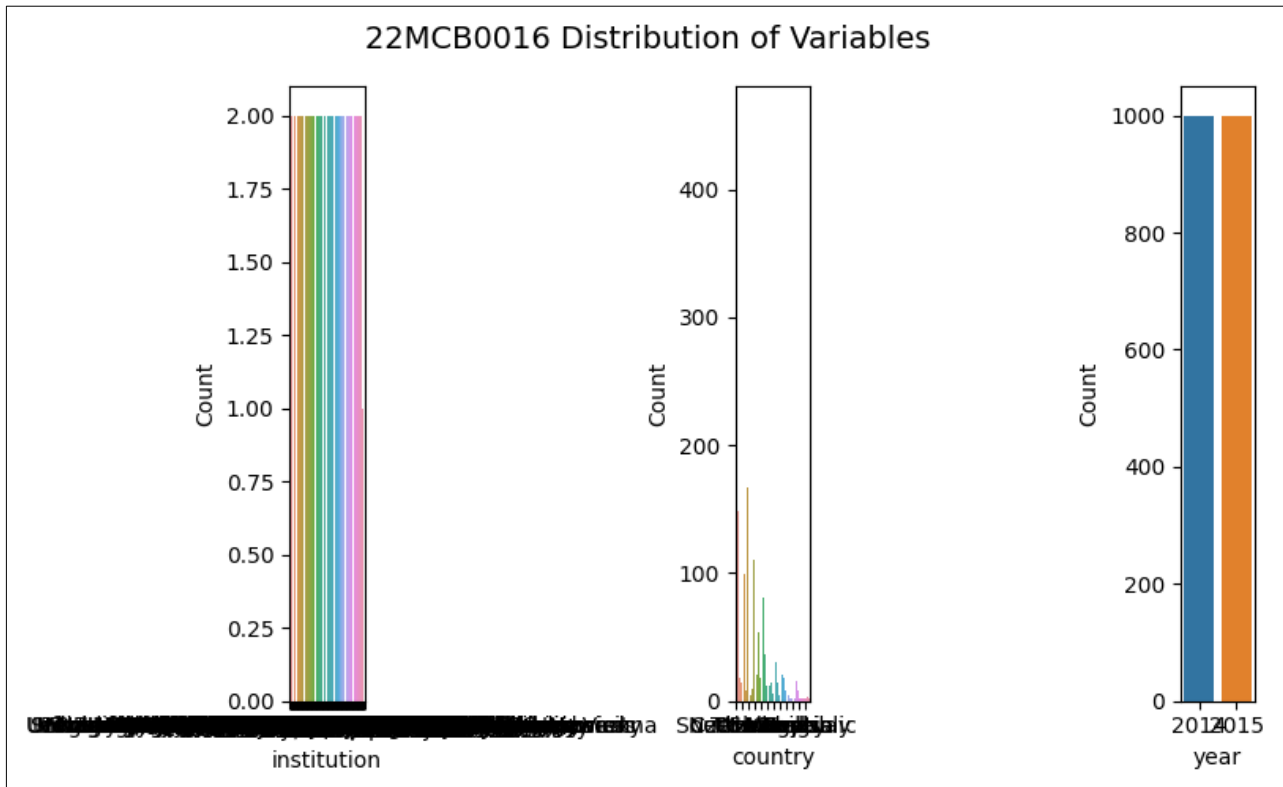


Fig.4. Shows the count plots for the selected columns ('Document Title', 'Authors', 'Publication Year')

CONCLUSION

The analysis and visualization of the bibliographic graph revealed valuable insights into the relationships between research topics and papers. The node and edge classification provided a structured representation of the heterogeneous network, while clustering aided in identifying cohesive research areas. The visualizations enhanced the understanding of the graph, making it easier to explore and interpret its intricacies. Researchers can leverage these findings to gain deeper insights into the research landscape and identify potential.

The methodology involves preparing and cleaning the dataset, extracting relevant information, and constructing a directed graph representing paper titles and citations. The graph is then visualized using a spring layout algorithm. Additionally, node features are extracted, and text processing and vectorization techniques are applied for clustering. Dimensionality reduction is performed using Truncated SVD, followed by applying K-means clustering and visualizing the clusters with a scatter plot. Variable distributions are evaluated using count plots for further insights.