Analysis of Air Transportation Network

Submitted in partial fulfilment of the requirements for the course of

IE532 ANALYSIS OF NETWORK DATA

By:

PROFESSOR CHRYSAFIS VOGIATZIS

Submitted by:

DHRUV NARESH BORDA

(UIN-675438725, NetID-borda2@illinois.edu)

DEEKSHA MANOHAR RAO

(UIN-, NetID-dm48@illinois.edu)

SAYANTAN MALLADEB

(UIN-, NetID-sm103@illinois.edu)

SUVRATA GAYATHRI KAPPAGANTULA

(UIN-, NetID-sk108@illinois.edu)

UNIVERISTY OF ILLINOIS URBANA-CHAMPAIGN

DECEMBER 2023

Table of Content

TABLE OF CONTENT **2**

LIST OF FIGURES **3**

ACKNOWLEDGMENT **4**

ABSTRACT **5**

1. INTRODUCTION **6**

1.1 GOAL STATEMENT 6

1.2 DESCRIPTION OF DATASET 6

1.2.1 Data Limitations *7*

2. METHODOLOGY AND COMPUTATIONAL IMPLEMENTATION **7**

2.1 DATA PREPROCESSING 7

2.2 DATA VISUALIZATION 7

2.2.1 Data Visualization of Flight Network *7*

2.2.2 Data Visualization of Flight Network based on User Preference *8*

2.3 DATA INSPECTION 10

2.4 NETWORK ANALYSIS 11

2.4.1 Mathematical Definition for Network Metrics *11*

2.4.2 Network Analysis Insights *12*

2.5 AIRLINES COMPARISON 13

3. CONCLUSION **15**

4. REFERENCES **15**

List of Figures

FIGURE 1 DATA VISUALIZATION OF NETWORK 8
FIGURE 2 PLOT FOR TOP 10 BIGGEST AIRLINES 9
FIGURE 3 DIRECTED NETWORK WITH DYNAMIC NODE SIZE 9
FIGURE 4 10 COUNTRIES WITH MOST AIRPORTS 10
FIGURE 5 DIRECTED NETWORK WITH DYNAMIC NODE SIZE 10
FIGURE 6 DEGREE OF NETWORK 12
FIGURE 7 FR AND WN AIRLINES COMPARISON 14
FIGURE 8 FR AND WN AIRLINES COMPARISON-METRICS 14

Acknowledgment

This report is based on the internship program as the partial fulfillment of course objective and requirements for the IE532 Analysis of Network Data. We perceive it as the best way to experience the implementation of academics in the real-life applications.

First, it is our radiant sentiment to place on record our deepest gratitude to Professor Chrysafis Vogiatzis for providing us a chance to implement our learnings from academics and for giving necessary guidance.

Last but not the least; we couldn't remain still without expressing our sincere thanks to all of our colleagues for their individual coordination, support, help and sincere cooperation during the project period.

Abstract

Air transportation plays a big role in global business operations and personal tourism. It is a way of supplying considerable financial opportunities connecting people and businesses all over the world. The important airports all over the globe are connected to many different airports. Here, we are going to consider every airport as a node and the routes between the airports as edges. The data was obtained from the "OpenFlights" source. Various centrality measures calculated for the data and a thorough network analysis conducted on the data helps in realizing few characteristics of the airports. Few characteristics includes visualization of data based on user given parameters, biggest airlines and airports, density of networks, betweenness, closeness centrality metric and many more. The characteristics aids in locating designated airports intended to boost the economy.

Keywords: Network Analysis, Centrality Measures, Biggest Airlines, Data Visualization

1. Introduction

1.1 Goal Statement

1. Identify the largest airlines and airports. Additionally, identify the nations that have the most airports.
2. Establish network statistics such as network degree, network density, average shortest path length, proximity and betweenness metrics for centrality, transitivity, and average clustering coefficient.
3. Determine the answers to insightful questions, for instance:
    a. *Airports with most incoming flights and outgoing flights*
    b. *Airports allowing to reach all other airports with the lowest and highest average number of airports in between.*
    c. *Airports that frequently and rarely serve as connectors between groups of airports*
4. Provide insights about the Flight Network by visualization on world map. Moreover, compare specific airlines to determine the area of operation.

1.2 Description of Dataset

Openflights is an opensource website which provides information on the worldwide scheduled air traffic. From this tool, we utilize the “Airports”, “Airlines” and “Routes” databases which contain flight data of various airlines operating between airports in the United States.

The Airports database contains data for over 10,000 airports until January 2017. This dataset provides information on 13 attributes of the airports like - Airport ID, name, city, country, IATA, ICAO, Latitude, Longitude, Altitude, Time zone, DST, Time zone database and Type. Airlines database contains data until January 2012 for 5888 airlines. This dataset provides information on 8 attributes like - Airline ID, Name, Alias, IATA, ICAO, Callsign, Country and Active. Routes database contains data until June 2014 contains 67663 routes between 3321 airports on 548 airlines. This dataset provides information on 9 attributes like - Airline, Airline ID, Source sirport, Source Airport ID, Destination Airport, Destination Airport ID, Codeshare, Stops and Equipment.

1.2.1 Data Limitations

The drawback with the current data is that the data has not been updated and it is available till 2017. Since, it's an old dataset it is missing the popular airlines like Etihad, Emirates and Lufthansa.

Another major drawback is we were not able to quantify the disruptions caused by the pandemic COVID-19. Since the pandemic would have major effect on all characteristics of the airport network.

2. Methodology and Computational Implementation

2.1 Data Preprocessing

Preprocessing data can increase the correctness and quality of a dataset, making it more reliable by removing missing or inconsistent data values brought on by human or computer error. It ensures consistency in data. In present analysis, three major processes are implemented under data preprocessing-Data Cleaning and Data Merging.

Firstly, Airlines, Routes and Airports dataset, as discussed in Section 1.2, are loaded in the program. Renaming and reindexing columns, eliminating missing values, and specifically removing complete entries with missing IDs between source and destination airports are just a few of the tasks involved in data cleaning.

Data Merging, or combining the cleaned Airlines, Routes, and Airports dataset into a data frame called "MergedData," is the final stage of data preprocessing.

2.2 Data Visualization

There are two main tasks that are implemented under data visualization as discussed in Section 2.2.1 and Section 2.2.2.

2.2.1 Data Visualization of Flight Network

The first, and most apparent information of network, is the network's nodes and edges. It can be derived that network contains 3222 nodes and 18862 edges. Second, the Basemap Matplotlib Toolkit has been called for the visualization of air routes and airports on a world map, depicted in Figure 1 Data Visualization of Network.

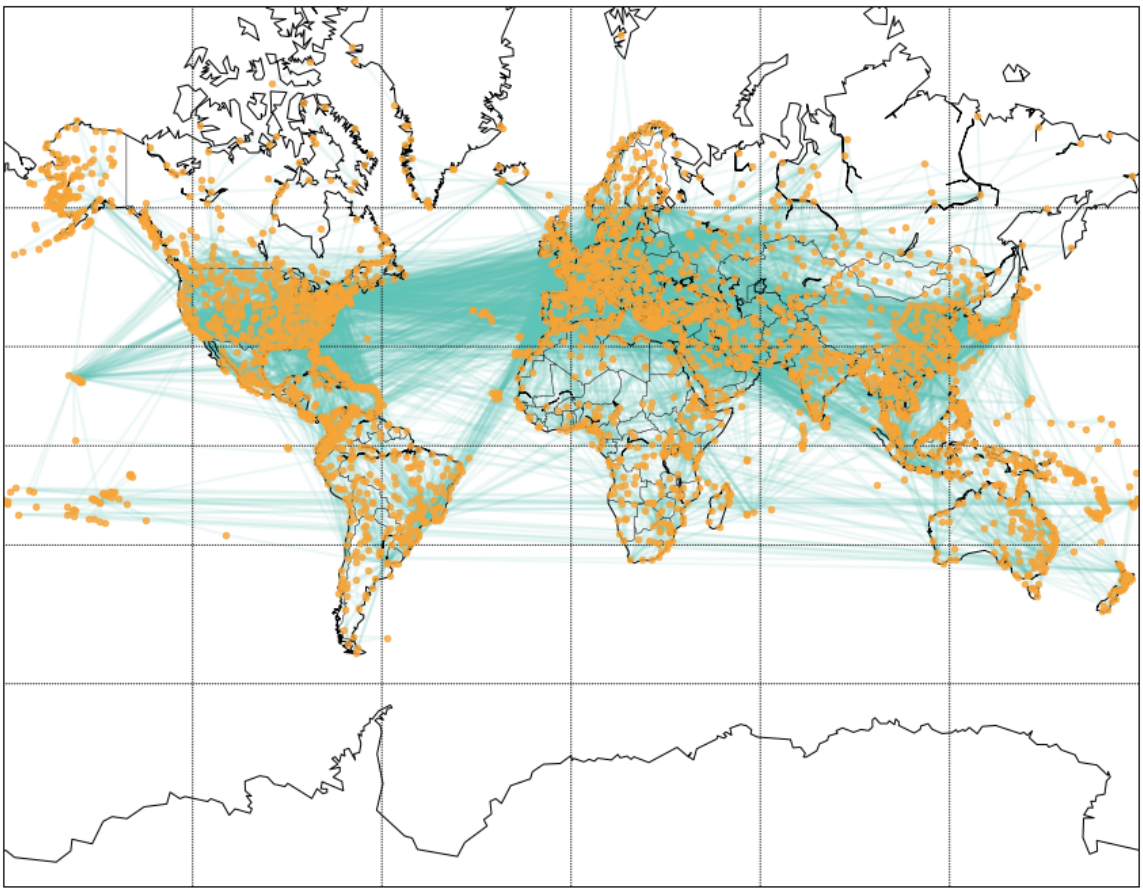

Figure 1 Data Visualization of Network

2.2.2 Data Visualization of Flight Network based on User Preference

The program incorporates the visualization depending on user preference in addition to the default Data Visualization of Flight Network, as stated in Section 2.2.1. Below is a representation of the code structure for the same.

1. Select all airlines and airports
2. Select specific airlines
    a. Select the biggest airlines
    b. Select a specific airlines
3. Select specific airports
    a. Select specifica airlines
    b. Select the biggest airlines

Additionally, options for creating directed or undirected graphs as well as options for static or dynamic node sizes are presented at the conclusion of each preference. Some instance for the same are depicted below:

Instance 1

Select the biggest airlines

- You chose to plot the biggest airlines
- How many of the biggest airlines do you want to plot? (1 to 15) 10
- You chose to plot the top 10 biggest airlines

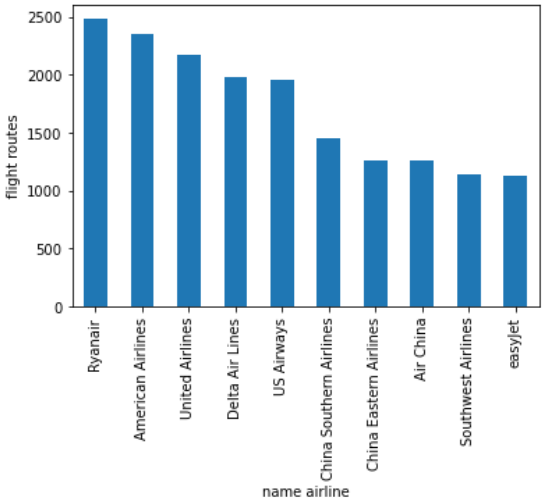

Figure 2 Plot for Top 10 Biggest Airlines

Instance 2

You chose to plot all airlines and airports

- Make a directed network
- Display size of airport depending on how many flight routes it has (degree)
- DiGraph with 3222 nodes and 36912 edges

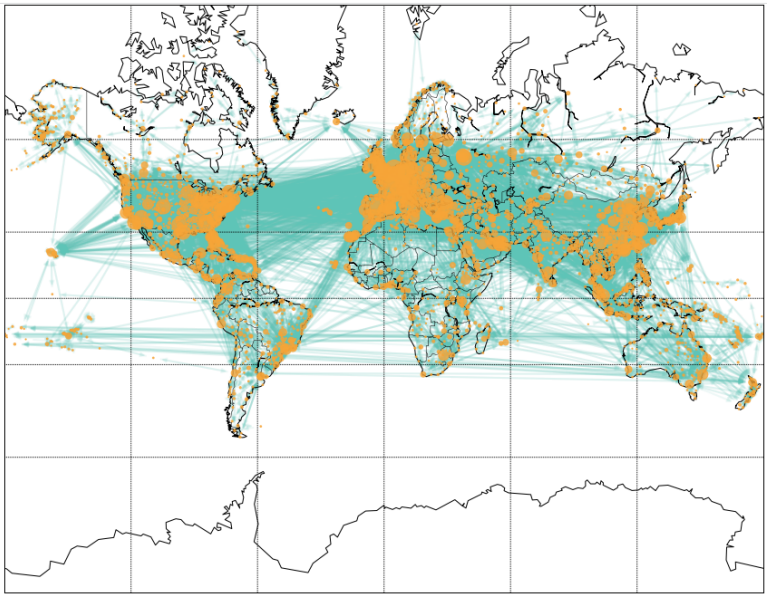

Figure 3 Directed network with Dynamic Node Size

2.3 Data Inspection

It entails identifying significant network-related insights. The following illustrates the code structure:

1. Show in which countries most airports are located
2. Show 10 biggest airports based on number of incoming flights
3. Show 10 biggest airports based on degree (most connected)
4. Show 10 biggest airlines

Instance 1

Show in which countries most airports are located

- You chose to plot the biggest airlines

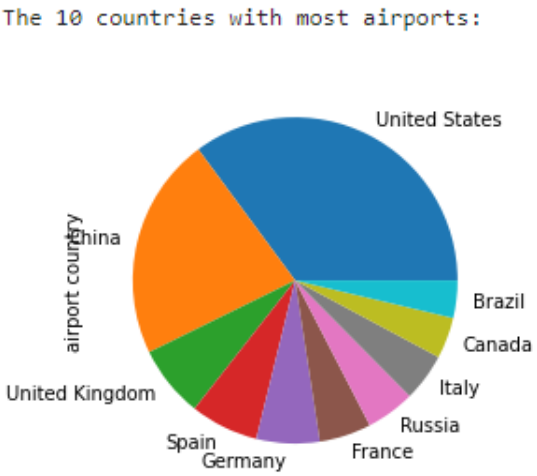

Figure 4 10 Countries with Most Airports

Instance 2

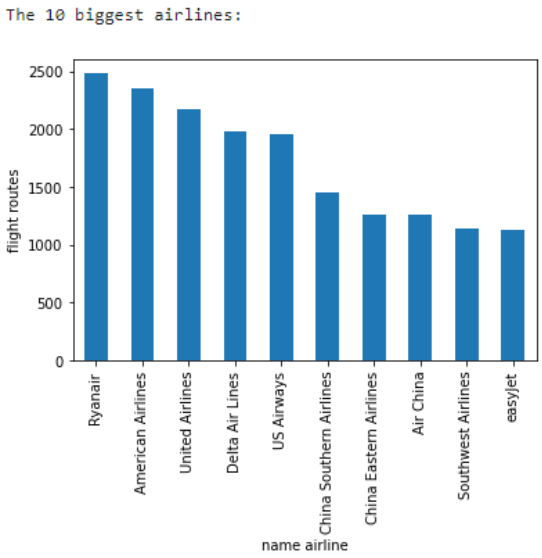

Figure 5 Directed network with Dynamic Node Size

2.4 Network Analysis

The main part of the project is the network analysis, measuring the network parameters.

2.4.1 Mathematical Definition for Network Metrics

1. Average clustering coefficient

A clustering coefficient is a measure of the degree to which nodes in a graph tend to cluster together in graph theory. The clustering coefficient for the graph is the average,

$$C = \frac{1}{n}\sum_{v \in G} c_v,$$

2. Closeness Centrality

Closeness centrality measures how short the shortest paths are from node i to all nodes.

It is the inverse of the average shortest distance between the vertex and the rest of the network's vertices. The formula is 1/(average distance to all other vertices).The inverse is used, so a higher closeness centrality score indicates a higher desirable centrality score (i.e., a shorter average distance to other vertices). Distance from node i to every other node j : dij .

$$c_i = \sum_j d_{ij}.$$

$$c_i = \left(\sum_j d_{ij}\right)^{-1} \implies c_i = \frac{n-1}{\sum_j d_{ij}} \text{(normalized)}.$$

3. Betweenness Centrality

Betweenness centrality is a measure that captures a completely different type of significance: the extent to which a particular vertex is on the shortest paths between other vertices.

total number of shortest paths connecting node k and `: $g_{kl}$ `.

total number of shortest paths connecting k and ` passing through i : $g^{i}_{kl}$ `.

$$b_i = \frac{\sum_{k,\ell \in V \setminus \{i\}} g^{i}_{k\ell}}{\sum_{k,\ell \in V \setminus \{i\}} g_{k\ell}}.$$

2.4.2 Network Analysis Insights

As shown in the code structure (followed by respective output) below, network analysis results in the development of deep insights about the Flight Network.

1. **Does a route exist between every two airports?** *False*
   Is every airport reachable from every other airport? *False*
2. **How many nodes are in the largest weakly connected component?**
   The percentage of nodes which are weakly connected: 0.991
3. **How many nodes are in the largest strongly connected component?**
   The percentage of nodes which are strongly connected: 0.985
4. **Average shortest path length**
   The average shortest path length for largest strongly connected network: 3.995
   The average shortest path length for largest weakly connected network: 3.979
5. **Density of Network**
   The density of original network: 0.0035567334247954964
   The density of largest strongly connected network: 0.003655488243553677
6. **Degree of Network**

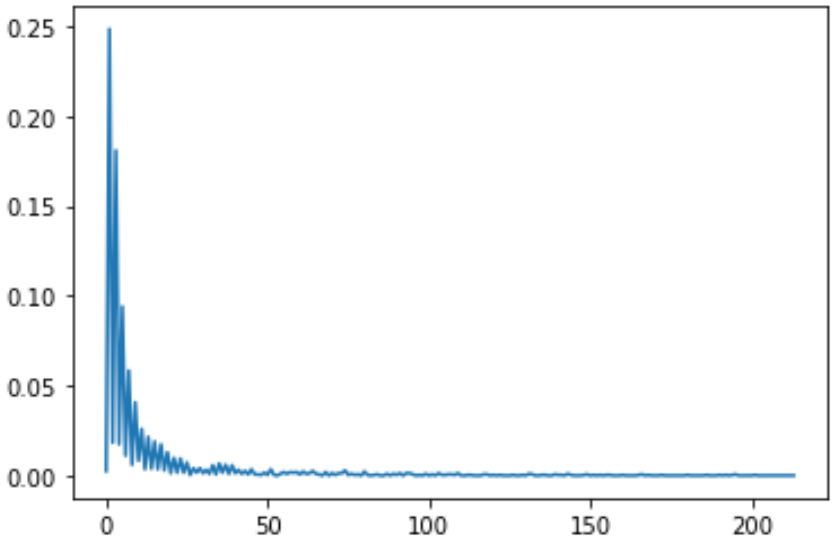

Figure 6 Degree of Network

7. **Transitivity and Average Clustering Coefficient**

   The transitivity for largest strongly connected network: 0.249683

   The transitivity for largest weakly connected network: 0.249681

   The average clustering coefficient for largest strongly connected network: 0.484168

   The average clustering coefficient for largest weakly connected network: 0.480729

8. **Indegree and Outdegree**

   The top 5 airports with most incoming flights: 'FRA', 'CDG', 'AMS', 'IST', 'ATL'

   The bottom 5 airports with most incoming flights: 'ODO', 'UKX', 'AYP', 'HUU', 'SCY'

   The top 5 airports with most outgoing flights: 'FRA', 'CDG', 'AMS', 'IST', 'ATL'

   The bottom 5 airports with most outgoing flights: 'ODO', 'UKX', 'AYP', 'HUU', 'SCY'

9. **Centrality Metrics: Closeness and Betweenness**

   Closeness Centrality

   Airports allowing to reach all other airports with the lowest average number of airports in between: 'FRA', 'CDG', 'LHR', 'DXB', 'AMS'

   Airports allowing to reach all other airports with the highest average number of airports in between: 'KPV', 'SRV', 'IRP', 'YHO', 'YBX'

   Betweenness Centrality

   Airports that frequently serve as connectors between groups of airports: 'ANC', 'LAX', 'CDG', 'DXB', 'FRA'

   Airports that rarely serve as connectors between groups of airports: 'ODO', 'UKX', 'ULK', 'AYP', 'PEM'

2.5 Airlines Comparison

The airline comparison sheds light on the regions that various airlines serve. For inspiration, the 10 airlines with the most flight routes considered are as below:

FR Ryanair

AA American Airlines

UA United Airlines

DL Delta Air Lines

US US Airways

CZ China Southern Airlines

MU China Eastern Airlines

CA Air China

WN Southwest Airlines

U2 easyJet

For instance, Figure 7 depicts the comparison among FR (Ryanair) and WN (Southwest Airlines).

Which airline do you want to visualise?

FR Ryanair

WN Southwest Airlines

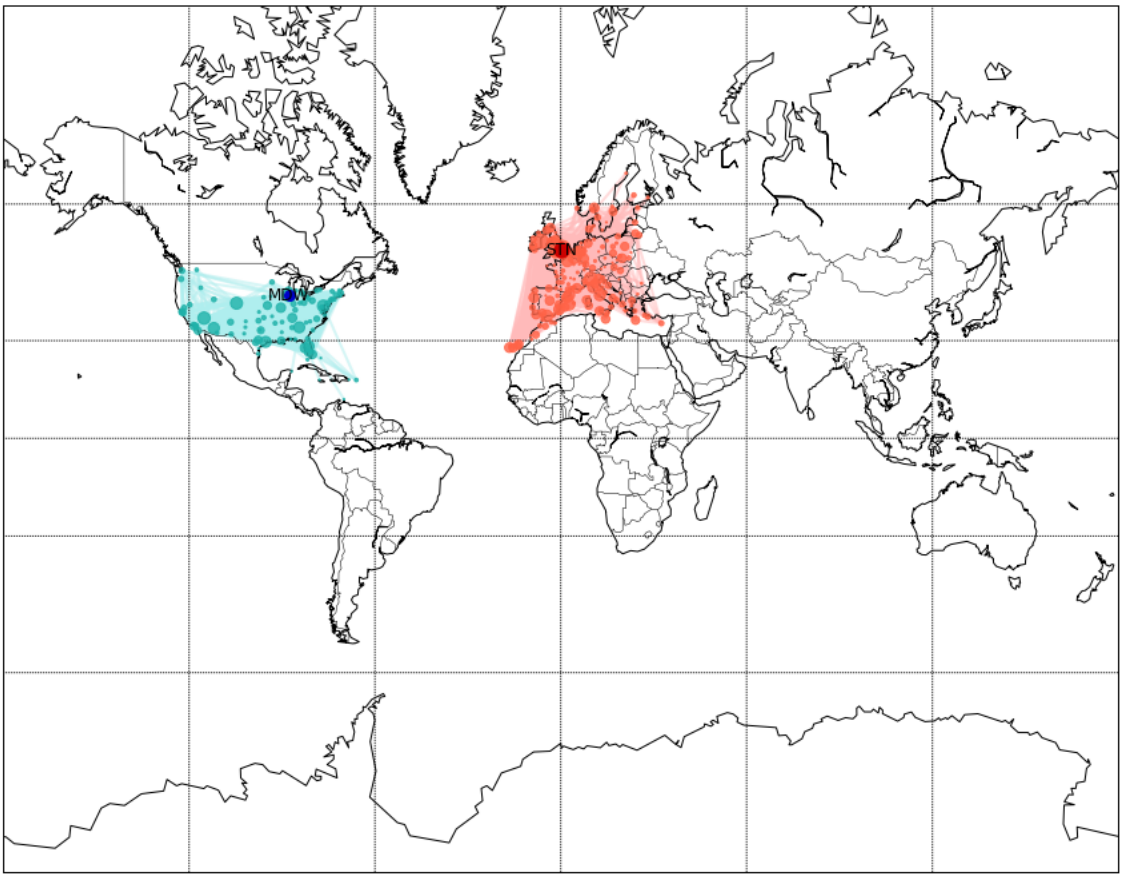

Figure 7 FR and WN Airlines Comparison

	metrics	airline 1	airline 2
0	nr of nodes	176	94
1	nr of edges	1242	574
2	density	0.080649	0.13132
3	global efficiency	0.460912	0.491621
4	biggest hub	STN	MDW

Figure 8 FR and WN Airlines Comparison-Metrics

3. Conclusion

1. The network analysis of the airline transportation data gives us the biggest and the most important airports. According to our analysis, USA has the greatest number of airports.

2. The airport which plays the most vital role is the one in Atlanta, Georgia based on the no. of incoming flights.

3. The closeness helps us to identify airports allowing to reach all other airports with the highest/lowest no. of airports in between. The betweenness centrality helps us identify the airports that frequently serve as connectors between groups of airports.

4. Identifying all the different characteristics will help plan ahead for the individual airports, for example: The busiest airport will require more workers. Finally, it will also help us in boosting the trade within and also with the other countries.

4. References

Data Source: https://openflights.org/data.html

Saleena, P. & Swetha, P.K. & D, Radha. (2018). Analysis and visualization of airport network to strengthen the economy. International Journal of Engineering and Technology (UAE). 7. 708-713. 10.14419/ijet.v7i2.9915.

Dorothy P. Cheung; Mehmet Hadi Gunes " A Complex Network Analysis of the United States Air Transportation". (2012)

Lee, Ju-Yang , Jang, Phil-Sik "Study on Research Trends in Airline Industry using Keyword Network Analysis: Focused on the Journal Articles in Scopus:" . (2017)