Identifying the Core Member in a Network

Pushpender Singh

Army Institute of Technology Pune, Dighi Hills 411015

Prof. N Balakrishnan

SERC, IISc Banglore, CV Raman Road, Bengaluru, Karnataka 560012

Table of Contents

Identifying the Core Member in a Network

Pushpender Singh

Army Institute of Technology Pune, Dighi Hills 411015

Guided by:

Prof. N Balakrishnan

SERC, IISc Banglore, CV Raman Road, Bengaluru, Karnataka 560012

Abstract

Social Network Analysis (SNA) analyses the structural properties of individuals or groups of individuals in a network. These measurements not only depict perspectives of the interconnections and relationships amongst various individuals, but also consider the effect of these interconnections on each other as well as on the group as a whole. In this project, we are analysing the network of airlines. We aim to study the causes of delays in flights in order to minimise them, and also to find new possible air-routes for flights, for better connectivity. To analyse this network, we are using the Airline Service Quality Performance dataset of the month of April 2018, provided by the Bureau of Transportation Statistics (BTS) and available on the public domain. The dataset has 479231 rows and 34 columns which contain information about the flight (such as flight name and number), origin of flight, destination, expected arrival time, expected departure time, actual arrival time, actual departure time and reason of delay (like security clearance, weather, carrier delay, late aircraft, National Aviation System(NAS) delay, etc.). By analysing this dataset, we are able to identify the network of flight routes with maximum delays, airports with the highest stay clearance delays and airports with the highest air traffic. The results will help us to reschedule flights so that the delays caused due to these factors can be reduced. It also helps us determine delays at airports sucha as security delays, carrier delays, etc., so that we can identify measures to minimise them. With the help of network visualisation techniques (using betweenness centrality measure, eigenvector centrality measure and degree centrality measure), we are able to determine the airports with maximum loads, and can set up new airports in nearby cities to reduce the burden. We can also determine the airports which play major roles in layovers using Prims and Kruskal algorithm, and also be able to find cities for which direct flight facilities can be provided. To implement this project and analyse the dataset, we are using programming language R and the packages dplyr (to manipulate the data frame), RNeo4j (to store data in graph form)

and igraph (visualisation, network analysis and centrality measures), and for plotting graphs, we are using R library ggplot2.

Keywords: network analysis, centrality measures, graph theory, visualisation

INTRODUCTION

1 Background

According to the report published in The Garfors Globe[1] the total number of flights takeoff in 2014 is about 37.4 million which means about 102,465 per day and according to Telegraph[2] the total number of people travel in these flights is around 3,328,000,000 (in 2014). Airways is one of the fastest growing industry. According to Statista[4] it is believed that the global aviation industry will reach up to 33.8 billion US dollars profit in 2018 which makes airways 750 billion dollar industry.

But with increasing passenger and revenue some great challenges are also growing. The flight delays are among some those challenges. According to the data received from the flightstats[3] last (June 2018) month, there were 192,925 flight delays. The average delay time is about 43 minutes[5] globally.

1.1 Statement of Problem

People find it extremely frustrating when they have to wait for something, especially when they do not know the reason for the delays. We saw the opportunity to provide people with information to improve their journeys, by suggesting various possibilities to avoid delayed flights. So here we are analysing flight data to find out different delays and how we can minimise these delays. We also study about the core-member(busiest airport) in the network so that we can suggest areas for new airports.

1.2 Objectives of Research

Using the flight dataset, we are trying to answer some of the questions mentioned below in our analysis.

1.Which is the busiest airport in Dec 2015?

2. Which airport has maximum delays (minutes)?

3. Categorizing which delay reason has caused the maximum delay (minutes).

4. Analyzing the top 20 well connected airports from the busiest airport.

5. Flight routes with the most delays.

6. Flight routes with the least delays.

7. Airports with high delays to stay clear of.

8. Airports with the least delays.

9. To find the core member in the network.

1.3 Scope

After successfully implementing this project we will be able to

1. Find the busiest airport (core member) and can suggest the area where another airport can be opened to reduce the burden or expansion of the airport. 2. We will able to suggest to start some direct flights between two cities to reduce layover time. 3. We can analyse the reason for delays and able to minimise them.

2 LITERATURE REVIEW

2.1 Information

2.1.1 Social Network Analysis[6]

Social Network Analysis (SNA) analyses the structural properties of individuals or groups of individuals in a network. These measurements not only depict perspectives of

the interconnections and relationships amongst various individuals but also consider the effect of these interconnections on each other as well as on the group of interconnected individual.

2.1.2 Centrality Measures[6]

Centrality, or prestige, is a general measure of how the position of an actor is within the overall structure of the social network and can be computed resorting to several metrics. The most widely used are degree, betweenness, closeness, and eigenvector centrality.

2.1.2.1 Degree Centrality Measures[7]

A node is important if it has many neighbors, or, in the directed case, if there are many other nodes that link to it, or if it links to many other nodes.

$$C_d(G) = \frac{\sum_{i=1}^{g} [\, C_d(v^*) - C_d(V_i) \,]}{H}$$

here, H= (n-1)(n-2) = n^2-3n+2 , C_d(G) is degree centrality of graph , v* the node with highest degree centrality, g is total number of vertex in graph.

2.1.2.2 Betweenness Centrality[7]

It is a measure of centrality in a graph based on shortest path. For every pair of vertices in a connected graph, there exists at least one shortest path between the vertices such that either the number of edges that the path passes through (for unweighted graphs) or the sum of the weights of the edges (for weighted graphs) is minimized. The betweenness centrality for each vertex is the number of these shortest paths that pass through the vertex.

$$C_B = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

where σ_{st} is the total number of shortest paths from node s to node t and $\sigma_{st}(v)$ is the number of those paths that pass through v.

2.1.2.3 Closeness Centrality[8]

$$C_c = \frac{1}{\sum_y d(y,x)}$$

where d(y,x) is the distance between vertices x and y.

2.1.2.4 Eigenvector Centrality[9]

In graph theory, eigenvector centrality (also called eigencentrality) is a measure of the influence of a node in a network. Google's Page Rank and the Kartz Centrality are variants of the eigenvector centrality.

$$C_E = \frac{1}{\lambda}\sum_{t \in M(v)} x_t = \frac{1}{\lambda}\sum_{t \in G} a_{v,t} x_t$$

3 METHODOLOGY

To find the core member in the network and depalys we are using two different approches

1. Using R language(using Neo4j server and inbuilt libraries and packages)

2. Using Gephi software

3.1 USING R LANGUAGE

DATASET

The dataset[10] is available on the Bureau of Transportation Statistics (BTS) website which is a Federal multi-organizational entity which focuses on all forms of U.S transportation data collection & analysis. The data used in this project is of month December 2015. The dataset contains 479231 row and 32 column.

DATASET PREPARATION

The dataset preparation process involves the importing libraries, importing dataset, cleaning dataset and creating a data frame.

PACKAGES USED

(I) dplyr package We will install and load the dplyr package that contains additional functions for data manipulation using data frames. It allows us to order rows, select rows, select variables, modify variables and summarize variables. We will be using functions like distinct , filter , group_by from this package.

(II) RNeo4j package Neo4j, a graph database, allows users to store their data as a property graph. A graph consists of nodes that are connected by relationships; both nodes and relationships can have properties, or key-value pairs. RNeo4j is Neo4j's R driver. It allows users to read and write data from and to Neo4j directly from their R environment.

(III) igraph package Routines for simple graphs and network analysis. It can handle large graphs very well and provides functions for generating random and regular graphs, graph visualization, centrality methods and much more.

(IV) magrittr package Provides a mechanism for chaining commands with a new forward-pipe operator, %> % . This operator will forward a value, or the result of an expression, into the next function call/expression. There is flexible support for the type of right-hand side expressions.

IMPORTING FILE INTO R DATAFRAME

The file we download from public domain was in CSV format. To process the data in the file we bring it into data frames. To import the file into dataframe we use read.csv() function. This function read the file and arrange data into the frame for further operations.

FILTERING AND CLEANING DATA

The dataset we get from public domain was not clean. It has some blank fields and missing values. To handle this problem we use mutate function of the dplyr library. With the help of this function, we calculate the mean of the column and replace it with the missing value. If some parameters whose mean is not possible like airport id and airport name or destination we completely remove that row. We replace NULL or NA with 0 to make calculation easier

STARTING NEO4J SERVER

NEO4J - Neo4j is an open source graph database management system developed by Neo Technology, Inc. It is designed for optimizing fast management, storage, and traversal of nodes and relationship. It stores data in graph forms.

We use Rneo4j library to access Neo4j database through R. The neo4j database was locally hosted on port 7474 ("http://localhost:7474/db/data/"). To start the neo4j server we use startGraph function of RNeo4j library to connect R programming platform with Neo4j database. To clear(delete any value that pre-exists) we use clear function.

NETWORK ANALYSIS

After starting the server we start network analysis to find out different pieces of information like the busiest airport, flights between two airports, delays from the data set. We break the dataset into different parts for analysis. The results of the network analysis are discussed in detail in the result section.

3.2 USING GEPHI SOFTWARE

INTRODUCTION TO GEPHI[11]

Gephi is an open-source software for network visualization and analysis. It helps data analysts to intuitively reveal patterns and trends, highlight outliers and tells stories with their data. It uses a 3D render engine to display large graphs in real-time and to speed up the exploration. Gephi combines built-in functionalities and flexible architecture to explore, analyze, spatialize, filter, cluster, manipulate, export all types of networks.

Gephi is based on a visualize-and-manipulate paradigm which allows any user to discover networks and data properties. Moreover, it is designed to follow the chain of a case study, from the data file to nice printable maps.

DATASET PREPARATION AND IMPORTING DATASET

To use dataset into Gephi data frame we need to rename some fields in the dataset. In Gephi we can import two datasets one is node dataset and other is edge dataset. The node dataset contains id that is node and label. The edge dataset contains source, destination and all other information about that edge.

DATA VISUALISTAION IN GEPHI

After importing the dataset into Gepthi data frame, It create the graph. Before applying any filters the graph look like below

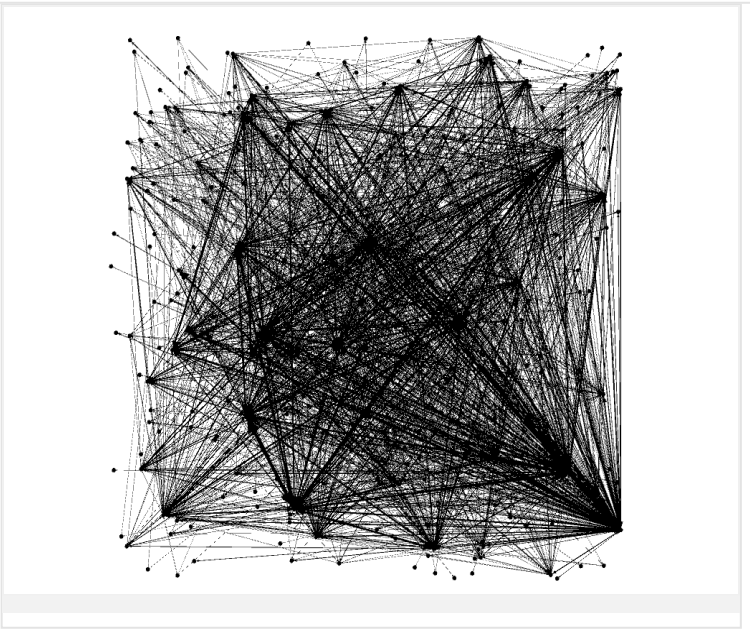


Fig 1 Graph generate by gephi after importing dataset

To visualise graph and network more effectively we increase the size of node and use fruchterman reingold layout. Now graph become more readable and informative. Now each node is visible clearly.

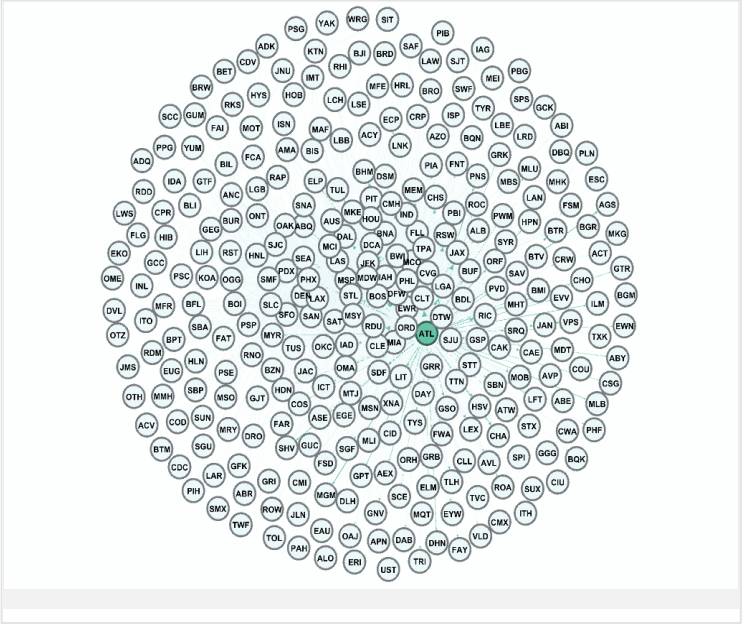


Fig 2 Graph after fruchterman reingold layout

4 RESULTS

USING R LANGAUAGE

1. ATL (Hartsfield-Jackson Atlanta International Airport) is the busiest airport in the US. It also means that you can virtually reach anywhere in US if you plan your trip via ATL. It has indegree 0f 159 and out-degree of 159 (total flights 318).

2. GUM (Guam), which is not on mainland US, but an island territory in western pacific, is only connected to US via Honolulu, another island territory. But, Honolulu has good connectivity, because of tourist attraction, reaching to Guam from Honolulu is pretty easy.

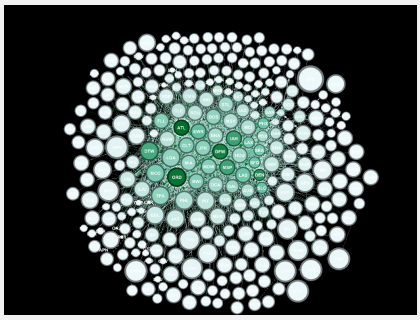
3. ATL(Hartsfield-Jackson Atlanta International Airport), DFW(Dallas/Fort Worth International Airport) and ORD (O'Hare International Airport) are the airports which have most flights departing from and arriving to. These are the core member of the network. ALT is the most important member in the network.

4. OO (SkyWest Airlines) is the carrier with the most delays.

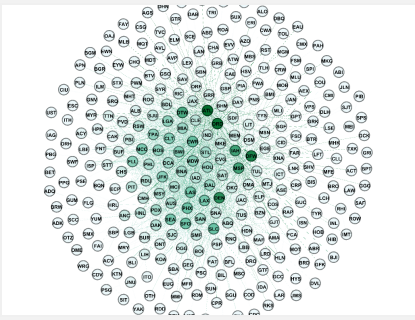
5. PPG 798airport had the highest average departure delay time, specifically 183.66 minutes, in the month of december 2015.

6. Carrier delays contribute to most of the flight delays.

Using Gephi Software



a) core-member in the network



b) core-member in the network

Fig 3 core-member in the network

1. The above figure showing the core-member in the network. The nodes with darker green colour are the most influential member or the most important member. Virtually you can reach to any nodes through these nodes. These node are ATL(Hartsfield-Jackson Atlanta International Airport), DFW(Dallas/Fort Worth International Airport) and ORD (O'Hare International Airport). With ALT core member of the network.(degree 318)

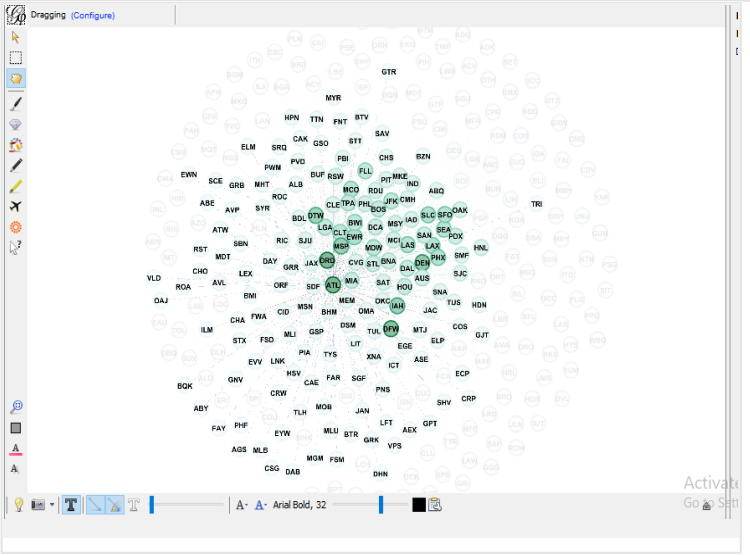
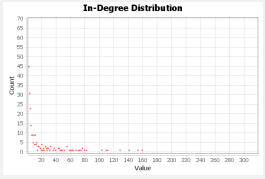
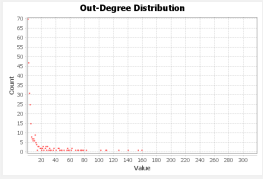


Fig 4 Adjacent nodes of core-member(ALT)

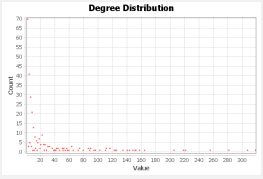
The above figure shows all the airports you can reached from the cental-node(ALT) airport. their are total 159(51.7 %) airports.



a) In-degree distribution of graph



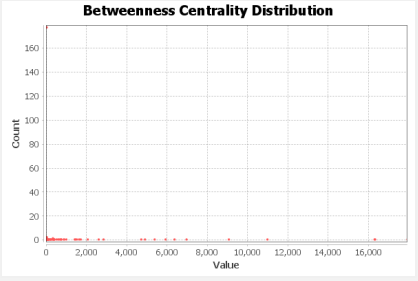
b) Out-degree distribution of graph



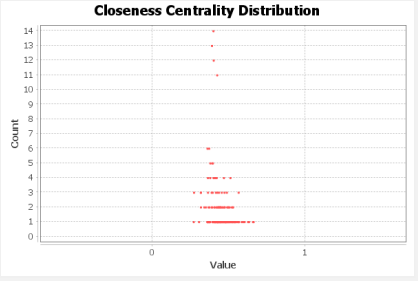
c) Degree-distribution of graph

Fig 5 Degree-distribution of graph

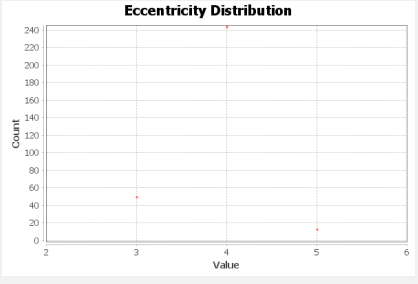
3. From the degree-distribution of the network we can conclude that maximum number of airpots cover 40 to 60 airports. The core-member covers more than 300 airports. So these are the busiest airports.



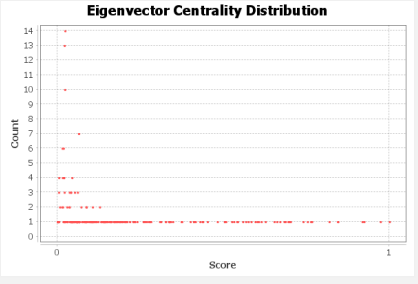
a) Betweenness Centrality Distributions



b) Closeness Centrality Distributions



c) Eccentricity Distribution



d) Eigenvector Centrality Distribution

Fig 6 Centrality of the network.

Centrality Measures

1.**Betweenness Centrality Distributions**

From the betweenness Centrality we can conclude that Maximum nuber of flights are for shorter distance(most flights lies between 0 to 2000 value).

2. **Closeness Centrality Distributions**

From the Centrality Distributions we can conclude that centrality of maximum airport lies between 0.2 to 0.6.

3.**Eccentrity Distributions**

From eccentrity distribution we can conclude that their are about 55 airpots which are close and connected with direct flights. Maximum airports(about 248) are connected to medium diatance airports and only (about 11) airports which are connected to far airports.

4. **Eigenvector Centrality Distributions**

From Eigenvector Centrality Distributions we can conclude that there are 4 airports ALT(1), ORD(0.9729), DEN(0.9236), DFW(0.9182). These are the imporatant airports in the network. You can virtually reach anywhere from these airports.

5 CONCLUSION AND RECOMMENDATIONS

From the data analysis we can conclude and recommend following things:-

1. ATL(Hartsfield-Jackson Atlanta International Airport), DFW(Dallas/Fort Worth International Airport) and ORD (O'Hare International Airport) are the airports which have most flights departing from and arriving to so we can establish new airports in nearby cities or increase their capacities.

2. GUM (Guam), which is not on mainland US, but an island territory in western pacific, is only connected to US via Honolulu,but this terrority has highest potential of tourism so we can connect it through more airports.

3. Carrier delay which is responsible for maximum times of delays can be completely reduce by taking proper actions.(reasons for carrier delays are aircraft cleaning, aircraft

damage, awaiting the arrival of connecting passengers or crew, baggage, bird strike, cargo loading, catering, computer, outage-carrier equipment, crew legality (pilot or attendant rest), damage by hazardous goods, engineering inspection, fueling, handling disabled passengers, late crew, lavatory servicing, maintenance, oversales, potable water servicing, removal of unruly passenger, slow boarding or seating, stowing carry-on baggage, weight and balance delays.)

5.1 FUTURE SCOPE

After analysing different datasets (Dec 2015,Dec 2016, Dec 2017). It is possible to train a model(using machine learning) to predict the possible delays and delayed time. From the result we obtain from our analysis and observation we can conclude that flight delay is a first -degree chaotic process, and it is possible to predict deays in flights.

REFERENCES

1.The Garfors Globe (https://garfors.com/2014/06/100000-flights-day-html/)

2. Telegraph (https://www.telegraph.co.uk/travel/travel-truths/how-many-planes-are-there-in-the-world/)

3. The Flightstats(https://www.flightstats.com/v2/global-cancellations-and-delays)

4. Statista(https://www.statista.com/statistics/193533/growth-of-global-air-traffic-passenger-demand/)

5. Economic time(https://www.economist.com/graphic-detail/2017/10/30/why-china-leads-the-world-in-flight-delays)

6. WIREs Data Mining Knowl Discov 2012, 2: 99–115 doi: 10.1002/widm.1048

7. Linton C. Freeman, Lehigh University, Centrality in Social Networks Conceptual Clarification, Social network 1 (1978/79) 215-239.

8. Alex Bavelas. Communication patterns in task-oriented groups.*J. Acoust. Soc. Am*,**22**(6):725–730, 1950

9. Dataset (https://www.bts.gov/browse-statistical-products-and-data/bts-publications/airline-service-quality-performance-234-time)

10. Bastian M., Heymann S., Jacomy M. (2009). *Gephi: an open source software for exploring and manipulating networks.*International AAAI Conference on Weblogs and Social Media.

ACKNOWLEDGEMENTS

I wish to express my sincere gratitude to my guide and mentor, Prof. N Balakrishnan for guiding and encouraging me during the course of my fellowship in Indian Institute of Sciences while working on the project on “Identifying the core member in the network”.

I also take the opportunity to thank Mrs K Nagarathna for helping me in various aspects of carrying out this project.

I sincerely thank the coordinator of Summer Research Fellowship 2018, Mr CS Ravi Kumar and his team for giving me the opportunity to embark on this project.