**Social Analytics**

Air transport network analysis

Under the Guidance of Prof. Ali Tafti

University of Illinois at Chicago

IDS 564 – Project Report

**Submission By-**

Shivam Duseja

Pawanjeet Kaur

taniya rajani

# 

# Introduction

Billions of people take air routes every year and the air industry have a great impact on the world's economy. The airport network is complex with great connectivity secondary if not primary. Hundreds of rural airports have connectivity via intermediate routes. Moreover, this is an interesting area to study network science to understand the world’s connectivity and identify community structures. Complex networks like these are ideal to study network structures aspects like small world phenomena and preferential or random attachment. With rapid growth in air travels, the interest in route optimization has increased. This study can aid in reducing air traffic and optimizing air routes. Our goal is to analyze the characteristics of this network, infer the most connected and visited airports and identify the community structures.

# Data Description

The route dataset is taken from <https://openflights.org/data.html>, It is a freely available dataset consisting of 67663 routes between 3321 airports and 548 airlines across the world. The datasets include following features:

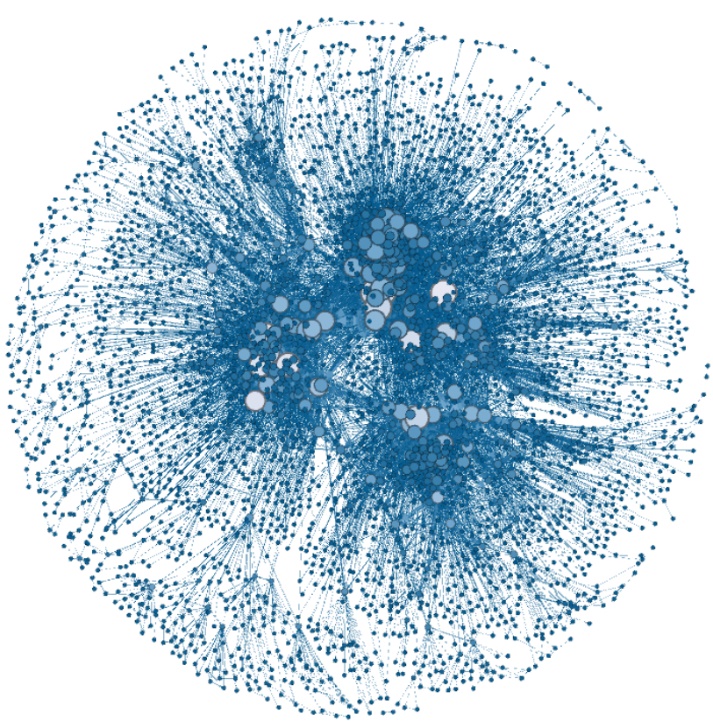
|  |  |
| --- | --- |
| Feature | Description |
| Airline | 2-letter (IATA) or 3-letter (ICAO) code of the airline. |
| Airline ID | Unique OpenFlights identifier for airline |
| Source airport | 3-letter (IATA) or 4-letter (ICAO) code |
| Source airport ID | Unique OpenFlights identifier |
| Destination airport | 3-letter (IATA) or 4-letter (ICAO) code |
| Destination airport ID | Unique OpenFlights identifier |
| Codeshare | "Y" if flight not operated by Airline, but another carrier |
| Stops | Number of stops on this flight |
| Equipment | 3-letter codes for plane type |

# Network Structure Analysis

The network graph will help us examine the extracted data. There are a total 3425 nodes or airports part of this network with 37595 unique routes or edges with varying weights. The weight here signifies the number of airlines operating on the route. It is difficult to analyze such a big network visually as there are overlapping nodes and edges. Hence, we would create a subgraph with minimum edge weight 6 and analyze the network of those routes which have at least 6 airlines operating.

The original network statistics are mentioned in the table below:

|  |  |
| --- | --- |
| Feature | Statistics |
| Number of routes | 67663 |
| Number of unique routes | 37595 |
| Number of unique departing airports | 3409 |
| Number of unique destination airports | 3418 |
| Number of nodes | 3425 |
| Number of edges | 37595 |

We will start our analysis by creating a graph of the airport routes network on the world map using the geo layout – post calculating various statistics wherein color of the nodes represents the degree and size of the node is basically the page rank.

For better understanding, we also created the network chart using the Fruchterman-Reingold algorithm as shown in the figure on the right.

Fort further insights, we have also listed the distribution of in-degree, out-degree and all-degree along with the edge weights in the tables below:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Degree | 0% | 1% | 5% | 10% | 25% | 50% | 75% | 90% | 95% | 99% | 100% |
| out | 0 | 1 | 1 | 1 | 1 | 3 | 8 | 28 | 53.8 | 134.52 | 239 |
| in | 0 | 1 | 1 | 1 | 1 | 3 | 8 | 28.6 | 53.8 | 134 | 238 |
| all | 1 | 2 | 2 | 2 | 2 | 6 | 16 | 57 | 107.8 | 268.52 | 477 |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Weight | 0% | 1% | 5% | 10% | 25% | 50% | 75% | 90% | 95% | 99% | 100% |
| Value | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 3 | 4 | 7 | 20 |

From the degree statistics table, we can infer that there are few nodes with higher out and in-degree and thus have maximum travel. It is interesting to see that some of the airports have zero in and out-degree.

Listing down some of the interesting airport codes with ***zero out*** **degree**:

BSS BVS CMP CZJ DLZ FMI KPR KYK KZB KZI MTE ORX QFX SPI TUA UII

The above airport codes are spread throughout the world – Balsas, Brazil (BSS) – Tupile, Panama (CZJ) – Greenland, Denmark (QFX)

Listing down some of the interesting airport codes with ***zero in*** **degree**:

IUE MSW VDA LJA PTJ SXX STZ

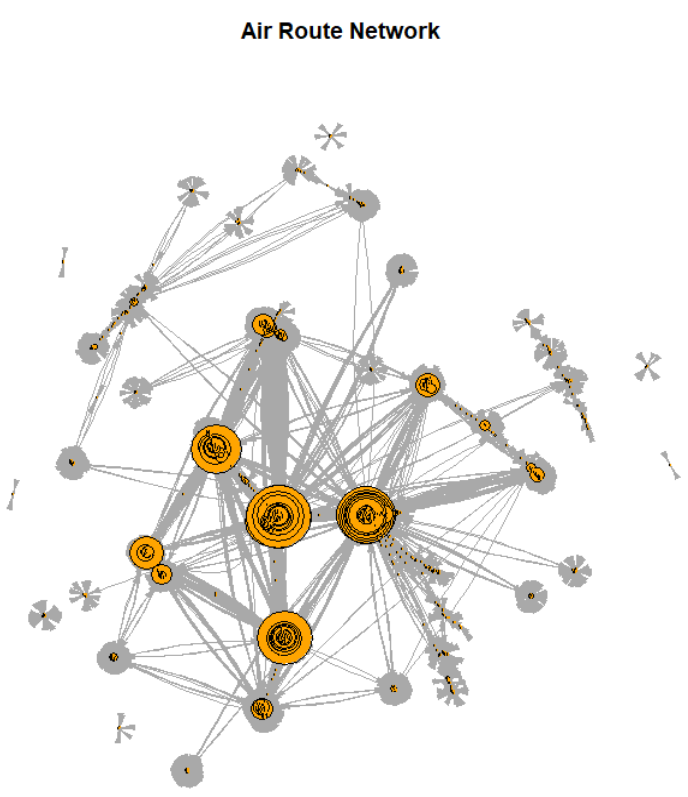
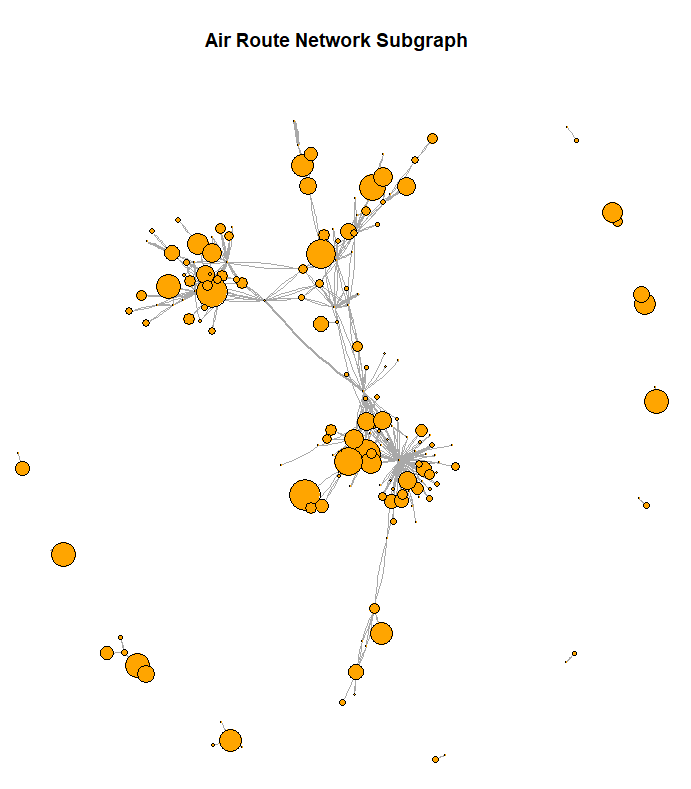
Similar to the out-degree cases, these airports are located across the world in countries like – Republic of Congo (LJA), Portland, Australia (PTJ), Brazil (SXX and STZ)

Looking at above statistics, it is clear - that certain routes have many airlines operating and few nodes/airports have higher connectivity. Hence, creating a subset of our graph would help us in analyzing and would make sense for better visualization.

For creating a subset of our graph – we will be filtering our network on the basis of number of airlines operating in the network. We will take the threshold value of 6 to filter our network i.e.– the weight of the Edge should be greater than 6. In this way we can look into routes that are busy and better connected.

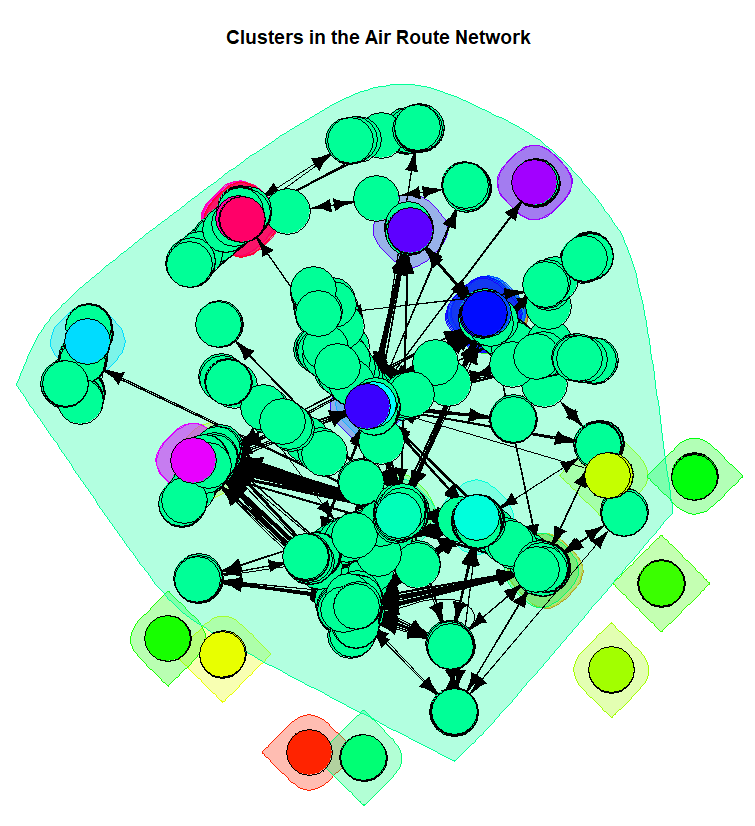
The airport route network with more than 6 airlines has 177 vertices and 476 edges. A network chart of this subgraph has been shown on the next page -

For better visibility and understanding, the vertex and edge weight size has been rescaled such that higher size indicates greater value.

Below table shows the distribution of in-degree, out-degree and all-degree and edge weight for subgraph network:

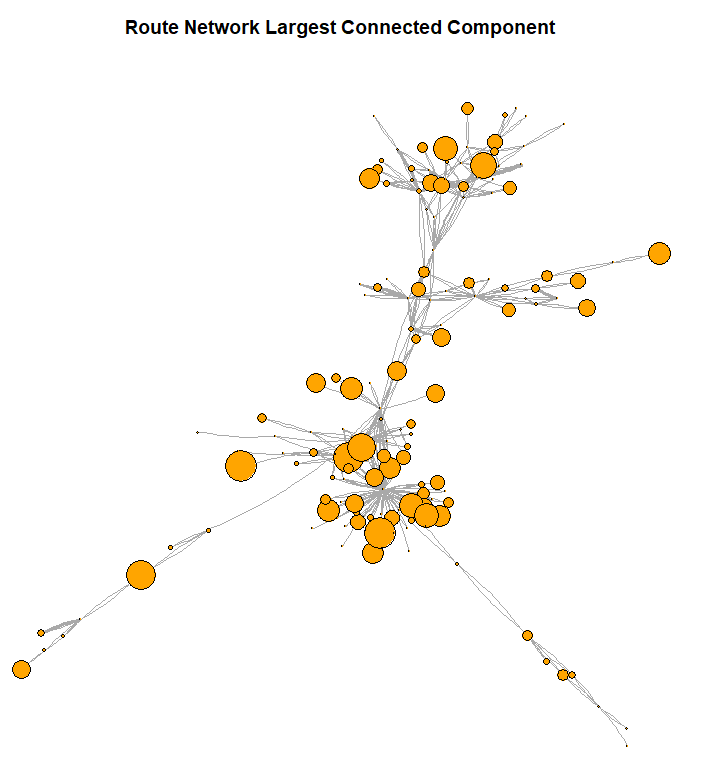
|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Degree | 0% | 1% | 5% | 10% | 25% | 50% | 75% | 90% | 95% | 99% | 100% |
| out | 0 | 0 | 0 | 1 | 1 | 1 | 3 | 5 | 9 | 15 | 48 |
| in | 0 | 0 | 0 | 1 | 1 | 1 | 3 | 5 | 9.2 | 15.48 | 46 |
| all | 1 | 1 | 1 | 1 | 2 | 2 | 6 | 10 | 18 | 30.48 | 94 |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Weight | 0% | 1% | 5% | 10% | 25% | 50% | 75% | 90% | 95% | 99% | 100% |
| Value | 7 | 7 | 7 | 7 | 7 | 8 | 9 | 10 | 11 | 12 | 20 |

**Clusters in the Air Route Network**

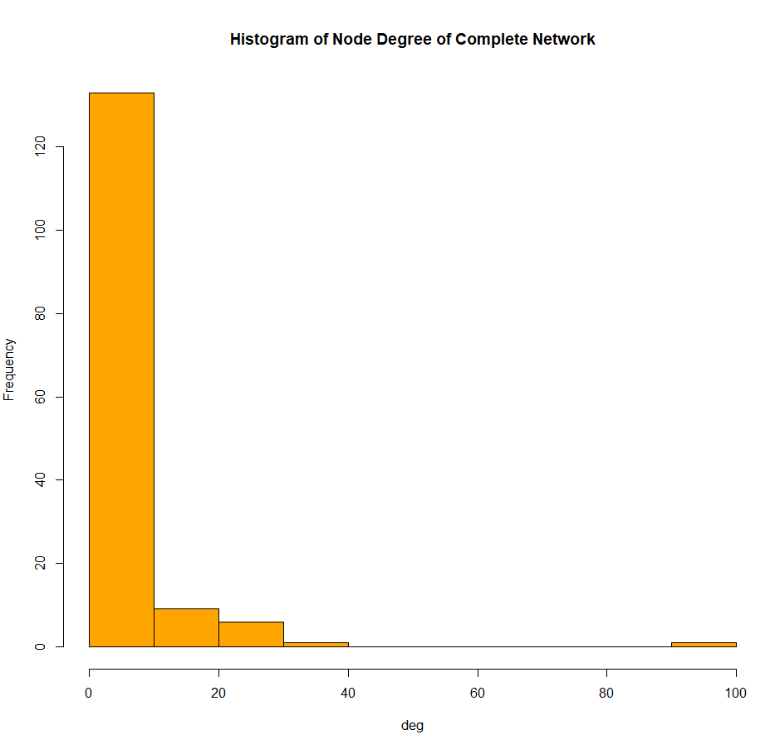
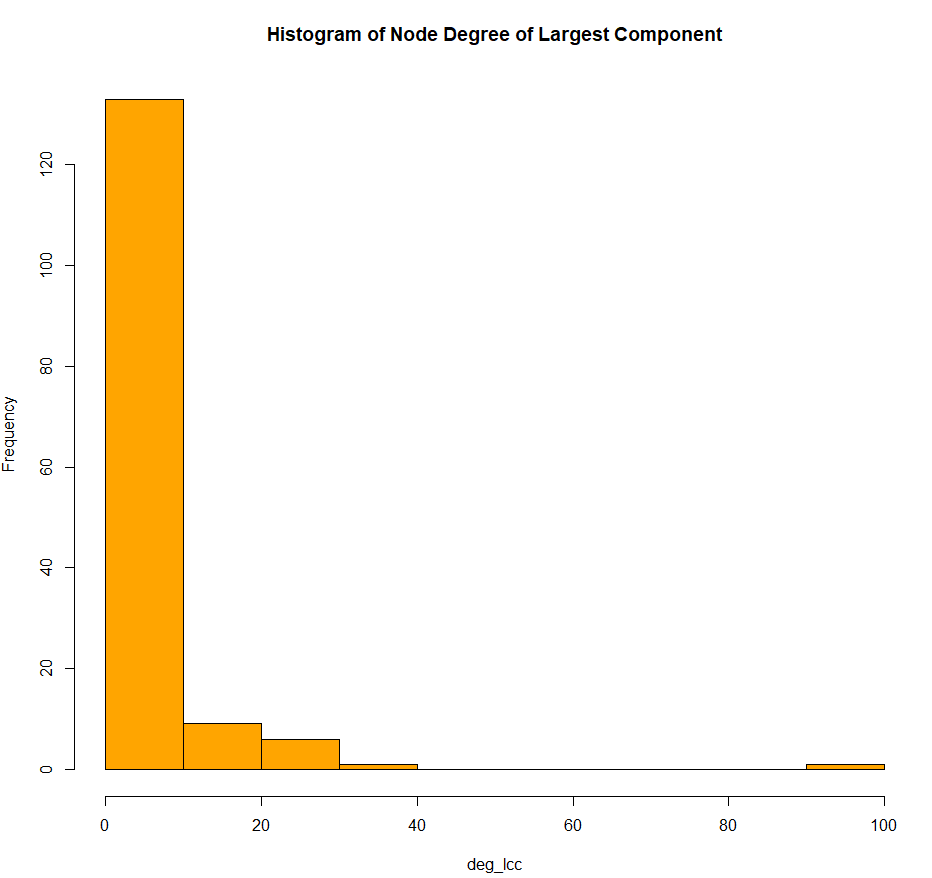
In order to further analyze our network, we created different clusters of our air route network based on strongly connected components as shown on the left–

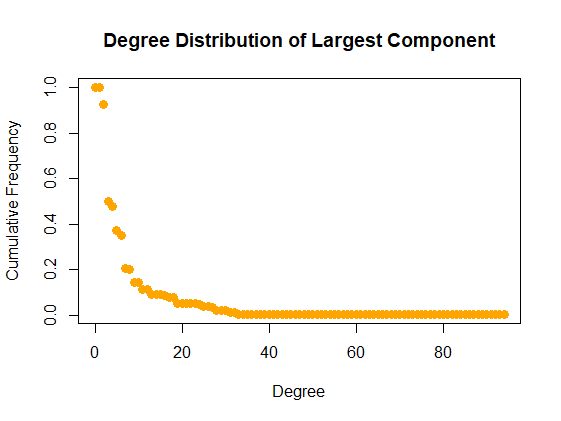
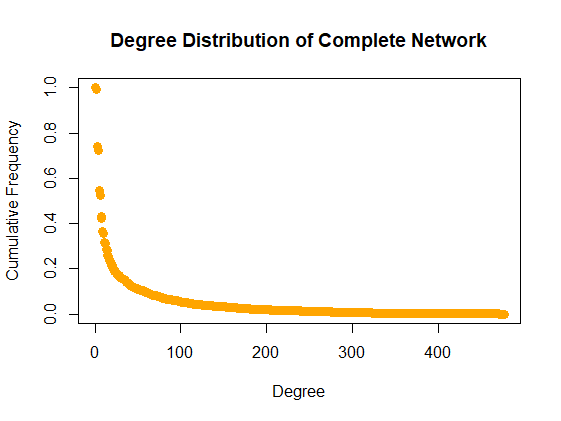
On looking at the network on the left, we can see that most of the vertices belong to a single cluster and on further investigating we found that around 98% of the vertices (3354) are present in a single cluster.

Hence, in order to further analyze this, we will create a subgraph consisting of this cluster (the one with the highest size - 98% of vertices).

The number of vertices and edges in the largest connected component are 150 and 451, respectively.

The network graph on the right displays the large connected component within the Air Route network-

The degree histogram and distribution were further analyzed for a complete network and largest connected component as shown below. Both the networks have few nodes with high degree but still the largest component is more balanced than the complete network. Majority of nodes in the network have less than 50 degrees.



In order to drive better insights, we also computed different network statistics like - highest degree nodes, routes, reciprocity, clustering coefficient, and diameter of the network. The below table displays the comparison of these statistics across the complete network and the largest component.

|  |  |  |
| --- | --- | --- |
| Statistics | Complete Network | Largest Component |
| Reciprocity | 0.97 | 0.95 |
| Transitivity | 0.24 | 0.10 |
| Diameter | 14 | 12 |
| Avg. path length | 4.14 | 4.45 |

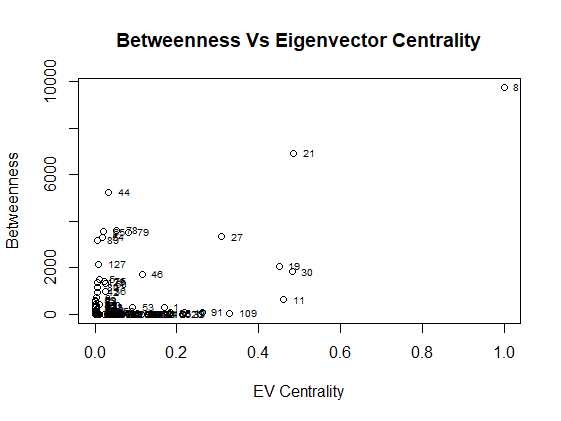
The longest shortest path of the complete network is 14 while that of the largest component is 12 which implies that the complete network is more connected. The global clustering coefficient or the transitivity for complete network is higher than the largest component thus two nodes with common connection are also connected. That is if there is a route from A to B and A to C then it is more probable in the complete network than in the largest component that there will be a route from C to B. The reciprocity, which means if there is a route from A to B then there will be a route from B to A, is higher in the complete network. However, the average path length of the largest component is higher than the complete network.

Two nodes or airports with the highest degree in the complete graph are FRA - Frankfurt airport and CDG - Paris Charles De Gaulle while the two airports with the highest degree in largest component are ATL -Atlanta airport and LAX -Los Angeles airport.

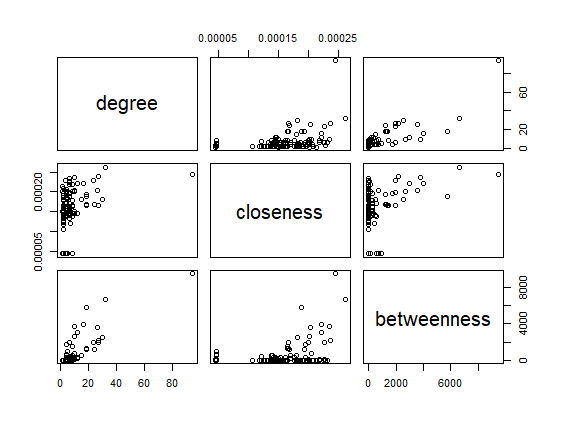
|  |  |  |
| --- | --- | --- |
| Centrality - Largest Component | Highest Value | Node/Edges |
| Closeness Centrality | 0.0025 | LAX -Los Angeles |
| Vertex Betweenness Centrality | 9739.54 | ATL -Atlanta |
| Edge Betweenness Centrality | 3176.66 | ICN Seoul Incheon->SFO San Francisco |
| Eigenvector Centrality | 1.00 | ATL-Atlanta |

**Centrality Measures**

We have also studied the nodes and edges with highest centrality measures as shown below:



Further, we created a plot between Betweenness and Eigen Vector centrality (as show on the right). From the figure, we can say that node 8 (Atlanta Airport) and node 21 (Los Angeles Airport). have the highest betweenness and eigenvector centrality. Hence, both the airports play an important role in the network.



For further analysis, we also created a graph (on the left) to analyze trends of degree, closeness and betweenness

From the graphs, we can say that on increasing the degree centrality, the betweenness centrality for those nodes - is increasing while the closeness centrality is relatively stagnant.

**Hubs and Authority**

For our further analysis on the largest connected component, we computed the hub and authority scores for the vertices and below is the tabular view of top 10 nodes having highest hub and authority scores, respectively.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Airport | ATL | LHR | LAX | ORD | JFK | SFO | MSY | MIA | DEN | LAS |
| Hub Scores | 1.00 | 0.47 | 0.46 | 0.40 | 0.39 | 0.29 | 0.29 | 0.23 | 0.19 | 0.19 |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Airport | ATL | LAX | ORD | JFK | LHR | MSY | SFO | MIA | BOS | DEN |
| Authority Scores | 1.00 | 0.52 | 0.510 | 0.51 | 0.49 | 0.36 | 0.32 | 0.29 | 0.24 | 0.24 |

A Hub is basically the vertex with high outdegree i.e. a vertex pointing to a high number of vertices. From the above table, we can see that ATL (Atlanta) airport has the maximum hub score i.e. the most number of outgoing routes. Atlanta airport is followed by LHR (London Heathrow Airport) in terms of the hub score and we can see that the Chicago O’Hare Airport is at number 4 in terms of the hub score.

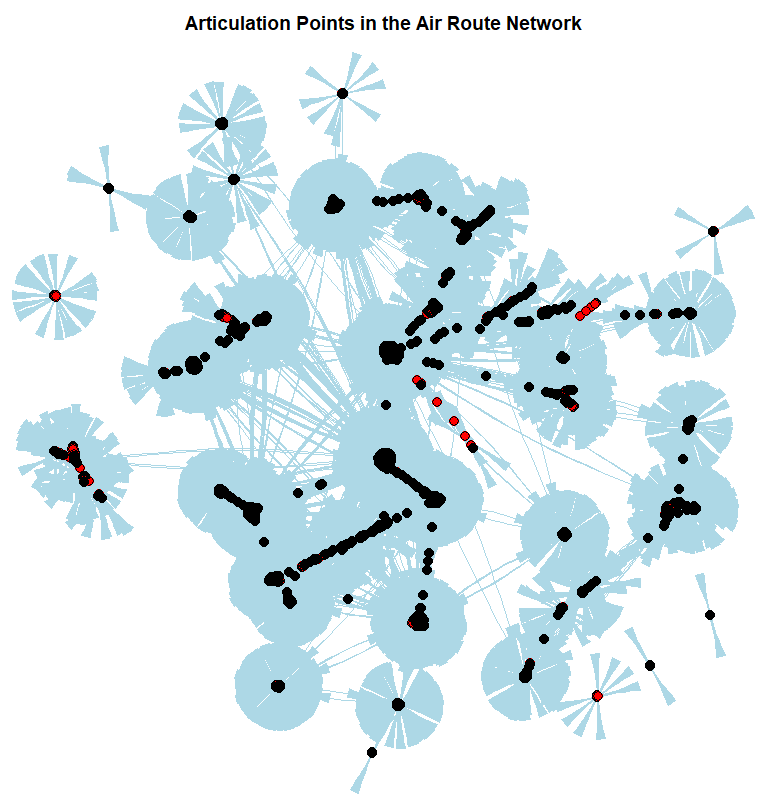
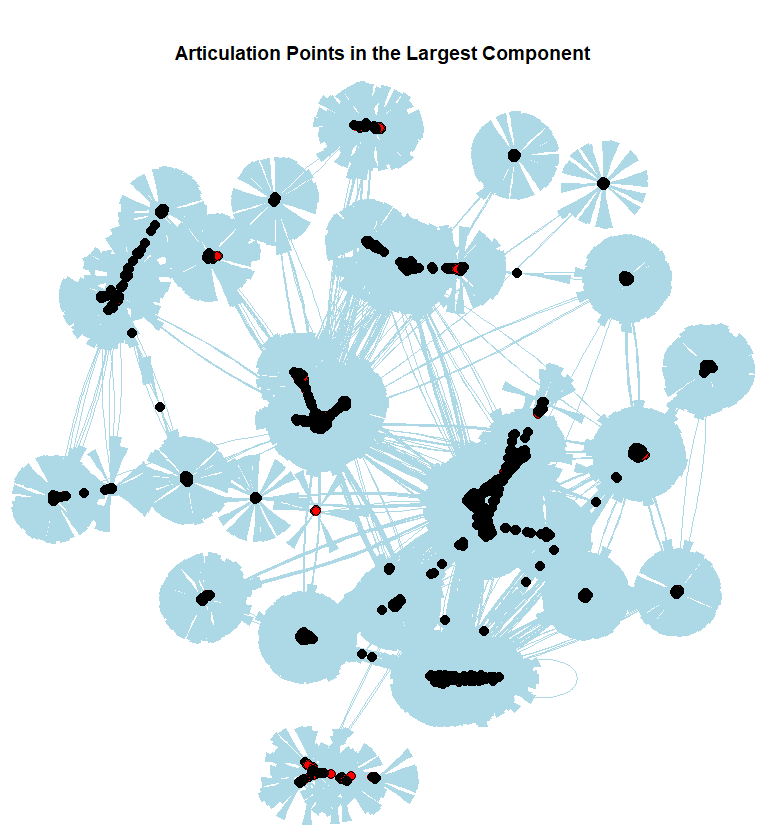
On the other hand, authority is basically a vertex with a high indegree i.e. a vertex getting linked by a high number of vertices. From the table above, we can see that ATL (Atlanta Airport) has the highest authority score as well which means that it has the most number of incoming routes as well. ATL is followed by Los Angeles Airport and the Chicago O’Hare Airport is at number 3 in terms of the authority score.

We also analyzed this network in terms of the hub and authority score of the vertices (as below) wherein the node size represents the hub/authority score (multiplied by a factor of 50 for better visualization).

Hence, we can say that these hubs and authorities play an important role in the network as they are responsible for maximum information flow within the network (in our case maximum incoming and outgoing routes).

**Articulation Points**

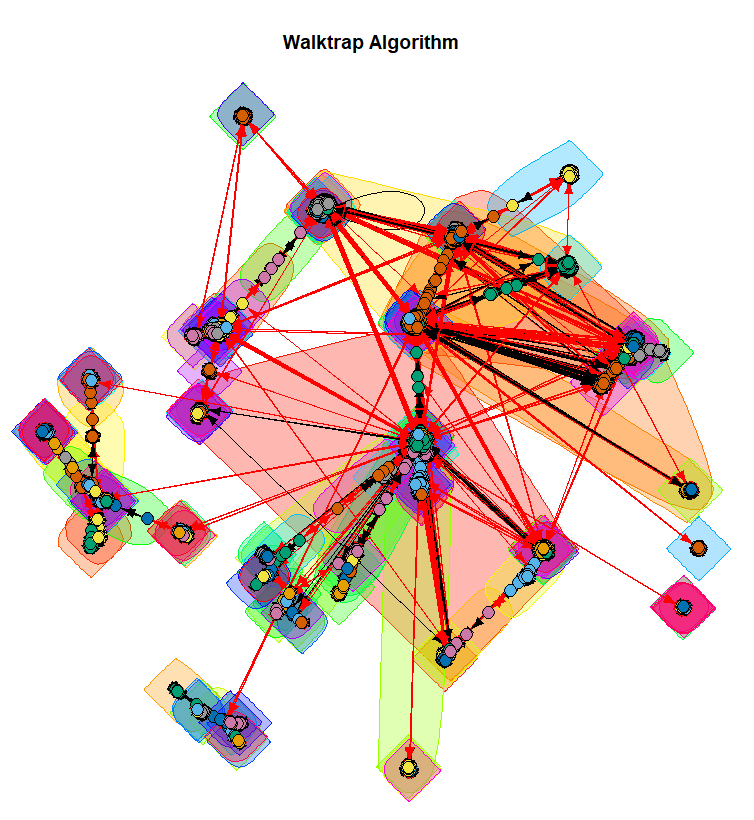
Articulation points are basically the vulnerabilities in the network, i.e. an articulation point is a vertex whose removal disconnects that network. These points are really important in a sense that if removed they can prevent the flow of information from one component of the graph to the other. These points are computed by brute-force method wherein each node is removed and the number of connected components within the network are counted. For our case, we have used the function ‘*articulation.points()*’ function.

The below plot displays the articulation points as highlighted in red. Also, the color, width of edges and size of nodes have been updated make the graph more interpretable.

The original Air Route network had 362 articulation points whereas the largest component has 358 articulation points or vulnerabilities. These points can therefore be used for designing more reliable route networks. Hence, we can say that the original air route network has more vulnerabilities when compared to the largest connected component.

**Community Detection**

Being a directed graph, we will use the walk trap algorithm to detect different communities within the network. In order to get a meaningful analysis, we will focus on the giant connected component instead of the complete network.



The plot on the right shows - the communities using the Walk trap algorithm. Note that the size of nodes, width of edges, arrows have been updated to make the graph more interpretable.

**Insights and Conclusion**

Post analyzing the Air Route Network – we came up with the following insights/conclusions:

* The largest connected component within the network consists of around 98% of the airports (3354 airports).
* Majority of the airports within the network have a degree of less than 20 i.e. most of the airports are connected to less than 20 other airports.
* Airports with the highest degree in the complete network are FRA - Frankfurt airport and CDG - Paris Charles De Gaulle whereas in the largest component we have - ATL -Atlanta airport and LAX -Los Angeles airport.
* Atlanta and LA Airport have the highest betweenness and Eigen vector centrality i.e. these two airports are important in a sense that these airports are linked between other airports most of the time.
* Atlanta Airport followed by London Heathrow Airport are the best hubs wherein Chicago O’Hare airport lies at number 4
* The Air Route network is more vulnerable as compared to the largest connected component within this network.

**Results**

**Bibliography**