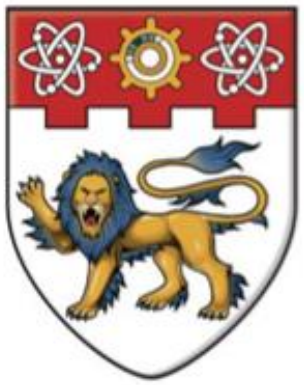

NANYANG TECHNOLOGICAL UNIVERSITY

MH8351 Web Analytics

Project Assignment Report

DATE

8 February 2020

TEAM MEMBERS

Sharmilly Thurai Raja (G1901884G)

Shingar Aggarwal (G1901889C)

Tham Chay Yong Gerald (G1902212G)

Teo Shin Siong Japheth (G1800706E)

Lee Wei Yang Shawn (G1801333C)

CONTENTS

1 INTRODUCTION

We analyze the global structure of the worldwide air transportation network, a critical infrastructure with an enormous impact on local, national, and international economies. The objective is to identify communities and key players in the international aviation network. Several popular algorithms were applied, and a comparison between the various algorithms' accuracy and applicability in aviation networks was done. We also attempted to derive insights from the network characteristics, such as degree distribution, clustering coefficient distribution, node betweenness, closeness centrality, authority and hub nodes.

2 DATA SOURCE

http://openflights.org/data.html

- Passenger flights data only
- Covers over 60,000 routes and 3000 airports
- Airport and airlines data updated as of January 2017
- Routes data updated as of June 2014

3 DATA PREPROCESSING

From the data source, we have 3 csv files: routes.csv, airlines.csv and airports.csv. In order for us to work with the data given, we will need to select certain columns from these routes.csv and airports.csv. We selected the source countries, destination countries, latitudes and longitudes of the source and destination countries for our project.

From the 6 columns, there were null values in the source or destination countries, represented by "\N" in the excel file. We dealt with the null values by removing them from the data set. The betweenness score of the network was computed at this point.

Furthermore, we have tuples where the source and destination countries were the same. This is presumed to be internal domestic flights and we removed these tuples as well in order to increase model performance and efficiency. As a result, we have about 33058 routes in the remaining tuples.

The data preprocessing method applied for the PageRank algorithm differs slightly and is elaborated below.

4 EXPLORATORY DATA ANALYSIS

From the remaining 33058 routes, we plotted the countries as nodes and the route as links in R.

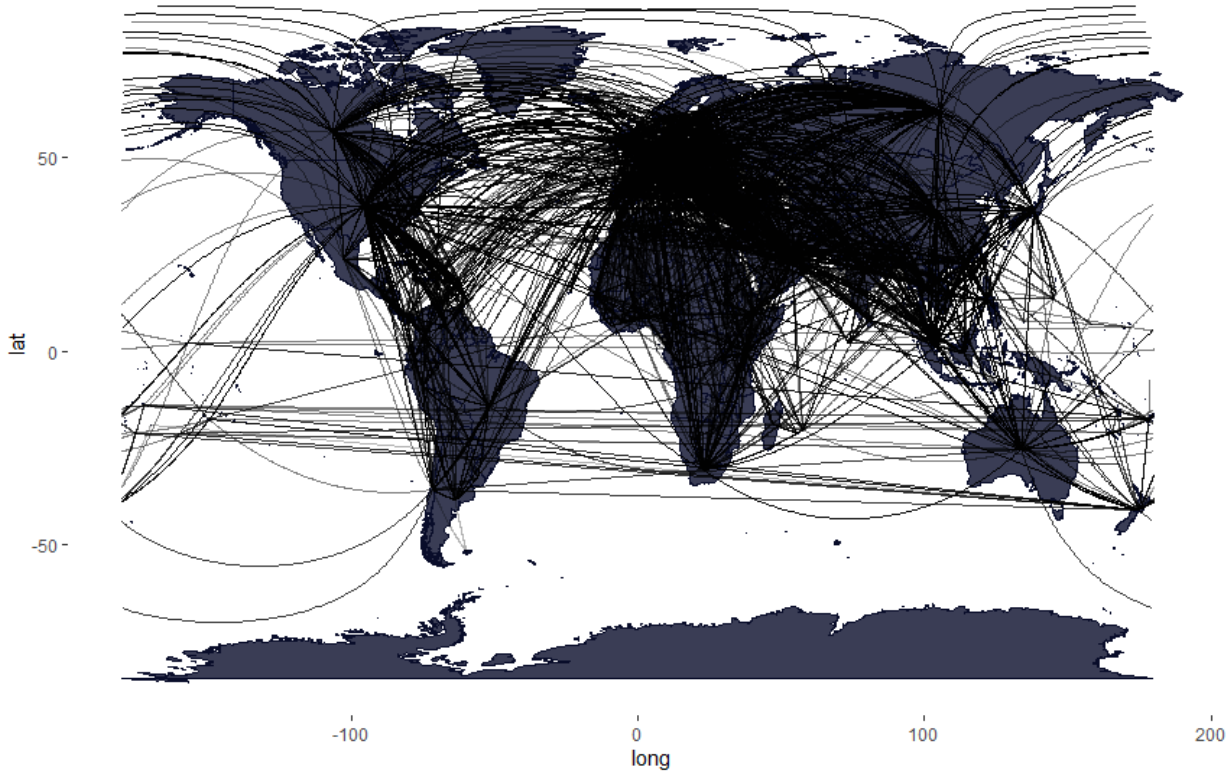

Figure 1. Aviation Network from Open Flights

From the preliminary findings, we can see that the routes are mainly concentrated in Europe and USA. We will discover more about the key players in our problem statement later in the report finding.

Using the igraph package in R, the following network summary statistics were derived:

- Number of nodes: 3330
- Number of edges: 67240
- Average total degree: 40.38
- Average in-degree/out-degree: 20.19
- Mean path length: 4.07
- Network Diameter: 13
- Global Clustering Coefficient: 0.25

From the degree distribution curve below, it closely approximates a power law, P(k)~k^-r. This means the aviation network is a scale-free network, where some airports act as highly connected hubs, but most airports are of low degree.

This also suggests that the phenomenon of preferential attachment is at play where nodes tend to connect preferentially to nodes with already high degrees. In addition, the spreading of P(k) values at high k represents noise in the tail, due to very low number of observations at high k.

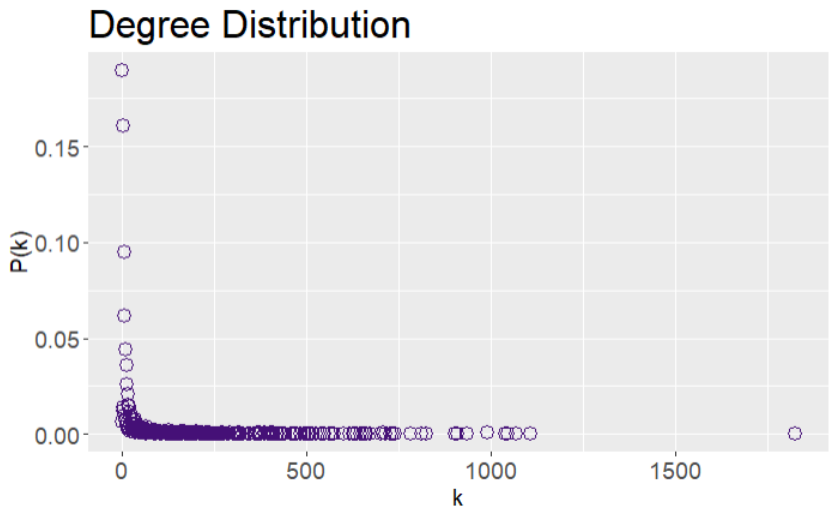

Figure 2. Degree Distribution

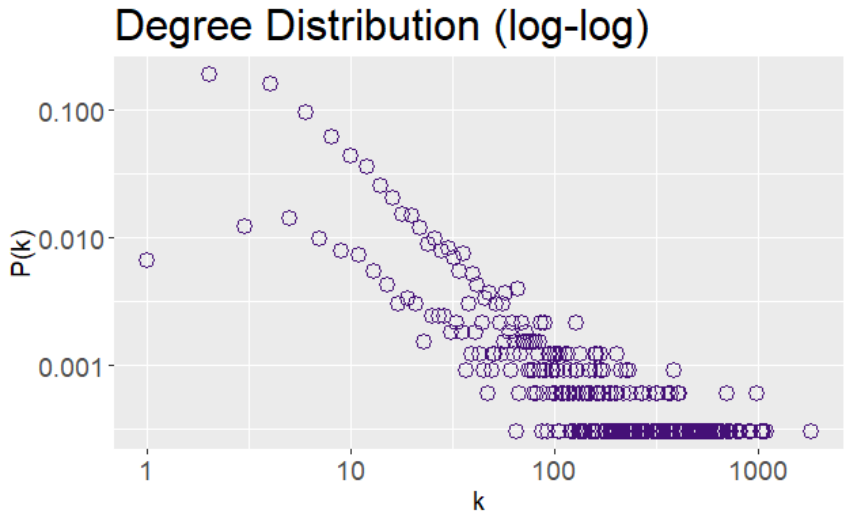

Figure 3. Degree Distribution (log-log)

The clustering coefficient characterizes the node's tendency to form clusters or groups. The clustering coefficient distribution plotted below strongly resembles a hierarchical network. In such networks, it is assumed that clusters combine in an iterative manner to account for the coexistence of modularity, local clustering, and scale free topology.

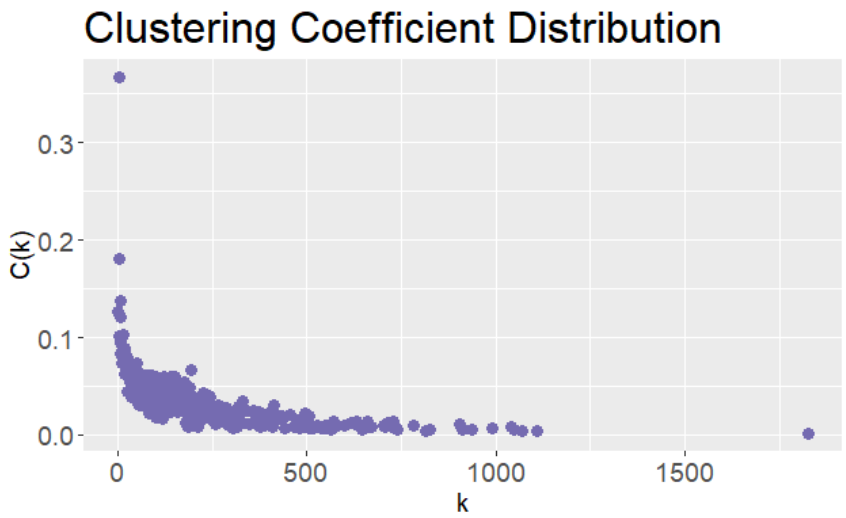

Figure 4. Clustering Coefficient Distribution

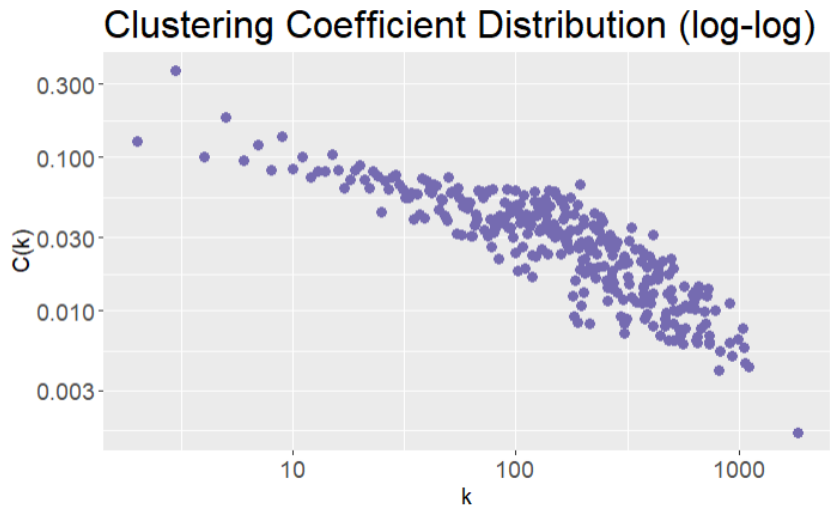

Figure 5. Clustering Coefficient Distribution (log-log)

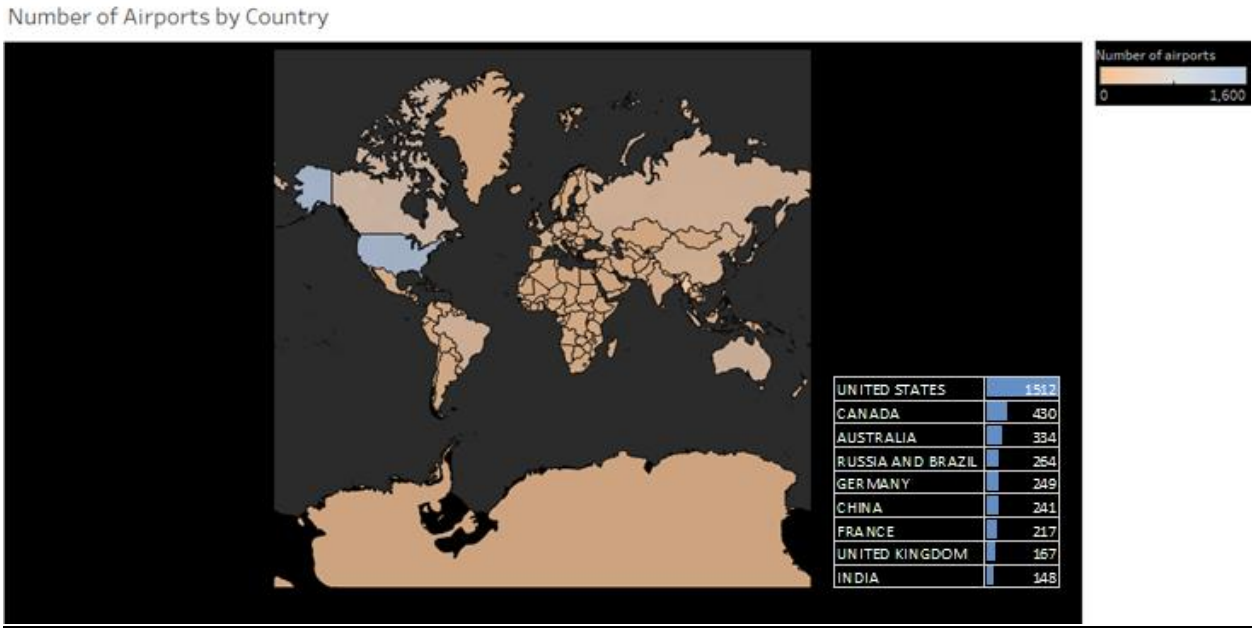

Figure 6. Number of Airports by Country

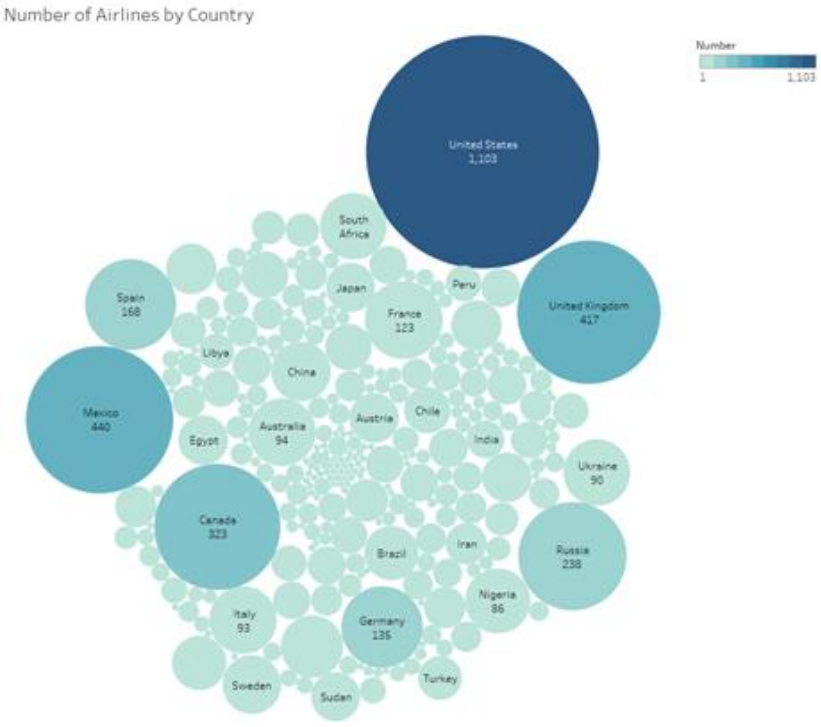

Figure 7. Number of Airlines by Country

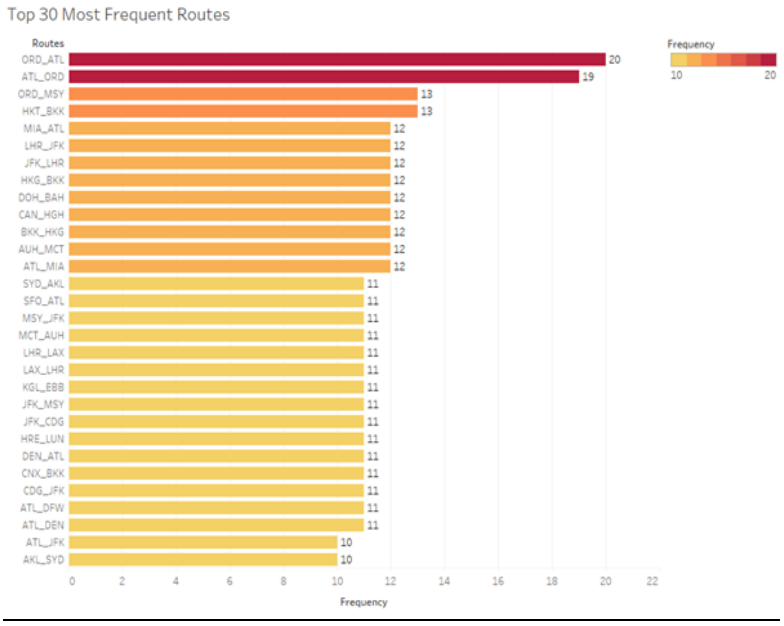

Figure 8. Top 30 Most Frequent Routes

Figure 6, 7 and 8 are constructed using Tableau. Figure 6 shows the map based on the longitude and latitude. Color shows sum of numbers of airports and the details are shown for the countries with the ten highest number of airports. From this figure it can be concluded that the USA has the highest number of airports. Figure 7 displays the number of airlines by country. Color and different sizes depict the sum of numbers of airlines. The marks are labelled by country and the sum of numbers. Apart from having the highest number of airports, USA also has the highest number of airlines. Figure 8 is created to visualize the frequency of the top thirty routes. Each bar chart is labelled by frequency of the specified route. The route from ORD to ATL has been travelled most by the airlines.

5 APPLYING THE PAGERANK ALGORITHM

The above map plot provided an understanding of popular routes. To further understand popular airports, multiple methods can be used.

One such simple method to determine popular airports is to count the number of flights that are handled. Most airlines make use of a hub-and-spoke system. However, counting flights inbound and outbound does not convey how important an airport is for routing airline traffic in general. This is due to the fact that certain hub airports might be a critical juncture for flights belonging to a certain airline (e.g. Singapore is a critical juncture for Singapore Airlines and all SQ flights will go via or from Singapore), and as a result, those hub airports might be considered more important than they are even if they have identical or even less flight counts.

As a result, we decided to model the data as a graph and make use of the PageRank algorithm to determine popular airports.

The PageRank algorithm is used by Google to rank web pages for an entered search query. The algorithm outputs a probability distribution used to represent the likelihood that a person randomly clicking on the links will arrive at any particular page. PageRank can be calculated for collections of documents of any size. The probability distribution is evenly divided among all documents in the collection at the beginning of the computational process. The algorithm undergoes several iterations through the collection to adjust approximate PageRank values to more closely reflect the theoretical true value.

The routes.csv and airports.csv file was imported and all null values represented by "\N" were replaced with NaN values. This was to allow us to have a visibility of the number of null values under each column for both data files.

The airports file consists of the following columns. The IATA code is a unique three-letter code that identifies an airport or its location. We will be making use of this dominantly in the algorithm.

```
airports.columns
```

```
Index(['AirportID', 'AirportName', 'City', 'Country', 'IATA', 'ICAO',
       'Latitude', 'Longitude', 'Altitude', 'Timezone', 'DST', 'Zone', 'Type',
       'Source'],
      dtype='object')
```

Figure 9. airports.csv Column Fields

The routes file consists of the following columns. The OrgIATA is the code for the airport of origin from which the flight takes off and DstIATA is the code for the destination airport at which the flight lands.

```
routes.columns
```

```
Index(['AirlineCode', 'AirlineID', 'OrgIATA', 'SourceAirportID', 'DstIATA',
       'DestinationAirportID', 'Codeshare', 'No.ofStops', 'Equipment'],
      dtype='object')
```

Figure 10. routes.csv Column Fields

For data cleaning, the airports data was checked for duplicate IATA codes (there were none), before IATA codes will null values were omitted. For the routes data, OrgIATA and DstIATA codes found in the hash table for the airports data is kept and the rest are omitted. Airports are referred to as nodes while the routes are referred to as edges.

Dead-end nodes, nodes which do not have any outgoing routes and non-incoming routes are identified.

The network is represented as a data frame which can be used as a hash table as well. The final result consists of the destination airports that have incoming routes. This is to allow for fewer iterations required to calculate the pagerank of each node.

The convergence and computational time of the algorithm depends on multiple aspects such as the number of nodes, damping factor and more. However, from literature, we have found that if a network is linear and the information of each node can be accessed linearly, the number of nodes affect the speed of computation but not the number of iterations. The damping factor, l is said to be proportional to the speed of convergence. The smaller the damping factor, the faster

the convergence. We decided to use l=0.85 as in literature it is the most commonly used damping value.

The algorithm provided us with the following results:

	Airport_Rank	Airport_Name
0	0.008312	Hartsfield Jackson Atlanta International Airport
1	0.005256	Chicago O'Hare International Airport
2	0.005100	Los Angeles International Airport
3	0.004848	Dallas Fort Worth International Airport
4	0.004450	Singapore Changi Airport

Figure 11. PageRank Algorithm Results

6 PROBLEM STATEMENTS

6.1 WHAT ARE THE TOP 3 MOST POPULAR AIRPORTS (MOST CONNECTED (HIGHEST DEGREE))?

Using the degree measure on the nodes, we have found out that the following airports are the top 3 in terms on highest degree.

1. USA (Chicago, O'Hare Airport)
2. UK (London Heathrow Airport)
3. France (Paris-Charles De Gaulle)

Reasons and insights to why these countries are the top 3 on the list:

1. O'hare not only serves commercial flights, but also military flights
2. O'hare is home to the largest airlines in USA and was ranked biggest mega hub in USA in 2017. This means that it has the highest connections between inbound and outbound flights within any 6-hour window. This is consistent with our findings.

6.2 WHICH NODES FORM A COMMUNITY? WHAT IS/ARE THE COMMUNITY STRUCTURES AND ANY SPECIAL FEATURES OF THE COMMUNITY?

To detect different communities, we applied the walktrap community algorithm and the leading eigenvector community algorithm.

With the walktrap community algorithm, we got a total of 13 communities. However, there are 5 main communities as there are 8 communities with very few members in it.
The main communities are:

1. America (comprising of members such as Columbia, United States, Argentina, Mexico, Brazil etc.)

2. Africa (comprising as members such as Somalia, South Africa, Ethiopia, Rwanda etc.)
3. Europe (comprising of members such as Italy, France, United Kingdom etc.)
4. Asia (comprising of members such as China, Hong Kong, Singapore, Japan etc.)
5. Middle East (Egypt, Afghanistan, India, Pakistan etc.)

It is surprising that India was grouped with the Middle Eastern countries and not Asia. This might be due to its location as it is nearer to the Middle Eastern countries.

With the leading eigenvector community algorithm, we got a total of 5 communities. However, this algorithm is less discerning as it groups countries that are much further apart together as well. (E.g. it grouped South Korea with Madagascar, New Zealand and Bhutan together.) These countries are rather far apart. This suggested to us that this algorithm is less than optimal.

We used the modularity score to compare the 2 algorithms that we used. Modularity reflects the concentration of edges within modules compared with random distribution of links between all nodes regardless of modules. For the walktrap community algorithm, the modularity score is 0.4366818 while the score for the leading eigenvector community is 0.4342501. While the scores are roughly similar, the score for walktrap community is slightly higher, which shows is in line with our conclusion.

Please refer to the codes for the list of the communities and the membership.

6.3 What Are The Top 10 Airports That Have The Highest Betweenness Scores? Why So?

The betweenness centrality in this project defines and measures the importance of an airport in the aviation network based on how many times it occurs in the shortest path between all pairs of airports in the network. The betweenness() function from the igraph package was used.

The top 10 airports with the highest betweenness scores are shown in Table 1 below. This essentially meant that these airports serve as very important hub and authority nodes in the aviation network.

Table 1. Top 10 Betweenness Scores

Rank	IATA Code	Airport	City, Country	Betweenness Score
1	LAX	Los Angeles International Airport	Los Angeles, USA	837,419.9
2	FRA	Frankfurt Airport	Frankfurt, Germany	819,061.1
3	ANC	Ted Stevens Anchorage International Airport	Anchorage, USA	779,059.3
4	CDG	Charles de Gaulle Airport	Paris, France	691,502.8
5	LHR	London Heathrow Airport	London, UK	610,857.4
6	ORD	O'Hare International Airport	Chicago, USA	610,586.4
7	DXB	Dubai International Airport	Dubai, UAE	587,573
8	SIN	Singapore Changi Airport	Singapore, Singapore	574,032
9	PEK	Beijing Capital International Airport	Beijing, China	554,409.8
10	GRU	São Paulo/Guarulhos–Governador André Franco Montoro International Airport	Sao Paulo, Brazil	510,378.7

Some interesting observations were made:

- The top 3 airports (Chicago O'Hare, London Heathrow, and Paris Charles de Gaulle) with the highest degree network measure (refer to section 6.1) are not the same top 3 here. This means that while Chicago, London and Paris airports has the most number of aviation connections, they are not the top three when it comes to exclusivity of air routes. Los Angeles, Frankfurt and Anchorage have the highest

betweenness scores because there is a high proportion of overall destinations in aviation travel that have to pass through these airports one way or another.

- Nonetheless, Chicago, London and Paris did not fare too badly in terms of betweenness because they were ranked 6, 5 and 4 respectively
- An interesting observation is the United States took up 3 spots in the top ten betweenness score whereas the other countries only appeared once. As such, the US can be said to be the leading key player in the aviation network. This can be further confirmed when a simple analysis on the total number of airports per country analysis was done based on the countries that appeared in this top 10, using the airports.csv raw data file (refer to Table 2).

Table 2. Number of Airports by Country

Country	**Number of Airports**
United States	1,512
Brazil	264
Germany	249
China	241
France	217
United Kingdom	167
United Arab Emirates	17
Singapore	6

6.3.1 BETWEENNESS SCORE RESULTS STUDY

To take a closer look at these results, two additional analyses were made to see if betweenness scores had a correlation to airport passenger numbers (2018) and cargo load (2018). Refer to Annex A for full results.

Two major assumption was made for these two analyses:

- While the aviation network consisted of data as of June 2014, the airport passenger and cargo load data used is in 2018. It is assumed that the proportion of passenger movement and cargo load versus the aviation network remain largely the same over the last few years.
- Only the top 10 airports with the highest betweenness scores were made in these analyses.

The first analysis was to observe if airports with higher betweenness scores served a higher number of passengers. The results are shown in Figure 12.

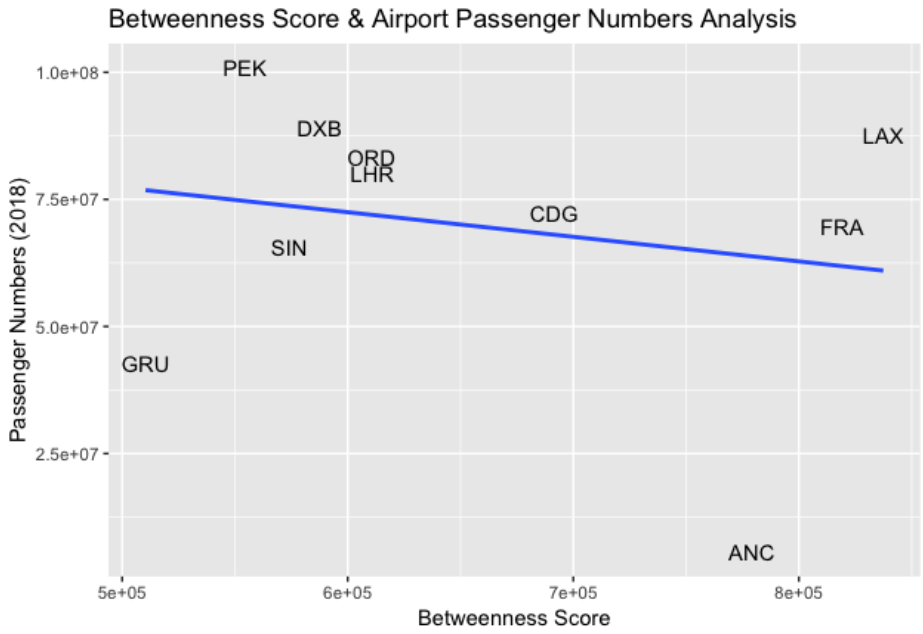

Figure 12. Betweenness Score vs Airport Passenger Numbers

The following observations can be made:

- As betweenness scores increased, passenger movements tended to decrease. Airports with higher betweenness scores did not mean higher passenger numbers.

- One hypothesis to explain this result can be that airports with very high passenger numbers and lower betweenness scores, such as Beijing and Dubai, tend to transit air passengers in large volumes to their final destination and this final destination has many alternate airport hub for the passenger to choose. These airports can be thought of as hubs with greater weights to nodes, which will also be reached by other hubs.

  On the other hand, airports like Los Angeles and Frankfurt, transport relatively fewer passengers on each trip but covers a greater variety of destinations that are exclusive. They can be thought of as hubs with smaller weights to nodes not commonly served by other hubs.

- Two outliers can be observed from this plot - Anchorage and Sao Paulo
  - Anchorage has a very high betweenness score but serve very few passengers. This can be because Anchorage in Alaska serve as the primary hub to connect passengers to other destinations in the state of Alaska. A further look into destinations served by this airport confirmed this point. From Figure 13 below, many destinations served by Anchorage airport are exclusive and exotic within the Alaskan state (such as Adak, St. Paul Island and Deadhorse). To get to these destinations by air, you have to transit at Anchorage Airport. Not surprisingly, the number of passengers on such routes is low. It is also interesting to note that out of the top three most popular airports obtained in section 6.1 using degree network measure, only Chicago have a direct connection to Anchorage.

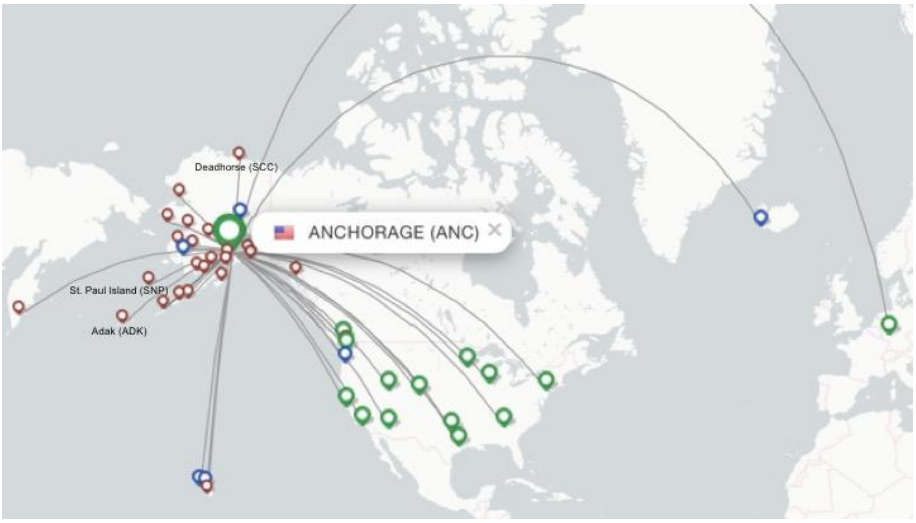

Figure 13. Ted Stevens Anchorage International Airport Destinations

- Sao Paulo comes in 10th in terms of the betweenness score. In terms of exclusivity to airport destinations, Sao Paulo pales in comparison to the other nine but wins all the other airports outside of the top ten. The number of passenger movements in this airport is substantially higher than Anchorage. As shown in Figure 14 below, Sao Paulo can be thought as an airport which has a delicate balance between serving a relatively high number of exclusive destinations (such as Rio Branco and Palmas) and serving other airport hubs (Chicago, London and Paris having the highest degree network measures are all served by Sao Paulo).

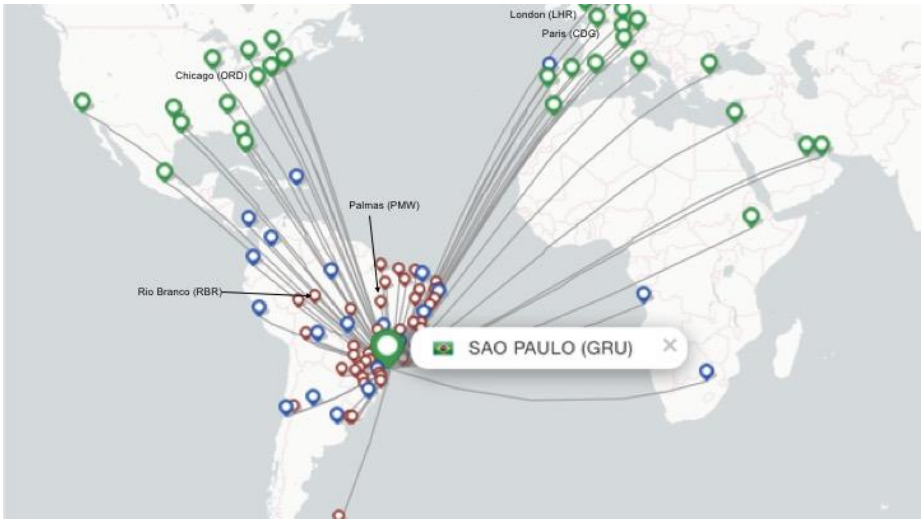

Figure 14. São Paulo/Guarulhos–Governador André Franco Montoro International Airport Destinations

The second analysis was to observe if airports with higher betweenness scores translated to more cargo being carried. The results are shown in Figure 15.

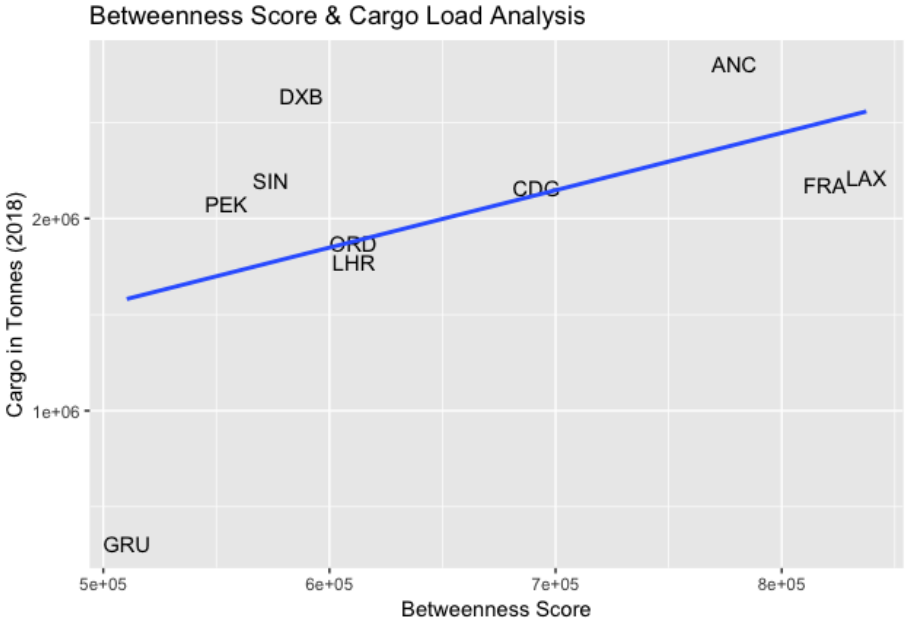

Figure 15. Betweenness Score vs Cargo Load

The following observations can be made:

- As betweenness scores increased, cargo load also increased. Airports with higher betweenness scores meant more aircraft cargo were being transported. From this result, it can be suggested that the movement of cargo is behaviorally different from the movement of passengers.
    - It can be suggested that the movement of cargo may be more point to point, taking place at multiple places until they get to their destination.
    - Passengers on the other hand have human needs and may opt to transit at fewer stops and their choice of an airport hub may be one with facilities and amenities for a more enjoyable aviation journey. These airport hubs can share common routes with other hubs to get them to their final destination, which explains in Figure 9 why an airport with a higher betweenness score did not translate into higher passenger numbers.
    - It can be suggested that some airports in the world are serving a higher cargo-to-passenger ratio than others. It is interesting to note that Anchorage carries the highest number of cargo in this list, even though their passenger numbers are very low, reinforcing this point. Further research shows that Anchorage, given its geographical location, is 9.5 hours away from 90% of the industrialised world, making it a popular pitstop for cargo aircraft.
- Sao Paulo appeared in the outlier once again. Its betweenness score is in the 10th position yet its cargo load does not quite match up with its peers. It can be suggested that the volume of passengers and cargo in and out of this airport is not as large, yet it serves a substantial number of exclusive destinations.

6.4 Applicability Of Closeness Centrality, Authority And Hub In The International Aviation Network

The authority and hub scores were obtained using the authority_score() and hub_score() functions in the igraph package, and the plot below shows that there is a very strong positive correlation between authority and hub scores. This means that each flight route roughly has the same number of incoming and outgoing flight for each pair of source and destination airports. This is logical as most of the passengers travel both ways unless there is a large scale permanent emigration. Furthermore capacity constraints in airports mean that there cannot be many more incoming flights than outgoing, or else it will form a major bottleneck for airport operations.

Therefore the authority and hub scores are not very useful as metrics for key player identification in the aviation network. It should be more applicable in the bibliographic network or social network analysis context.

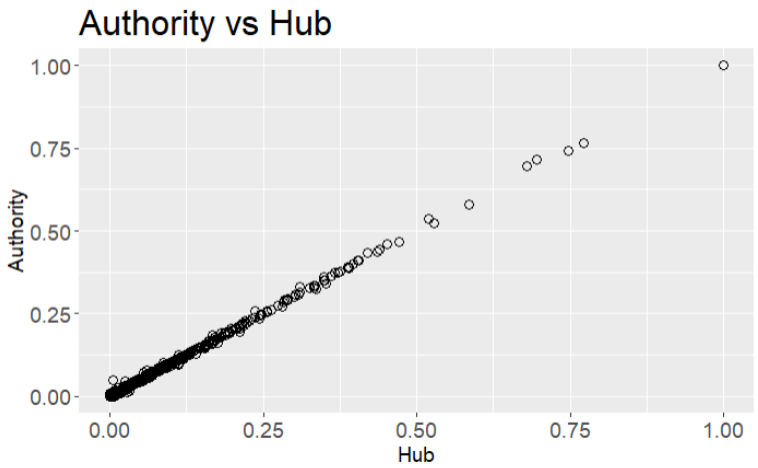

*Figure 16.*Authority vs Hub

Closeness centrality measures how easily a node can reach other nodes in the network. The closeness centrality was derived using the centr_clo() function of igraph package. From the table below of 10 airports with the highest closeness, most have a relatively low connectedness (degree of 2). The high closeness value is probably due to the limited routes to and from these

airports, which means they are disconnected from the larger, more connected network. Therefore, closeness centrality is not a suitable and accurate representation of the key players in this aviation network.

Name of Airport	Closeness	Degree	IATA	Country
Coffman Cove Seaplane Base	1.0	2	KCC	USA
Kongolo Airport	1.0	2	KOO	Democratic Republic of the Congo
El Porvenir	1.0	2	PVE	Panama
Conceição do Araguaia Airport	1.0	2	CDJ	Brazil
Larsen Bay Airport	1.0	2	KLN	USA
Amook Bay Seaplane Base	1.0	2	AOS	USA
Seal Bay Seaplane Base	1.0	2	SYB	USA
Unalaska Airport	1.0	6	DUT	USA
Charlotte Amalie Harbor Seaplane Base	1.0	6	SPB	Virgin Islands
Christiansted Harbor Seaplane Base	1.0	2	SSB	Virgin Islands
Eros Airport	1.0	2	ERS	Namibia
Grand Canyon West Airport	1.0	2	GCW	USA
Boulder City Municipal Airport	1.0	2	BLD	USA

Figure 17. Closeness Centrality Results

7 CONCLUSION

Our analysis of the community structure of the air transportation network is important for two additional reasons. First, it allows us to identify the most efficient ways to engineer the structure of the network. Specifically, having identified the communities, one can identify which ones are poorly connected and the ways to minimize that problem. Second, cities that connect different communities play a disproportionate role in important dynamic processes such as the propagation of infections such as the recent Wuhan corona-virus. As we described, finding the communities is the first step toward identifying these cities.

The betweenness score gave us a good overview of important key players (in terms of countries) and key nodes in the aviation network. A cross-reference with passenger and cargo volume data provided us interesting information on some airports and the behavioral differences between passenger and cargo travel.

8 PROJECT ACHIEVEMENT

Our project achievement includes:

- Using a variety of network measures (degree measure, degree distribution, scale-free networks, mean path length, network diameter, clustering coefficient), community detection algorithms (walktrap community, eigenvector community) and link analysis concepts (PageRank algorithm, betweenness centrality, closeness centrality, Authority and Hub) to analyze the aviation network
- Performing a deeper study from results of the betweenness score to further understand the underlying reasons by linking with external datasets.

9 RECOMMENDATIONS

Possible work for future improvement includes:

- Using passenger movement and cargo load from the same timeframe for the betweenness score results study
- Perform network analysis at a broader level: city level or country level

ANNEX A

Rank	IATA Code	Airport	City, Country	Passenger Movement (2018)	Cargo Movement in Tonnes (2018)
1	LAX	Los Angeles International Airport	Los Angeles, USA	87,534,384	2,209,850
2	FRA	Frankfurt Airport	Frankfurt, Germany	69,510,269	2,176,387
3	ANC	Ted Stevens Anchorage International Airport	Anchorage, USA	5,600,000	2,806,743
4	CDG	Charles de Gaulle Airport	Paris, France	72,229,723	2,156,327
5	LHR	London Heathrow Airport	London, UK	80,126,320	1,771,342
6	ORD	O'Hare International Airport	Chicago, USA	83,339,186	1,868,880
7	DXB	Dubai International Airport	Dubai, UAE	89,149,387	2,641,383
8	SIN	Singapore Changi Airport	Singapore, Singapore	65,628,000	2,195,000
9	PEK	Beijing Capital International Airport	Beijing, China	554,409.8	2,074,005
10	GRU	São Paulo/Guarulhos–Governador André Franco Montoro International Airport	Sao Paulo, Brazil	510,378.7	305,904

REFERENCES

- business.uzh.ch/dam/jcr:7ebaf68b-69a4-4f6e-9cdc-1d73c3fe0abc/srep30750.pdf
- https://stackoverflow.com/questions/9471906/what-are-the-differences-between-community-detection-algorithms-in-igraph/9478989#9478989
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1142352/
- https://www.sciencedirect.com/science/article/pii/S209252121730041X
- https://towardsdatascience.com/graph-analytics-introduction-and-concepts-of-centrality-8f5543b55de3
- https://www.flightsfrom.com/ANC
- https://airwaysmag.com/best-of-airways/ted-stevens-anchorage-international-airport/