

Flight Network Analysis

Student:

Li Liqian (G1800749K)

Tan Choon Kiat (G1700699B)

Tan Hong Guang (1800654K)

Course:

MH8351 Web Analytics

Contents

[Introduction 3](#_Toc32064191)

[Motivation 3](#_Toc32064192)

[Problem Statement 3](#_Toc32064193)

[Data Description 3](#_Toc32064194)

[Methodology 4](#_Toc32064195)

[Data Exploration 6](#_Toc32064196)

[Community Detection 11](#_Toc32064197)

[1. Spinglass 11](#_Toc32064198)

[2. Label Propagation 11](#_Toc32064199)

[3. Walktrap 11](#_Toc32064200)

[4. Louvain 12](#_Toc32064201)

[Results 12](#_Toc32064202)

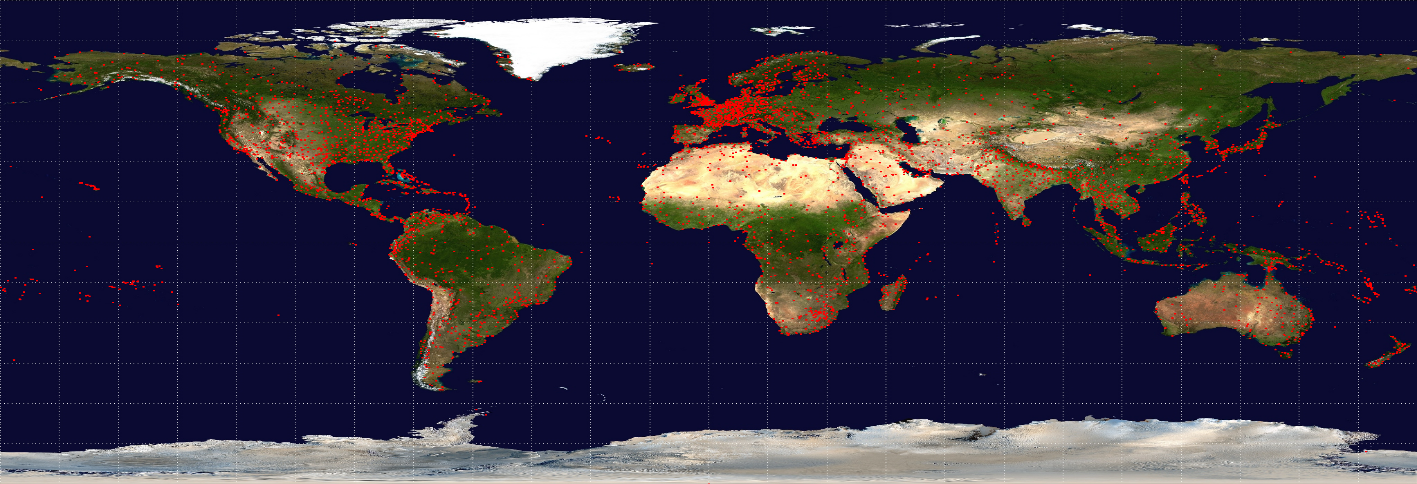
[Key Players 13](#_Toc32064203)

[Conclusion 17](#_Toc32064204)

[Limitations 17](#_Toc32064205)

[Future work 17](#_Toc32064206)

[References 18](#_Toc32064207)



# Introduction

## Motivation

Aviation started in the 18th century when the hot air balloon idea first began, an apparatus capable of atmospheric displacement through buoyancy. It then moved on to controlled gliding flying in 1896, and then the construction of the first ever airplane in the early 1900s.

In today’s world, air travel is one of the most important aspects of international travel as traveling via air flights is becoming more affordable. The present trends in air transport suggested that air passenger numbers could be double to 8.2 million in 2037 [1]. Identifying the key players within the aviation industry would be beneficial for airlines and airport operators, to optimally allocate their resources and time more efficiently.

## Problem Statement

We aim to identify Air Transport Hub(s) (i.e. Key Player(s)) in the world using OpenFlights [2] Airport and (Air Flight) Route datasets. After identifying air transport hubs, airlines and airport operators could optimise their resources and expand their businesses.

## Data Description

Airport and (air flight) Route datasets were extracted from OpenFlights. Airport Database contains over 10,000 airports, train stations and ferry terminals spanning the globe. Only *type=airport* is included in this project. Each entry contains the following information as shown below:

|  |  |
| --- | --- |
| **ID** | Unique OpenFlights identifier for this airport. |
| **Name** | Name of airport. May or may not contain the **City** name. |
| **City** | Main city served by airport. May be spelled differently from **Name**. |
| **Country** | Country or territory where airport is located. |
| **IATA** | 3-letter IATA code. Null if not assigned/unknown. |
| **ICAO** | 4-letter ICAO code. Null if not assigned. |
| **Latitude** | Decimal degrees, usually to six significant digits. Negative is South, positive is North. |
| **Longitude** | Decimal degrees, usually to six significant digits. Negative is West, positive is East. |
| **Altitude** | In feet. |
| **Timezone** | Hours offset from UTC. Fractional hours are expressed as decimals, eg. India is 5.5. |
| **DST** | Daylight savings time. One of E (Europe), A (US/Canada), S (South America), O (Australia), Z (New Zealand), N (None) or U (Unknown). |
| **Tz database time zone** | Timezone in ["tz" (Olson) format](http://en.wikipedia.org/wiki/Tz_database), eg. "America/Los\_Angeles". |
| **Type** | Type of the airport. Value "airport" for air terminals, "station" for train stations, "port" for ferry terminals and "unknown" if not known. *In airports.csv, only type=airport is included.* |
| **Source** | Source of this data. "OurAirports" for data sourced from [OurAirports](http://ourairports.com/data/), "Legacy" for old data not matched to OurAirports (mostly DAFIF), "User" for unverified user contributions. |

Route dataset was updated as of June 2014. There are over 3,000 airports and 60,000 air flight routes in the datasets. Each entry in the route datasets contains the following information as shown below:

|  |  |
| --- | --- |
| **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 of the source airport. |
| **Source airport ID** | Unique OpenFlights identifier for source airport。 |
| **Destination airport** | 3-letter (IATA) or 4-letter (ICAO) code of the destination airport. |
| **Destination airport ID** | Unique OpenFlights identifier for destination airport. |
| **Codeshare** | "Y" if this flight is a codeshare (that is, not operated by *Airline*, but another carrier), empty otherwise. |
| **Stops** | Number of stops on this flight ("0" for direct) |
| **Equipment** | 3-letter codes for plane type(s) generally used on this flight, separated by spaces |

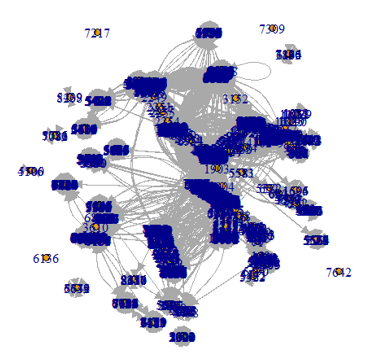
## Methodology

After exporting the datasets from OpenFlights, we dropped those records with NULL values for source airport ID field and destination airport ID field. We treated airports as nodes and created routes from source airport ID to destination airport ID in R Programming using *igraph* package. Preliminary data exploration was done to better understand the datasets. We further performed several community detection algorithms and used modularity as a measure of quality to select the best algorithm. Lastly, various evaluation metrics were utilised to identify air transport hub(s)/ key player(s) within the communities and analysis was done on the characteristics of the hubs.

# Data Exploration

Using *Route* dataset, we plotted *Graph 1*, which shows that the dataset is a disconnected graph with some airports not connected with the others. There are some highly connected sub-graphs as shown.

Graph 1: Network Graph with nodes (airports ID) connected by edges (routes)



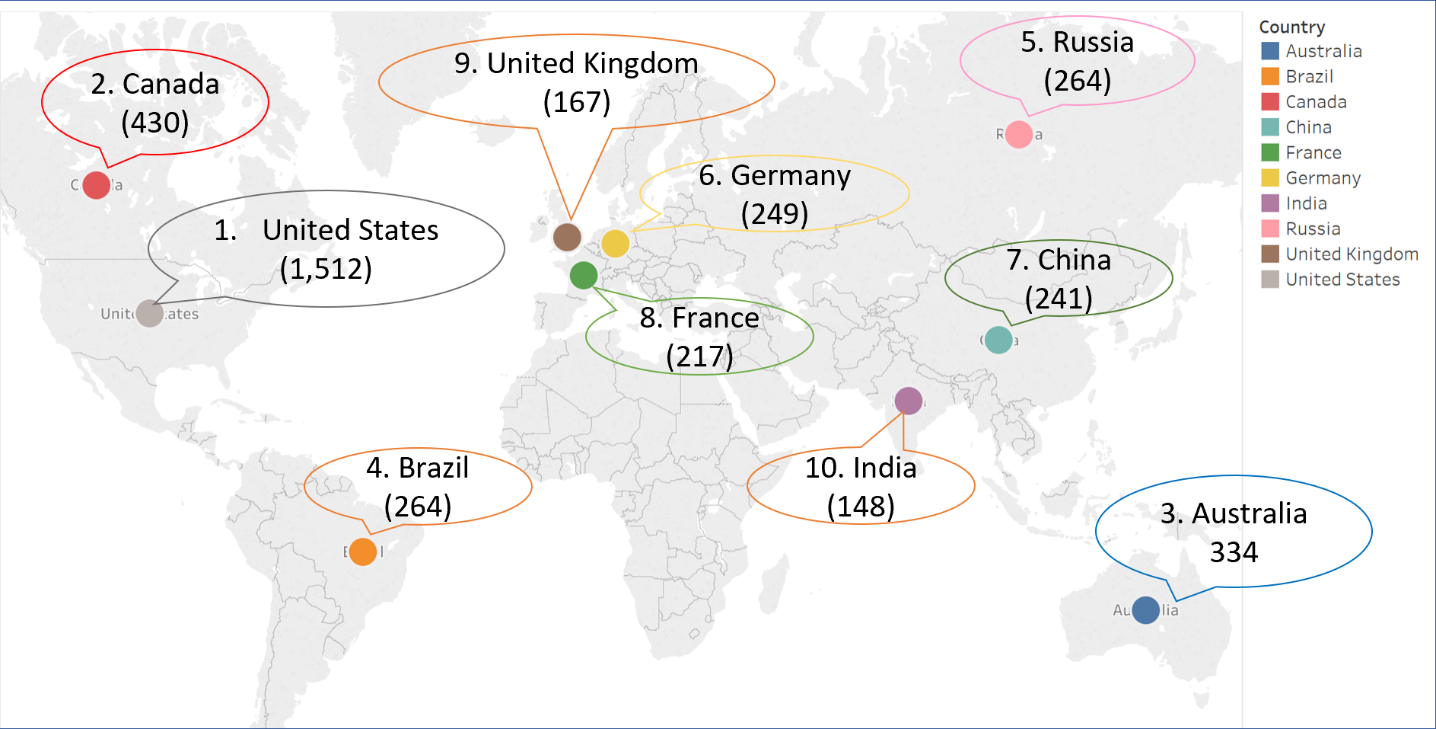
There are 3,218 airports (nodes) with 66,771 routes (edges) operated by unique 547 Airlines. We proceeded to analyse the datasets and obtain summary statistics for the network. *Table 1* shows the summary statistics of the network.

Table 1: Summary Network Statistics

|  |  |
| --- | --- |
| Average total degree | 41.55 |
| Average in/out degree | 20.78 |
| Mean path length | 3.98 |
| Network diameter | 13 |
| Density | 0.006 |
| Global clustering coefficient | 0.25 |

We then conducted further data exploration by evaluating the *Airport* dataset using Tableau and R programming. The airports are classified into different countries and *Figure 1* indicates that the United States has the highest number of airports (1,512) followed by Canada with 430 airports. This can be attributed to the perception that the United States is the richest and most powerful nation in the world which draws in a lot of people looking for business opportunities. As such, more airports were built to serve the incoming investors and visitors.

Figure 1: Top 10 Countries with highest number of Airports



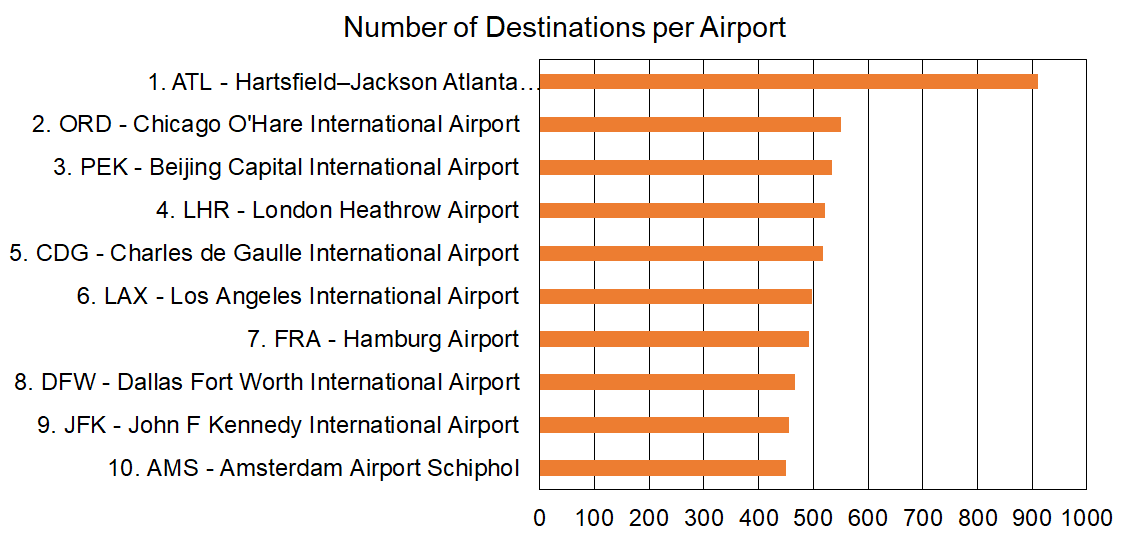
We conducted a preliminary analysis of Route dataset to identify the popular routes that airlines like to fly. Domestic air travels are the most popular routes that airlines fly, as shown in *Table 2* where top four routes are domestic flights. This could be more profitable to airlines as the traveling distance is shorter and more passengers would be likely to travel domestically than internationally.

Table 2: Popular Routes operated by different Airlines

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **S/N** | **Start Airport** | **Country** | **End Airport** | **Country** | **Number of Airlines** |
| 1 | Chicago O'Hare International Airport | United States | Hartsfield Jackson Atlanta International Airport | United States | 20 |
| 2 | Hartsfield Jackson Atlanta International Airport | United States | Chicago O'Hare International Airport | United States | 19 |
| 3 | Phuket International Airport | Thailand | Suvarnabhumi Airport | Thailand | 13 |
| 4 | Chicago O'Hare International Airport | United States | Louis Armstrong New Orleans International Airport | United States | 13 |
| 5 | Hamad International Airport | Qatar | Faro Airport | Canada | 12 |
| 6 | Abu Dhabi International Airport | United Arab Emirates | Muscat International Airport | Oman | 12 |
| 7 | Hong Kong International Airport | Hong Kong | Suvarnabhumi Airport | Thailand | 12 |
| 8 | Guangzhou Baiyun International Airport | China | Hangzhou Xiaoshan International Airport | China | 12 |
| 9 | Miami International Airport | United States | Hartsfield Jackson Atlanta International Airport | United States | 12 |
| 10 | Hartsfield Jackson Atlanta International Airport | United States | Miami International Airport | United States | 12 |

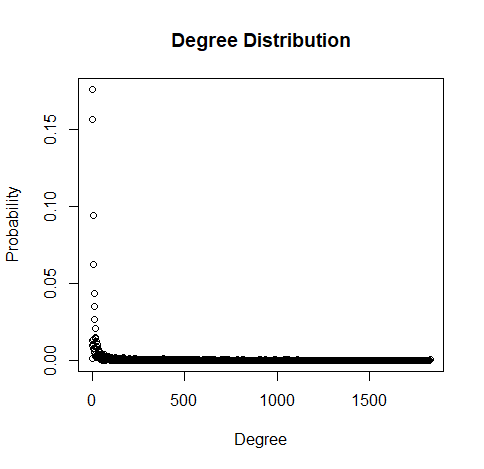
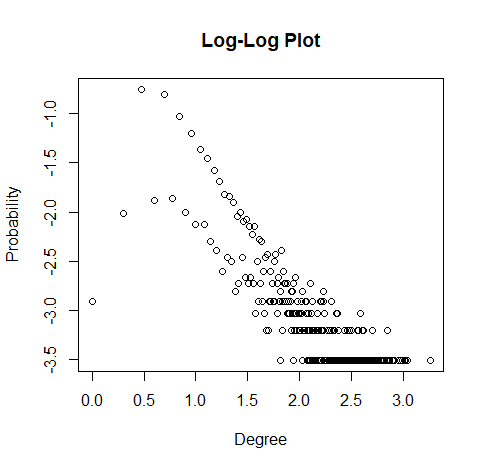
*Figure 2* shows that Hartsfield-Jackson Atlanta International Airport is the most popular and busiest airport as an airport destination by most of the airlines. This could give us a heads up that the airport may be an air transport hub within the region. We will further analyse the datasets later in the report. It is also worth noting that Beijing Capital International Airport is the third most popular and Asia busiest airport based on our datasets as China is developing Beijing Capital to attract more tourists and investors to their country.

Figure 2: Number of Destinations per Airport



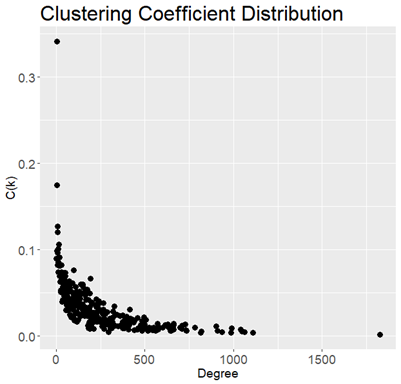
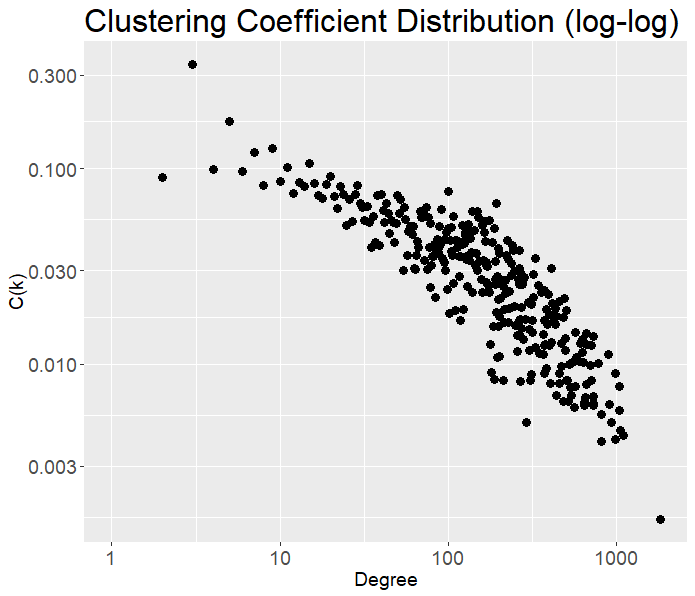
*Figure 3* shows that the degree distribution mimics a scale free network and is similar to the power law. This means that some nodes are highly connected hubs, which is also shown in *Figure 1*, although most of nodes are of low degree and there exist few hubs with many service routes by airlines. This also suggests that the new nodes (airports) tend to connect to nodes (airports) with already high degrees.

Figure 3: Degree Distribution and Degree Distribution (Log-Log) Plot

The clustering coefficient distribution (Log- Log) is approximately linear (with r approximating 1) as shown in *Figure 4*. This shows that it could portray itself as a typical hierarchical network. As the clustering coefficient are mainly below 0.1, which could indicate that most of the neighbours of each node does not connect with one another.

Figure 4: Clustering Coefficient Distribution and Clustering Coefficient Distribution (Log-Log) Plot

# Community Detection

From the preliminary assessment of the data, we found that the graph network consists of nodes with higher degrees and we might be able to identify hubs within the graph. However, we have to detect communities within the network before we can find the key hubs (players).

Four algorithms were used for community detection using R programming. *Spinglass*, *Label propagation*, *Walktrap* and *Louvain*.

## Spinglass Algorithm

Spinglass algorithm is from statistical physics approach using the concept of particle spin states. In this model, each particle (i.e. node) can be in one of *c* spin states, and the interactions between the particles (i.e. edges) specify which pairs of nodes would prefer to stay in the same spin state and which pairs prefer to have different spin states. After a defined number of steps, and the spin states of the particles in the end define the communities.

Spinglass is generally a good detection method and can be used for negatively-weighted edges. However, it is not particularly fast and not deterministic due to the simulation itself.[3]

## Label Propagation Algorithm

Label Propagation algorithm is a simple approach in which every node is assigned with unique labels. The method proceeds iteratively by assigning the labels to the majority of their neighbors and stops when the label of each node is one of the most frequent labels in its neighborhood. [3]

Label Propagation is very fast but yields different results based on the initial configuration, which is decided randomly.

## Walktrap Algorithm

Walktrap algorithm is an approach based on random walks. Random walks are performed on the graph, where the walks are more likely to stay within the same community as there are only a few edges that lead outside a given community. Walktrap runs short random walks depending on parameters set and uses the results of these random walks to merge separate communities in a bottom-up manner.

Walktrap is a fast algorithm, performs well in the presence of lower in-degree to out-degree ratios (Orman and Labatut, 2009[6]) and at graph sizes as small as 100 (Pons and Latapy, 2006[7]), hence, it may not perform well for large network.

## Louvain Algorithm

The Louvain algorithm is a heuristic approach, based on the modularity measures, which measures the density of edges inside communities to edges outside communities. This algorithm consists of two phases that are repeated iteratively. First small communities are found by optimizing modularity locally on all nodes. Followed by each small community is grouped into one node. It stops when maximum modularity is obtained.[8]

The Louvain method is a simple, efficient and easy-to-implement method for identifying communities in large networks.[4] The method has been used with success for networks of many different types such as fast unfolding of communities in large networks,2008 [9]. It is one of the most widely used method today for detecting communities in large networks.[5]

## Results

Several measures for quantifying the quality of communities have been proposed such as Modularity, Betweenness Centrality, Clustering coefficient, Density etc.

In this project, modularity is used. Modularity is one of the most used known functions to quantify community structure in graph [10]. Larger modularity indicates better communities. The community structure would be better if the number of internal edges exceeds the expected number. Modularity is normalized and it ranges from -1 to 1. Negative value indicates there is no community structure, whereas positive value means that the number of edges within groups exceeds the number expected on the basis of chance.[11]

*Table 3* shows the number of communities detected and modularity for each of the four algorithms used.

Table 3: Number of communities detected and Modularity

|  |  |  |
| --- | --- | --- |
| **Methods/ Algorithms** | **No of Communities Detected** | **Modularity** |
| 1. Spinglass | 22 | 0.387 |
| 2. Label Propagation | 87 | 0.604 |
| 3. Walktrap | 186 | 0.629 |
| **4. Louvain** | **39** | **0.640** |

As shown in *Table 3*, Louvain algorithm has the highest modularity 0.640 with 39 communities detected. With this, we identify the top 3 largest communities using the Louvain method. The largest community is North America & South America. Followed by Europe, and then Asia & Oceania.

# Key Players

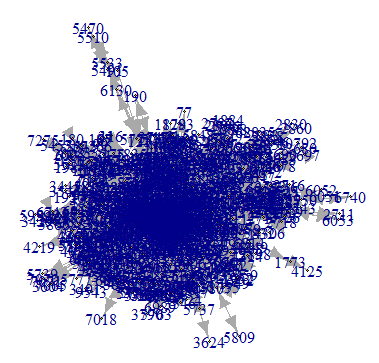
Different companies have their own motivations. From Airports’ perspective, it will allow them to learn from the key hubs on how they manage their flights, air navigation systems and procedures to handle the heavy air traffic. The country may also want to build additional airport in the vicinity of the hub to ease the air traffic load and eventually increase more flights to the country as new airport can cater more passengers. From Airlines’ perspective, they would prefer to park their planes in a key hub where there are more reliable services such as maintenance of their planes and refueling services. Also, key hubs would have more human traffic, leading to more potential passengers and increase in revenue.

The tables below show the key hubs identified for each of the 3 main communities that we previously identified using Louvain algorithm. They are ranked according to the various evaluation metrics (i.e. Degree, Closeness, Betweenness, PageRank, Authority score and Hub score).

Table 4: Key players within North America and South America (759 airport nodes)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Ranking** | **Airport identified** | **Degree** | **Closeness** | **Betweenness** | **PageRank** | **Authority score** | **Hub score** |
| **1** | **Hartsfield Jackson Atlanta International Airport** | **1,703**  **(Rank 1st)** | **0.533**  **(Rank 1st)** | **0.000970**  **(Rank 1st)** | **0.0415**  **(Rank 1st)** | **1**  **(Rank 1st)** | **1**  **(Rank 1st)** |
| 2 | Dallas Fort Worth International Airport | 834  (Rank 3rd) | 0.532  (Rank 2nd) | 0.000621  (Rank 2nd) | 0.0229  (Rank 3rd) | 0.476  (Rank 4th) | 0.466  (Rank 4th) |
| 3 | Chicago O'Hare International Airport | 900  (Rank 2nd) | 0.519  (Rank 4th) | 0.000530  (Rank 5th) | 0.0234  (Rank 2nd) | 0.593  (Rank 2nd) | 0.603  (Rank 2nd) |

Graph 2: Network Graph of the North America and South America Community

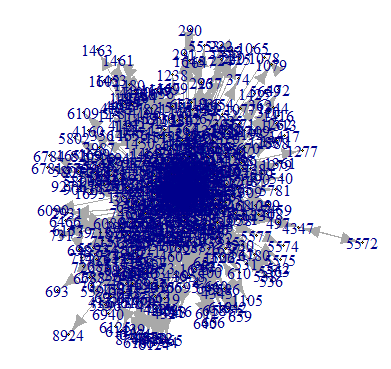


Hartsfield Jackson Atlanta International Airport has been the world's busiest airport by passenger traffic since 1998. It is the primary hub of Delta Air Lines, and is a focus city for low-cost carriers Frontier Airlines, Southwest Airlines, and Spirit Airlines. With just over 1,000 flights a day to 225 domestic and international destinations, the Delta hub is the world's largest airline hub.[12]

Table 5: Key players within Europe (601 airport nodes)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Ranking** | **Airport identified** | **Degree** | **Closeness** | **Betweenness** | **PageRank** | **Authority score** | **Hub score** |
| **1** | **Barcelona International Airport** | **706**  **(Rank 1st)** | **0.2633**  **(Rank 1st)** | **0.000206**  **(Rank 5th)** | **0.0155**  **(Rank 1st)** | **1**  **(Rank 1st)** | **1**  **(Rank 1st)** |
| 2 | Amsterdam Airport Schiphol | 604  (Rank 2nd) | 0.262  (Rank 2nd) | 0.000146  (Rank 10th) | 0.0130  (Rank 3rd) | 0.832  (Rank 2nd) | 0.844  (Rank 2nd) |

Graph 3: Network Graph of the Europe Community



Most of the traffic at Barcelona Airport is domestic and European, in which Vueling has an operational base. Low-cost airline traffic grew significantly, especially after the creation of operating bases by Vueling and Clickair at the airport.[13] Number of tourists increased significantly from 1990 to 2017 as shown below in *Figure 5*.

Figure 5: Number of Tourists in Hotels in Barcelona City from 1990 to 2017

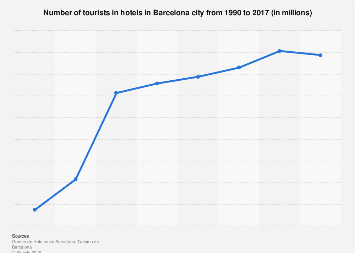
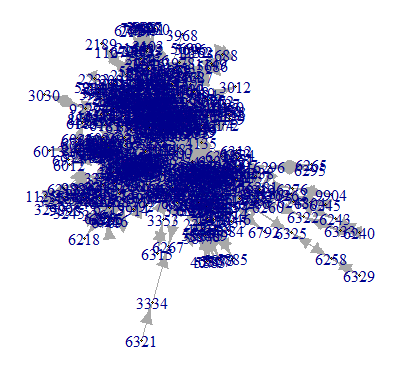


Table 6: Key players within Asia and Oceania (554 airport nodes)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Ranking** | **Airport identified** | **Degree** | **Closeness** | **Betweenness** | **PageRank** | **Authority score** | **Hub score** |
| **1** | **Singapore Changi Airport** | **496**  **(Rank 1st)** | **0.164**  **(Rank 3rd)** | **0.000596**  **(Rank 1st)** | **0.0262**  **(Rank 1st)** | **1**  **(Rank 1st)** | **1**  **(Rank 1st)** |
| 2 | Kuala Lumpur International Airport | 350  (Rank 5th) | 0.164  (Rank 1st) | 0.000314  (Rank 6th) | 0.0180  (Rank 7th) | 0.743  (Rank 2nd) | 0.748  (Rank 2nd) |

Graph 4: Network Graph of the Asia and Oceania Community



Singapore Changi Airport is one of the largest transportation hubs in Asia. Air traffic growth in the Asia-Pacific region is one of the fastest in the world. Singapore is strategically located at the heart of the region, and is well-placed to benefit from the growth potential of aviation-related activities.[14]

# Conclusion

## With the advancement of technology and in this closely connected world, Aviation has become an important industry that will last us many generations. While domestic airports generally tend to have higher degrees (due to edges being the number of flights) as compared to International airports, it may not be a good gauge to be considered as key hub since key hub intuitively should be a hub that connects with many different countries. Visualising air flights as a web graph, we could tell that it is a scale-free network, having few airports with many degrees and many airports with few degrees. Among the different community detection algorithms, the Louvain method seems to yield the best modularity. The communities detected seem intuitive as when visualised with a map, the communities seem to be based on continents. Different key hubs have their individual strong points which makes them a key player within their communities.

## Lastly, we have identified some of our project’s limitations and future works to further improve our analysis.

## Limitations

1. The edges we used in our analysis are not weighted, as we did not have comparable available data.
2. Database used did not have data on cargo flights.
3. More methods could be used in detecting communities

## Future work

1. More attention could be given to clusters which are very small. These could help airlines company to go to these destinations as other airlines did not cover these locations.
2. Classify flights as either Domestic or International flights. Airports with high degree may be misleading due to many Domestic flights.
3. Use other community detection algorithms such as Girvan-Newman, fast and greedy method or Leiden algorithm to see which better suits the aviation network.

# References

1. <https://www.iata.org/en/pressroom/pr/2018-10-24-02/>
2. https://openflights.org/data.html
3. <https://stackoverflow.com/questions/9471906/what-are-the-differences-between-community-detection-algorithms-in-igraph>
4. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5102890/>
5. <https://perso.uclouvain.be/vincent.blondel/research/louvain.html>
6. Orman G. K., Labatut V. (2009). A comparison of community detection algorithms on artificial Networks, in Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 5808 (Heidelberg: Springer; ), 242–256.
7. Pons P., Latapy M. (2006). Computing communities in large networks using random walks. J. Graph Algorithms Appl. 10, 191 10.7155/jgaa.00124
8. <https://en.wikipedia.org/wiki/Louvain_modularity>
9. Fast unfolding of communities in large networks  
   Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre  
   Journal of Statistical Mechanics: Theory and Experiment (10), P10008, 2008.
10. <https://en.wikipedia.org/wiki/Community_structure>
11. <https://en.wikipedia.org/wiki/Modularity_(networks)>
12. <https://en.wikipedia.org/wiki/Hartsfield%E2%80%93Jackson_Atlanta_International_Airport>
13. <https://en.wikipedia.org/wiki/Barcelona%E2%80%93El_Prat_Josep_Tarradellas_Airport>
14. <https://www.mot.gov.sg/about-mot/air-transport/aviation-hub>