CSC 495 – FINAL PROJECT

PSYDUCK GROUP

JESUS VALBUENA **KEVIN KAI CHUNG YING**

WORLD AIR TRAVEL NETWORK ANALYSIS

DePaul University

JUNE 06, 2017

Contents

Introduction

Air travel is a constantly growing industry that has doubled in size in the past decade per the International Air Transport Association, with an estimated value of 746$ billion dollars in 2015. Much of this growth has come from the appearance of low-cost carriers, which is now more than 25% of the worldwide market. The importance of this market and the variation that the Low-Cost Aviation has brought to the industry makes the analysis of the air traffic transportation industry an interesting subject of analysis and the transportation networks related to this industry become a compelling type of network to study.

The commercial aviation sector is complex and has several players in its value chain: airports, airplane manufactures, travel agents and others. To simplify the analysis, we decided to focus our interest on the airports, the passengers going through them and the amount of traffic between airports. We are using airport usage and airport flow (to/from) data to determine which airports, cities and countries are the major hubs of air travel, how the airports, cities and countries relate to each other in terms of air travel (where are the largest movements of people and between which places) and determine if we can extract communities of airports (clusters) that will allow us to discover unseen characteristics in the air travel industry like smaller airports that have high importance because of its utilization by low cost companies, or major hubs that are extremely important to the network and whose breakdown would make the entire network suffer.

Data Description

The data for this project was taken from the Open Flights Data Organization in https://openflights.org/data.html and Air Transport World in http://m.atwonline.com/airports/world-airport-traffic-december-2016. The dataset is a collection of airport flow from airport to airport with 10668 airports and 37338 edges between them. Other characteristics were added to the dataset manually to have a more complete dataset to analyze. These include for the vertices:

- Airport name and code (From/to)
- City (From/to)
- Country (From/to)
- Average Number of passengers of each airport per month

And for the Edges:

- Edge Weight based on amount of flights between destinations

The size of the network was rather significant and the data was gathered and put together using R to create an IGRAPH object that could be analyzed. This required an exploration, cleaning and filtering process that is explained below.

1. Data Exploration and Cleaning

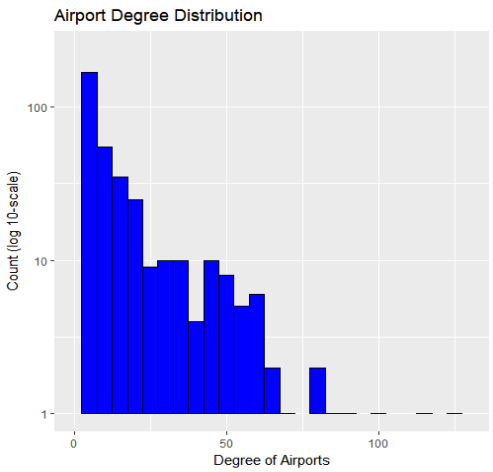
The data cleaning process involved the following steps:

- Dealing with NA values.
- Filtering Vertices with very low weights in the network (Delete Weight <=3)
- Filtering Remaining singleton vertices (Delete degree=0)

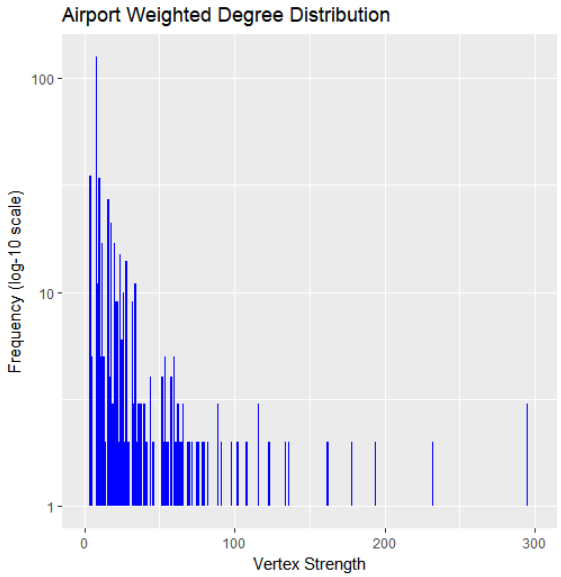
To do this, all rows with NA values were removed from the data set before the creation of the graph to be analyzed. Filters were then applied to reduce the size of the network to a more manageable size.

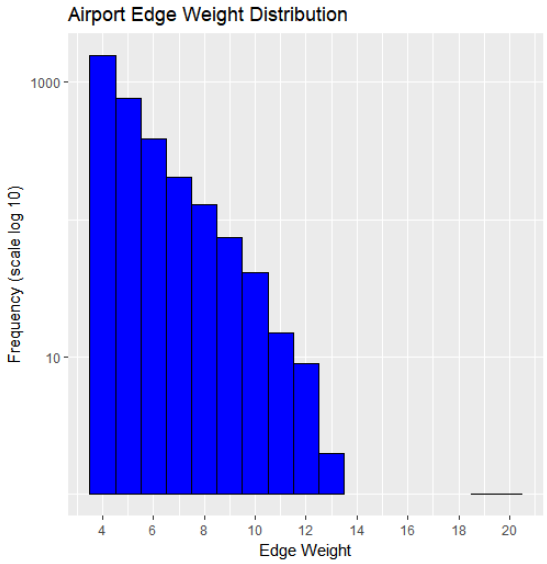
After his initial processing, we obtained a network which contained 594 nodes and 3192 vertices. Airport degree varied the following way:

Airport Degree - 5 Number	
min	7
1st quarter	2
median	4
mean	10.75
3rd quarter	11
max	216

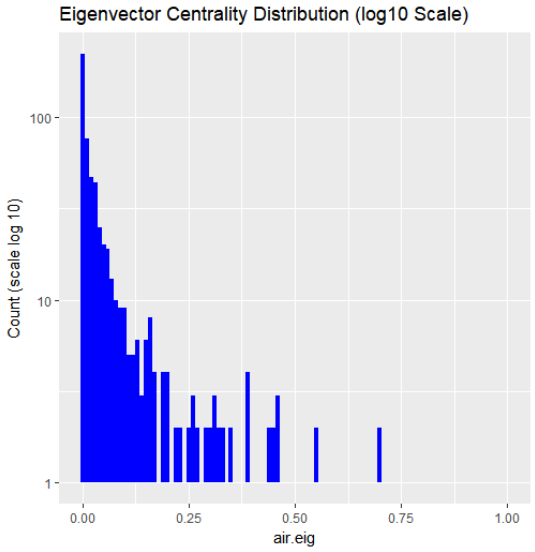


The degree variation indicates that there are few airports with high degrees that act as big hubs for the network and a very large number of airports that that act locally.



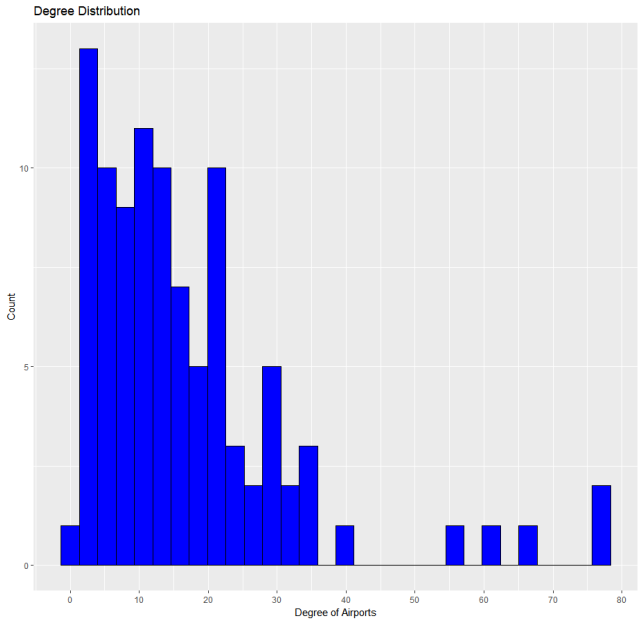


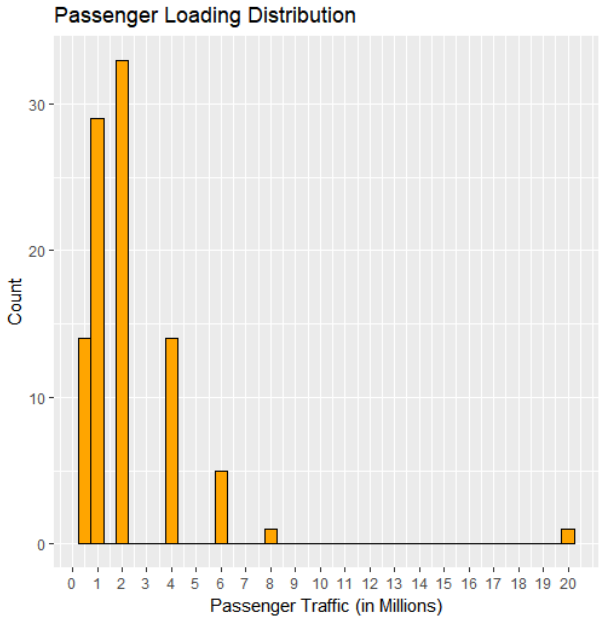
If we assign weights to the edges in the distribution, we can see that the great majority of airports has a weight that is less than 75. So, we have many airports that act as local transit hubs in the network. The weighted distribution allows us to see that there is only a few of airports that have weights over 250 (which represent the airports like Atlanta or London) that act as major hub for the entire network. The Edge weight of nodes mostly fall under 10 which means most of airports have less than 10 airlines flying on the same route.



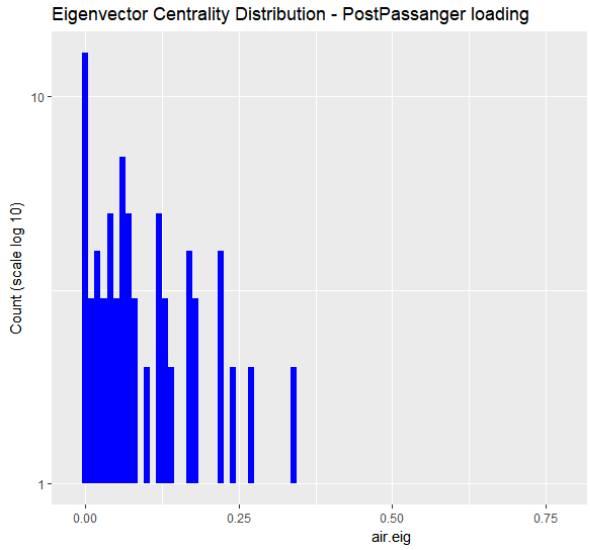
The eigen-centrality of most of the nodes in the network is below 0.5 indicating that most nodes do not have a large level of influence in the network. A small number of network have a level of influence near 0.75, which influence the network a lot. These nodes are probably large airports like Atlanta or London. This can be viewed more clearly in the overall network visualization below.

After the initial filtering, the passengers were added to the IGRAPH and the values were normalized since they were in the millions. By diving each value by 10^6. The network was cleaned for singletons again. The network after adding the passenger loadings per airport per month and adding had 97 nodes and 808 edges.





The degree distribution of the network does not change too much after adding the passenger loadings. Passenger loading distribution follows the same pattern as the degrees. Only a few airports have monthly passenger loadings over 20 million and most of them have loadings lower than 2 million passengers a month.



The eigen-centrality of most of the nodes in the network after adding the passenger loadings is below 0.5 indicating that most nodes do not have a large level of influence in the network. A small number of network have a level of influence near 0.5, which excerpt a greater influence.

Data Analysis

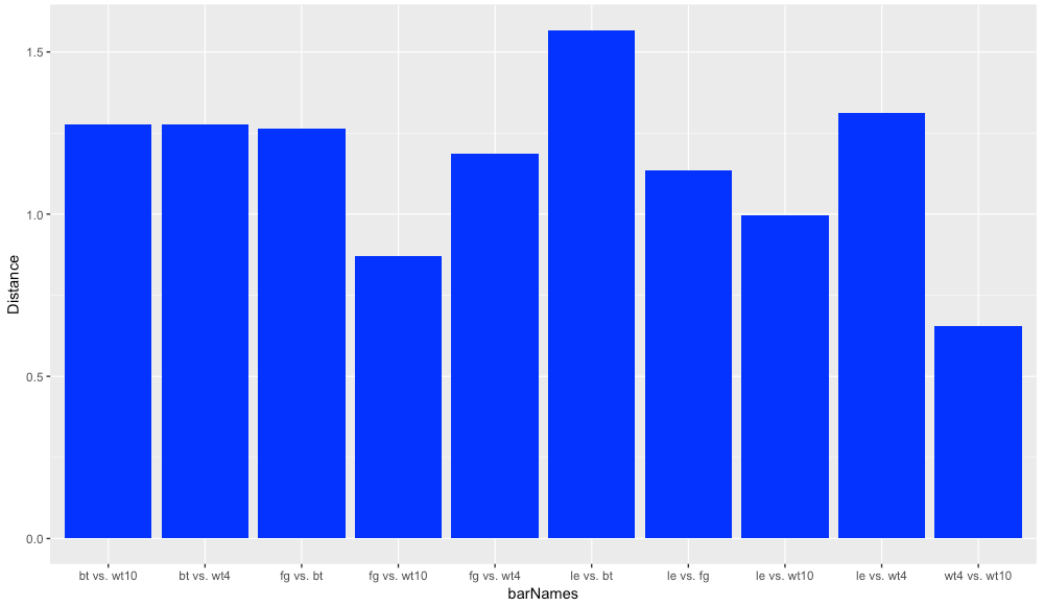
The analysis of the network was done both on the network with the passenger loadings and without the passenger loadings included since the addition of passenger loadings and subsequent filtering of nodes left a much smaller network that was not as adequate for certain types of investigation such as community detection. The analysis focused on identifying and comparing communities of airports that are mostly geographical in nature. The different communities represent regions of the world with the major air transport hubs that allow for interconnectivity between communities and connect to smaller airports. Because of these characteristics the node betweenness was a point of attention in the analysis of this network.

2. Community detection analysis

We did community detection analysis on the network before passenger loadings and obtained the following results using R and GEPHI.

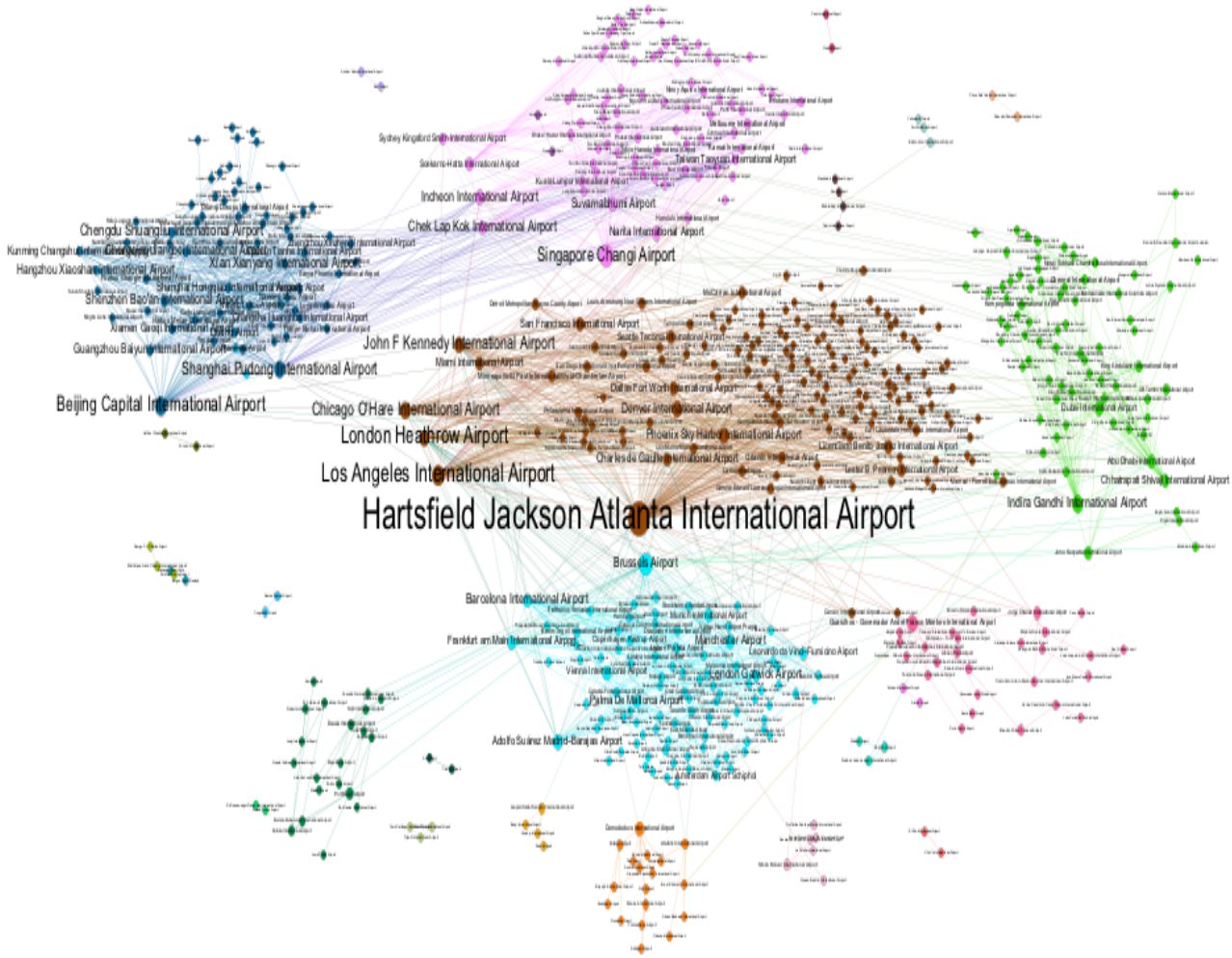
	Number of Cluster	Modularity
Gephi Modularity Algorithm	27	0.670
Lead Eigenvector	39	0.628
Fastgreedy	28	0.652
Edge Betweenness	34	0.613
Walktrap (4 Steps)	54	0.649
Walktrap (10 Steps)	41	0.653

Comparison of Distance Between Community Detection Algorithems



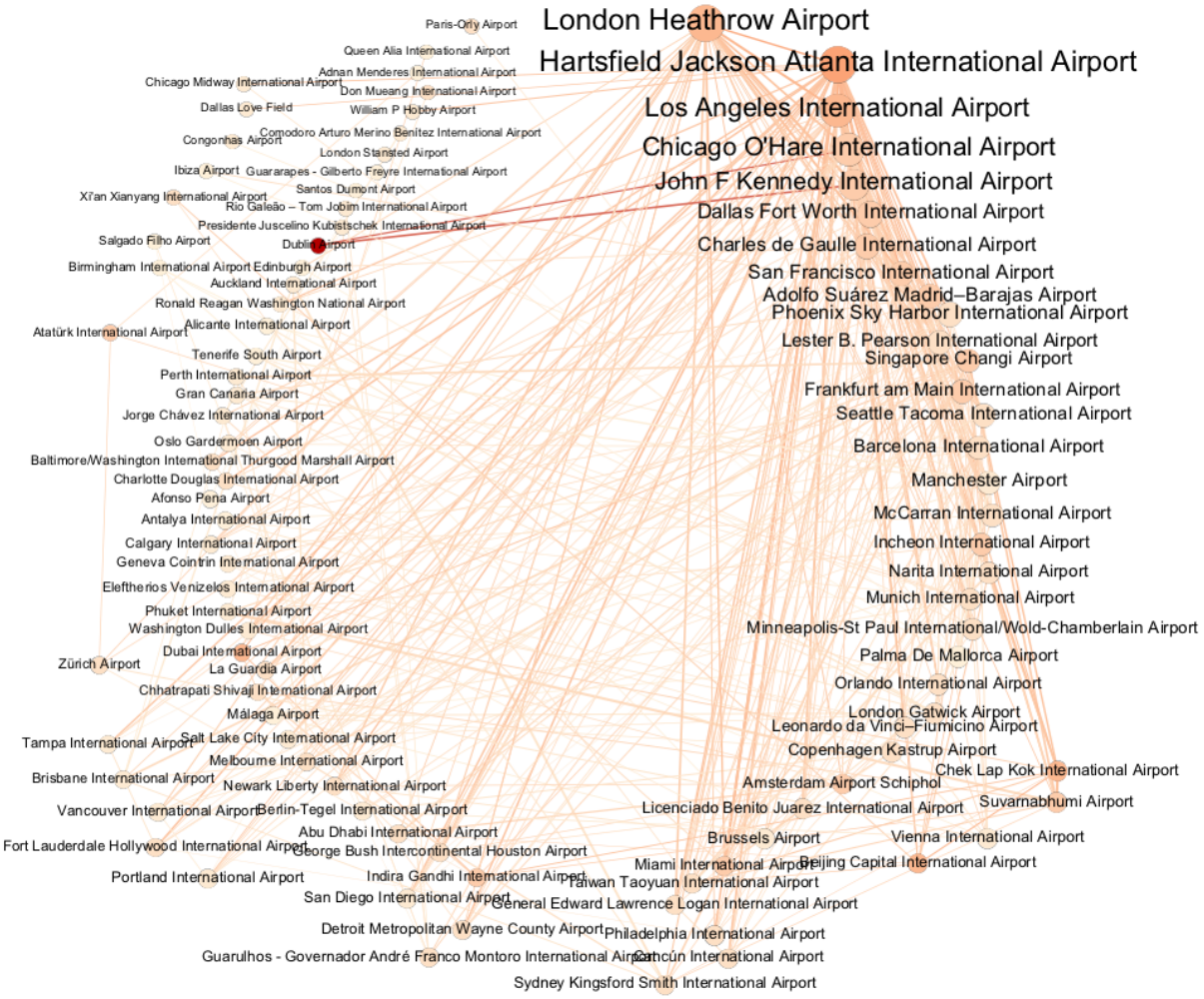
We obtained the modularity generated by Gephi algorithm. This process generated 27 communities from the network. Nodes and text size are by degree, color is by communities, the communities are ranging from 2 to 160. And now we would use Gephi modularity detection result for further analysis.

The Modularity Detection of Entire Network by Gephi



The display of the entire network above shows a full view of the network after the data has been filtered. All the nodes in the network represent airports and the size of each node represents its degree (the amount of connection each airport has). The edges in the network vary its size per the weight of the connection, which represents the amount of traffic there is between 2 airports. The color of the nodes represents the modularity of the nodes. There are 27 communities (i.e. 27 colors) based on modularity minimization. These regions mostly form geographical parts of the planet.

Using the network that contained passenger loadings, we did a special focus on the high traffic airports where the node size, text size was plotted by degree, and the colors represented by passenger loading using a circular layout to represent the clear link between passenger monthly load in airports and degree (see visualization below).



From the economical point of view, the more traffic an airport has, the higher degree values (More connection flight to different airport) it has as well. By observing the previous circular plot, we can see this relationship (degree vs loading) taking the following airports as examples:

- London Heathrow Airport
- Hartsfield Jackson Atlanta Intl. Airport
- Chicago O'Hare Intl Airport

However, some of the airports do NOT have this high degree have high passenger loading. This kind of circumstance is probably created by geographical reasons (position of the node in the network).

For example:

- Chek Lap Kok Intl Airport
- Dubai Intl Airport
- Dublin Airport

- Singapore Changi Airport

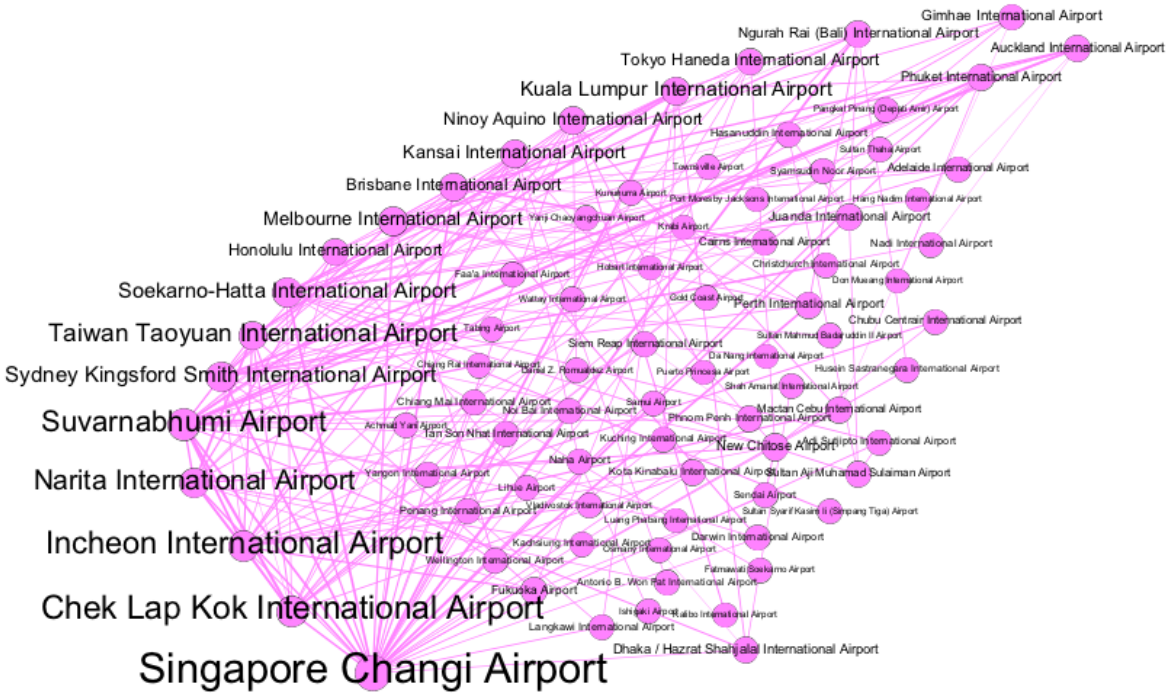
Based on the above characteristics, we chose 2 communities to process for further analysis. The 2 communities are chosen based on their high degrees and passenger traffic stats. These communities are detailed below.

Community1: Mostly US airport region

In this visualization, node colors indicate the modularity (Western airports) and the nodes /text size is scaled by degree centrality.

The community for most US airports was obtained from the network before adding the passenger loadings to have a larger network to find communities in. The plot below depicts this community of airports where node size and color are defined by degree. Based on this plot we can see that those high degree nodes play the role of hubs for the other airport in the same region. The above top high traffic airports (London, Atlanta, Chicago) are basically heavily connected to each other. They also act as key hubs to connect other airports the close-in the region. Most these airports are based in the North America and Europe region. This also implies that this heavy connection might be due to close business relationships and cultural proximity.

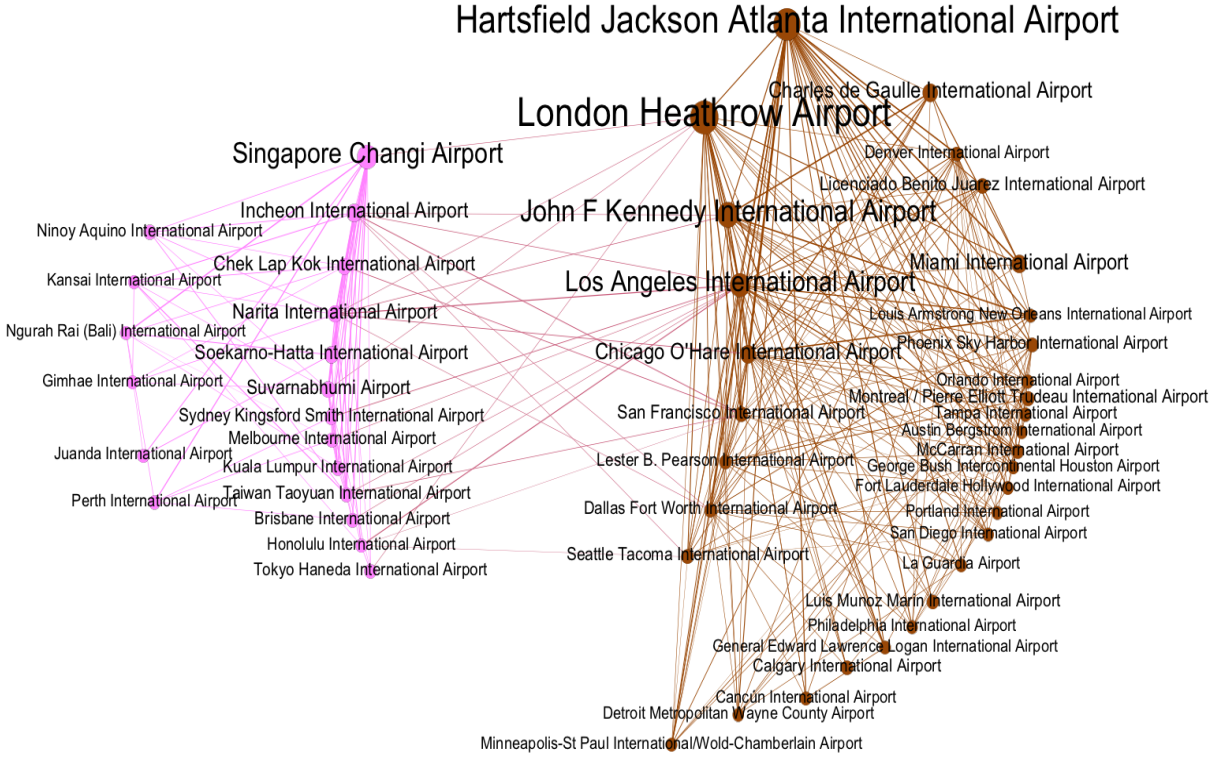
Community2- Asia Region Airport - High Traffic



In this visualization, node colors indicate the modularity (Eastern airports) and the nodes /text size is scaled by degree centrality.

The community for most Asia region airports was obtained from the network before adding the passenger loadings to have a larger network to find communities in. The plot below depicts this community of airports where node size and color are defined by degree. We can see a similar behavior in this cluster compared to what we saw in the US network where large airports (like Singapore or Suvarnambhumi airports) act as major hubs. However, the Mostly US network has a clear central point with very high degree that acts as major hub for the entire network. The distribution of airport traffic in the Asian cluster is more divided among several different airports in different sub-regions (japan, Australia, Singapore, India) that then act as jobs for smaller airports in the network.

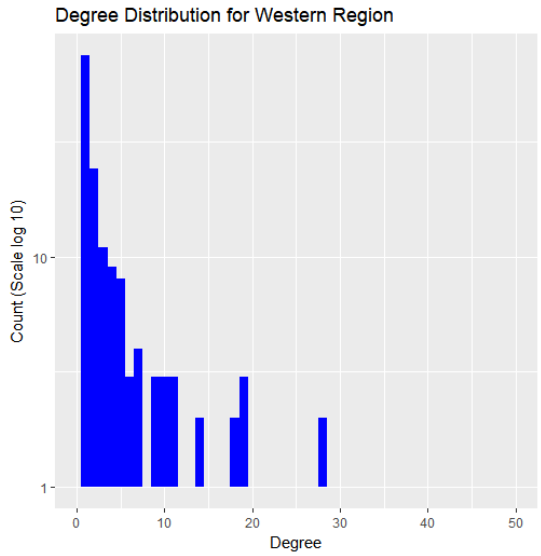
Inter-community connectivity analysis

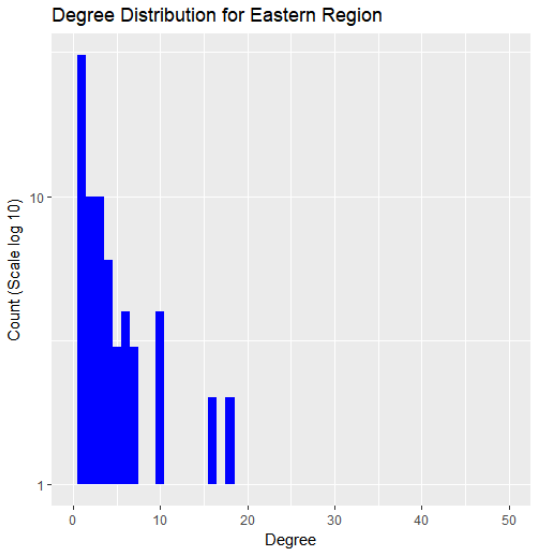


In this visualization, node colors indicate the modularity and the nodes /text size is scaled by betweenness centrality.

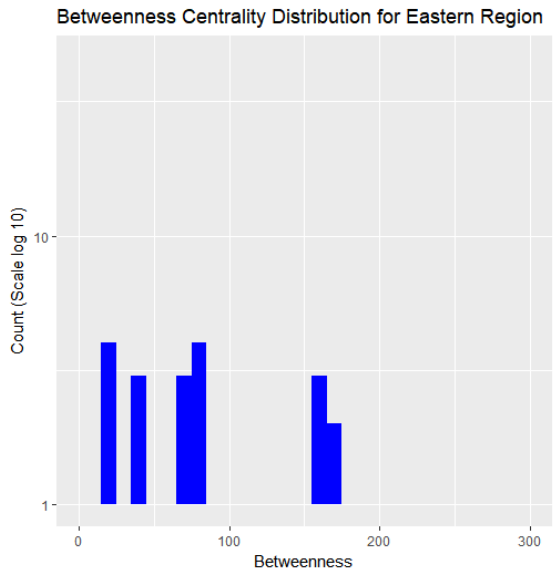
There is an interesting connection that can be analyzed between high traffic airports in different communities (based on geography). Major hubs in different communities and geographical regions connect to each other. We can see an interesting connection between the Los Angeles airport and most of the Asian high traffic airports that is based on geographical closeness. The Atlanta airport despite is importance in the Mostly US network is not as heavily connected as the LA airport to the Asia region community. This indicates that the LA airport might be of great importance in terms of betweenness in the network which will be analyzed in the following section of this report.

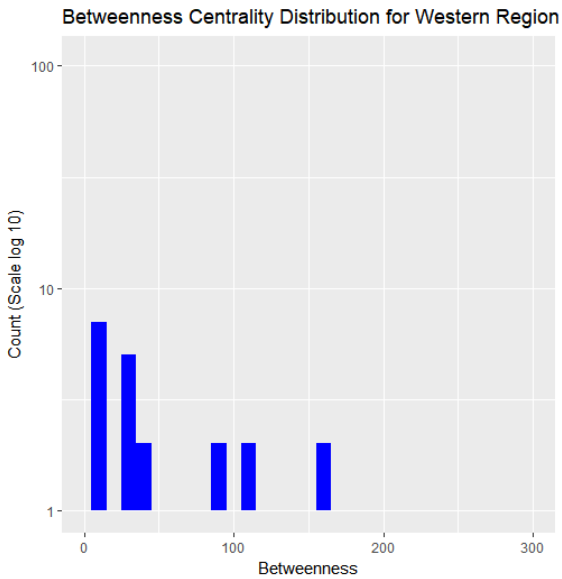
The degree distribution of the two communities shows more high degree nodes in the western (Mostly US) cluster. However, both communities share the fact that both geographical sections of the world have most nodes with a degree lower than 10. Most airports act as local transportation sharing degree values between 15 and 20. These hubs allow for the flow of a great number of passengers in the entire network.





The high betweenness nodes are visible in the betweenness-centrality distribution graphs where only a few high betweenness nodes are present in each geographical region. It is interesting to note that even though the western community has more nodes with high degree, there are fewer nodes with high betweenness than in the Eastern region. We can conclude from this that the Asia region is more decentralized in terms of air traffic with a higher proportion of its nodes acting as connects to other sections of the entire network.





The pattern of degree and betweenness are very close. This effect could give us some hints how the network forms. The high degree node (airport) play the key role of connectors connects to other airports regardless of the region. Most of the airports has very low (close to zero) betweenness values, and only several of high degree airports have high betweenness values.

By comparing the western and Eastern chosen communities, we found that the pattern of their degree and betweenness are very similar. The high degree nodes (airports) act as the bridges between airports, so that those airports having high betweenness values can help communicate the two different regions. Other airports have lesser degrees and betweenness because they depend on those connecting airports with high betweenness.

3. CUG tests and Assortativity

CUG testing indicates that all the networks display assortativity patterns that are significant with respect to a set of 1000 random networks generated. The transitivity of the networks was also tested and turned out to be significant with respect to random networks. Across all four networks, the geographical region does affect the likelihood that an airport's location with similar characteristics has a greater chance to have a route connection.

The table below sums out the results for assortativity.

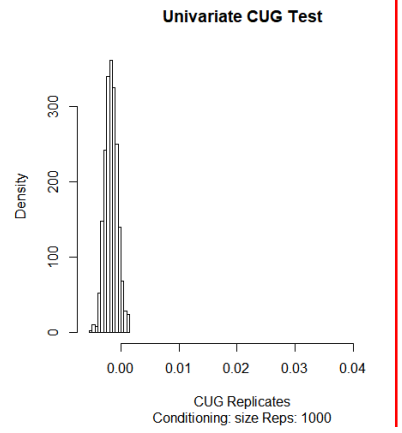
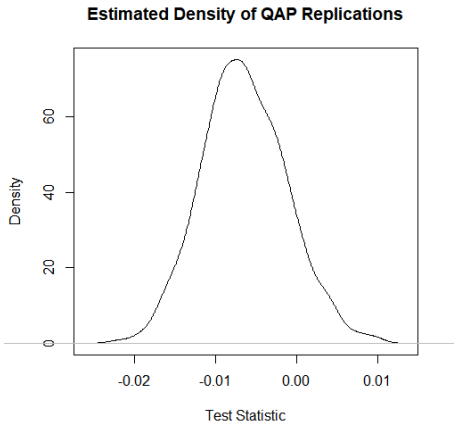
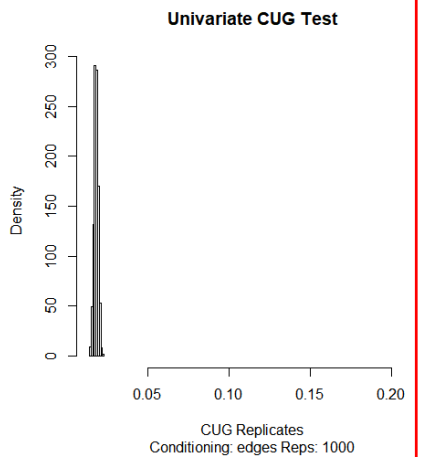
	CUG TEST		
	Assortativity	Pr(X>=Obs)	Pr(<=Obs)
Entire network (before passenger)	0.457	0	1
Entire network (after passenger)	0.18	0	1
US community	-0.033	1	0
Asia Community	-0.37	1	0

It was interesting to note that the assortativity was not significant once we dealt with the smaller communities for the Western and Eastern geographical areas. The tendency of node to attach to other nodes of the same type is less significant once we are dealing with smaller clusters.

	CUG TEST		
	Transitivity	Pr(X>=Obs)	Pr(<=Obs)
Entire network (before passenger)	0.21	0	1
Entire network (after passenger)	0.33	0	1
US community	0.17	0	1
Asia Community	0.25	0	1

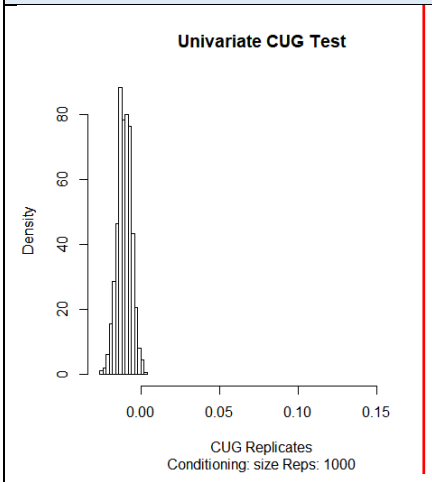
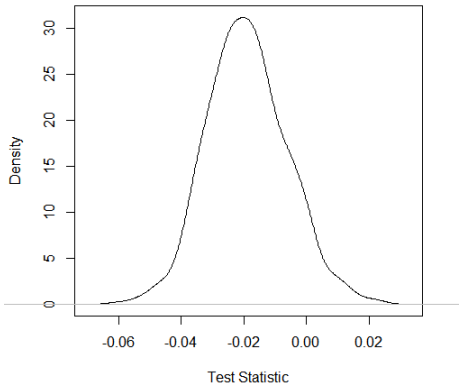
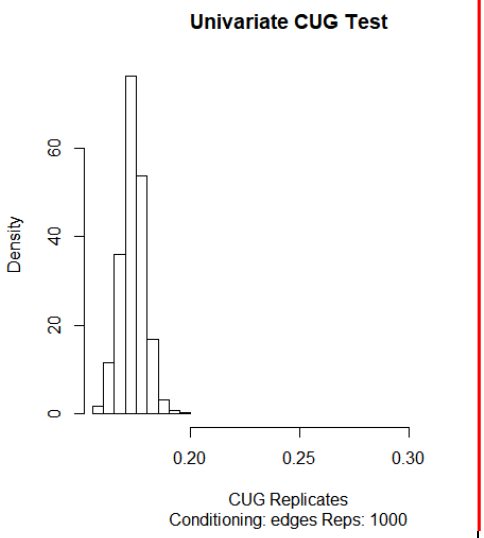
a. Entire Network Before Passenger Filtering

The airport CUG and QAP test for assortativity are displayed below for the entire networks before filtering, showing the significance of the assortativity and transitivity for the entire network.

Airport CUG test for assortativity	Airport QAP test for assortativity	Airport CUG test for Transitivity
		

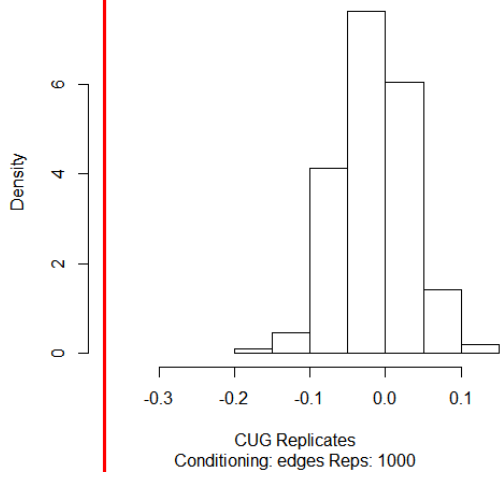
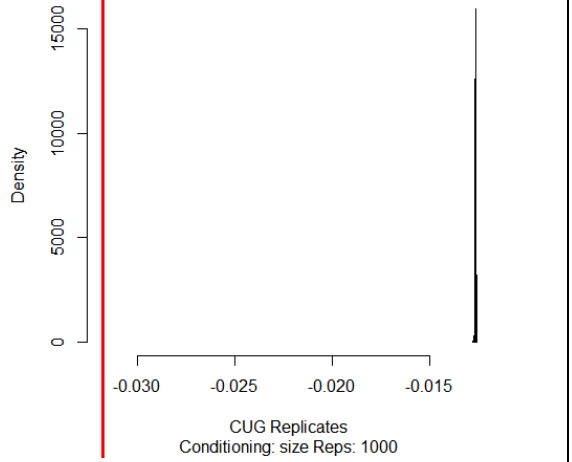
b. Entire Network After Passenger filtering

The airport CUG and QAP test for assortativity are displayed below for the entire networks before filtering, showing the significance of the assortativity and transitivity for the entire network after filtering for passenger loadings.

Airport CUG test for assortativity	Airport QAP test for assortativity	Airport CUG test for Transitivity
		

c. Western and Eastern Community

The assortativity is not significant when dealing with the communities detected.

Airport CUG test for assortativity	Airport QAP test for assortativity
	

Conclusion

Air transport plays a very significant role in our world. The commercial sector can be influenced seriously by the aviation industry. Consequently, airport and route design act as a key player on this matter. Based on this line of though, we analyzed how airport networks are being built. There are various route network characteristics associated with different regions or the network that we were able to visualize and analyze in detail.

The entire network shows that major airport hubs appear on large communities. One key characteristic is that those hubs possess very high degree and heavy edge weight. This implies that those airports play a very important role connecting different airports within their regions. Also, this also implies that those low degree airports are very depends on those hubs in order to have connection to other airports.

By observing the analysis findings, we could probably have an idea of which airports have the busiest traffic. However, we are also able to visualize the picture behind this more obvious finding. Busy logistics makes deliveries to this airports' region or through them to other regions.

The more traffic the place has, the more business/trade they have going through it. Therefore, we could conclude that those high-performance airports' city mostly act as a financial or trading centers.

The two communities (Western & Eastern) gave us a closer look on how those major hubs are connecting with other airports within the network. Those top hubs are not just responsible for bonding with the airports within the region, but also plays the key role connecting other airports from other communities. This finding is easy to predict and foresee if we also consider the business-related reasons for massive air traffic through a node. US or Europe are very large economic systems; Eastern countries mostly play a key role of supplying / manufacturing. These airports are connected and pull these communities closer. This has a huge effect on trade and the economy, but more research is needed to be sure of this point.

In the CUG test, the result shows that airport network is not randomly built. On the contrary, the connection is governed by the Geographic factors (e.g. Business establishment and culture etc). This property would be a very good guide for corporations to forecast or expand their business geography maps. With this relation, we could have better understanding on how the passenger traffic flow behaves and where the flow is coming from. Then the forecasting of what type of business could be implement based on this information would become possible. Nevertheless, more research must be performed to discover more evidence on this matter.