Web Analysis

Group Project

Liang Jiaqi G1901813B

Wang Hongyu G1902385H

Xu Zifan G1901842K

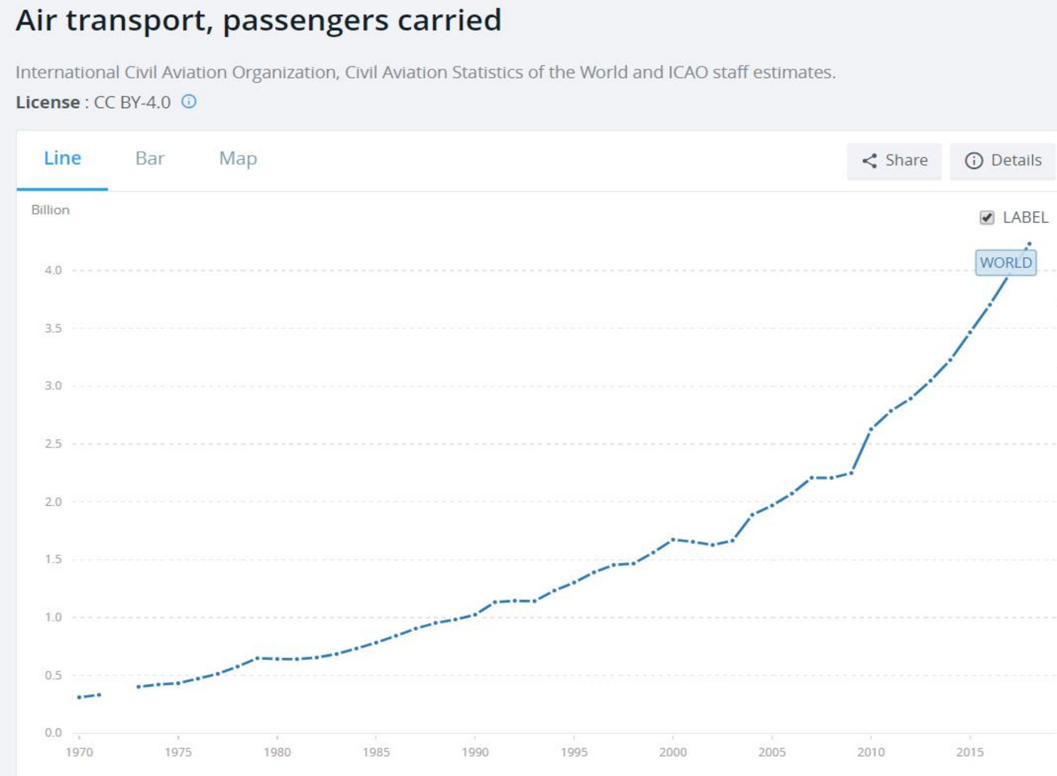
Xiao Hanhua G1901844D

Content

1. Introduction

1.1 Background

We can see in the figure below that there is an increasing trend if passengers carries by air transport. That is to see, air transport takes a important role in today's transportation. So it is necessary to analyze the community and key players of flights data.



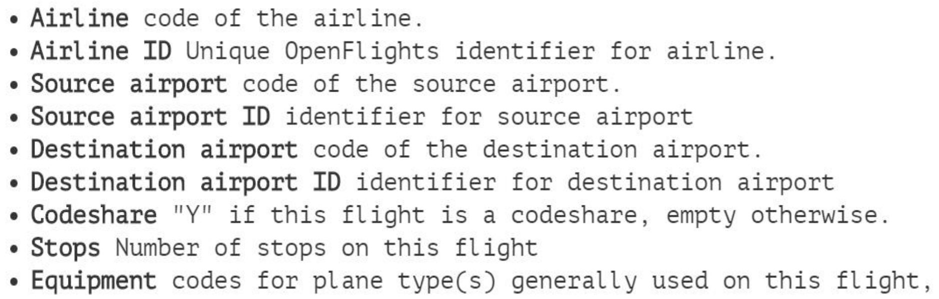
Air transport, passengers carried

International Civil Aviation Organization, Civil Aviation Statistics of the World and ICAO staff estimates.

License : CC BY-4.0

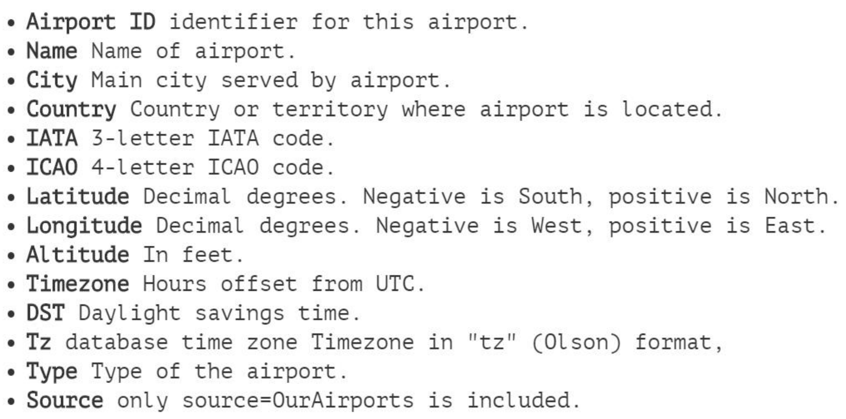
1.2 Dataset Description

We have two tables in the dataset.
The first is route.dat
This Airline Route Database contains 67,663 routes between 3,321 airports on 548 airlines spanning the globe.



- **Airline** code of the airline.
- **Airline ID** Unique OpenFlights identifier for airline.
- **Source airport** code of the source airport.
- **Source airport ID** identifier for source airport
- **Destination airport** code of the destination airport.
- **Destination airport ID** identifier for destination airport
- **Codeshare** "Y" if this flight is a codeshare, empty otherwise.
- **Stops** Number of stops on this flight
- **Equipment** codes for plane type(s) generally used on this flight,

The second is airports.dat
The Airports Database contains over 10,000 airports, train stations and ferry terminals spanning the globe.

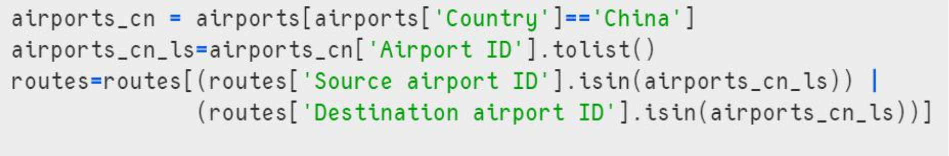


- **Airport ID** identifier for this airport.
- **Name** Name of airport.
- **City** Main city served by airport.
- **Country** Country or territory where airport is located.
- **IATA** 3-letter IATA code.
- **ICAO** 4-letter ICAO code.
- **Latitude** Decimal degrees. Negative is South, positive is North.
- **Longitude** Decimal degrees. Negative is West, positive is East.
- **Altitude** In feet.
- **Timezone** Hours offset from UTC.
- **DST** Daylight savings time.
- **Tz** database time zone Timezone in "tz" (Olson) format,
- **Type** Type of the airport.
- **Source** only source=OurAirports is included.

1.3 Data Preprocessing

1.3.1 Find airports in China and routes from or to China

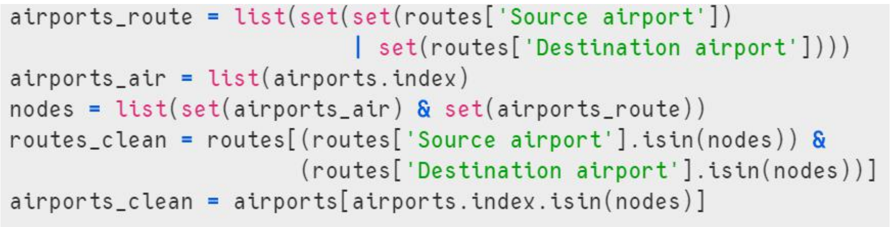
First choose the airports in China to make a list.
Then find routes from or to China.



```
airports_cn = airports[airports['Country']=='China']
airports_cn_ls=airports_cn['Airport ID'].tolist()
routes=routes[(routes['Source airport ID'].isin(airports_cn_ls)) |
              (routes['Destination airport ID'].isin(airports_cn_ls))]
```

1.3.2 Find common airport between both datasets

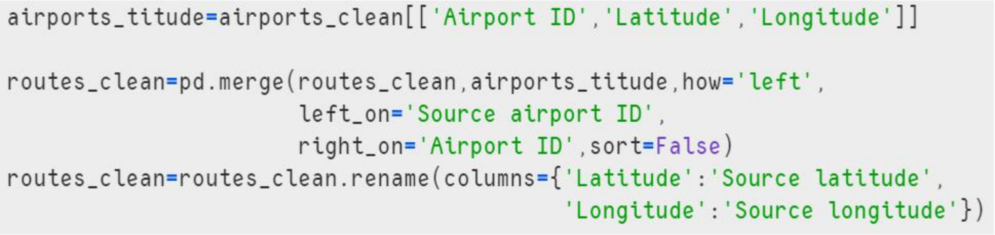
The airports of these two datasets are not just the same. To better analyze the data, we find common airports between both datasets.



```
airports_route = list(set(set(routes['Source airport'])
                      | set(routes['Destination airport'])))
airports_air = list(airports.index)
nodes = list(set(airports_air) & set(airports_route))
routes_clean = routes[(routes['Source airport'].isin(nodes)) &
                      (routes['Destination airport'].isin(nodes))]
airports_clean = airports[airports.index.isin(nodes)]
```

1.3.3 Altitude and longitude of airports.

For doing the visualization and the calculation of distance, we add new columns to describe the altitude and longitude of the airports. We use merge function to do this.

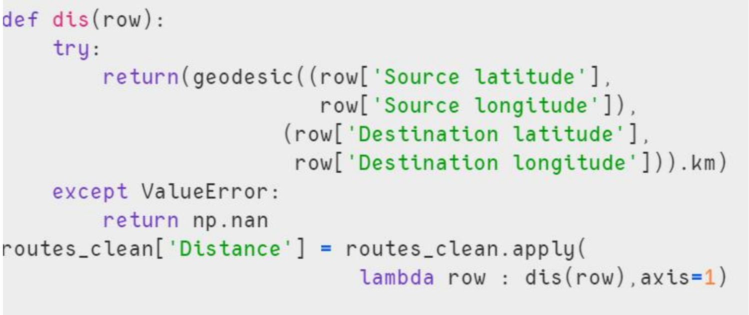


```
airports_titude=airports_clean[['Airport ID','Latitude','Longitude']]

routes_clean=pd.merge(routes_clean,airports_titude,how='left',
                      left_on='Source airport ID',
                      right_on='Airport ID',sort=False)
routes_clean=routes_clean.rename(columns={'Latitude':'Source latitude',
                                         'Longitude':'Source longitude'})
```

1.3.4 Distance of the route.

To better analyze the flights data, we calculate the distance of each route. We use geodesic function to return the distance value, which is imported by geopy library.



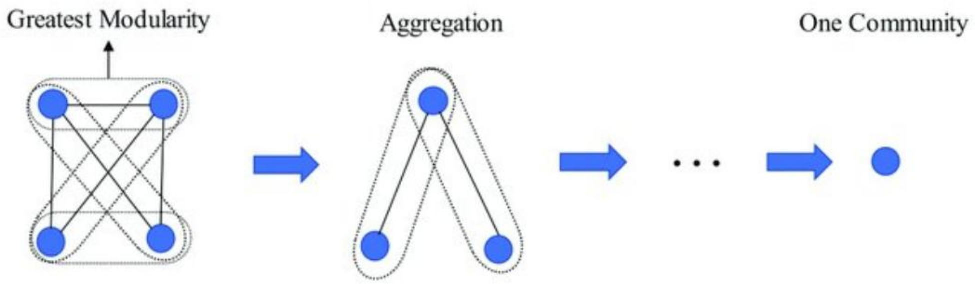
```
def dis(row):
    try:
        return(geodesic((row['Source latitude'],
                         row['Source longitude']),
                      (row['Destination latitude'],
                       row['Destination longitude'])).km)
    except ValueError:
        return np.nan
routes_clean['Distance'] = routes_clean.apply(
                            lambda row : dis(row),axis=1)
```

1.4 Community Detection Algorithms

1.4.1 Greedy Modularity Maximization

In this very simple approach we start out with each vertex in our network in a one-vertex group of its own, and then successively amalgamate groups in pairs, choosing at each step the pair whose amalgamation gives the biggest increase in modularity, or the smallest decrease if no choice gives an increase. Eventually all vertices are amalgamated into a single large community and the algorithm ends. Then we go back over the states through which the network passed during the course of the algorithm and select the one with the highest value of the modularity. This method is implemented in igraph in the function fast greedy.community.

A naive implementation of this idea runs in time O(n^2), but by making use of suitable data structures the run time can be improved to O(n \log^2 n) on a sparse graph. Overall the algorithm works only moderately well: it gives reasonable divisions of networks, but the modularity values achieved are in general somewhat lower than those found by the other methods described here. On the other hand, the running time of the method may be the best of any current algorithm, and this is one of the few algorithms fast enough to work on the very largest networks now being explored.

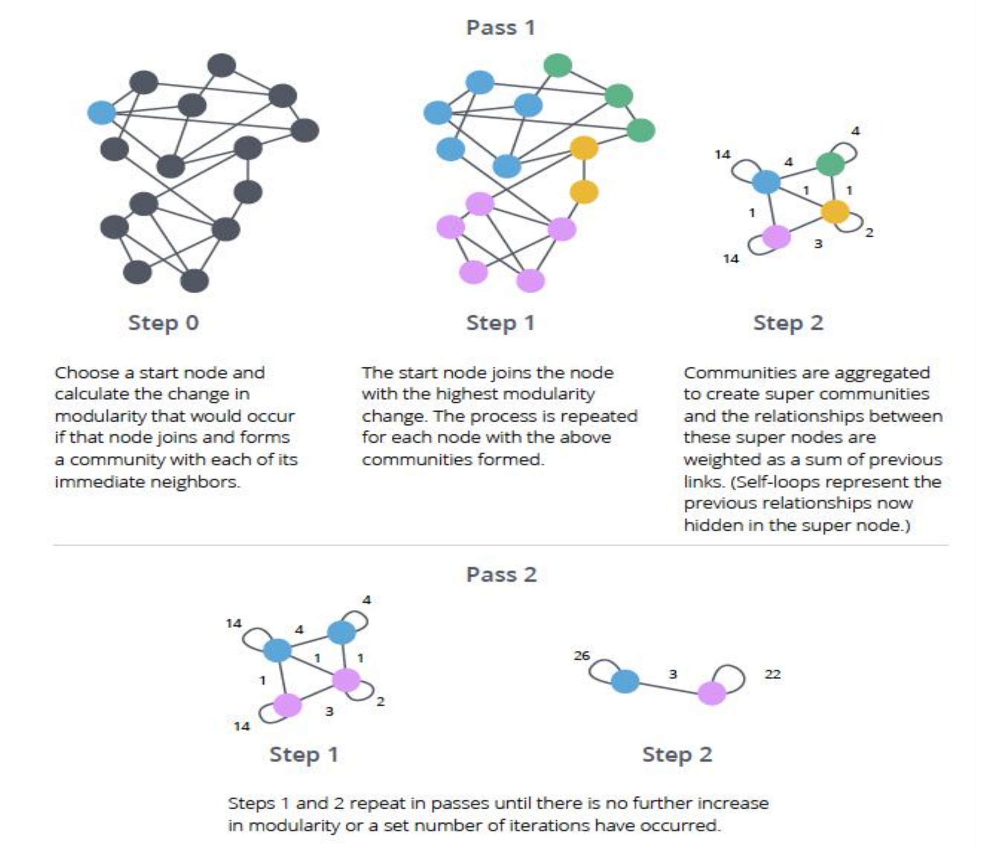


1.4.2 Louvain Modularity Algorithm

The Louvain method of community detection is an algorithm for detecting communities in networks. It maximizes a modularity score for each community, where the modularity quantifies the quality of an assignment of nodes to communities by evaluating how much more densely connected the nodes within a community are, compared to how connected they would be in a random network.

The Louvain algorithm is one of the fastest modularity-based algorithms and works well with large graphs. It also reveals a hierarchy of communities at different scales, which is useful for understanding the global functioning of a network.

The Louvain algorithm was proposed in 2008. The method consists of repeated application of two steps. The first step is a “greedy” assignment of nodes to communities, favoring local optimizations of modularity. The second step is the definition of a new coarse-grained network based on the communities found in the first step. These two steps are repeated until no further modularity-increasing reassignments of communities are possible.



Choose a start node and calculate the change in modularity that would occur if that node joins and forms a community with each of its immediate neighbors.

The start node joins the node with the highest modularity change. The process is repeated for each node with the above communities formed.

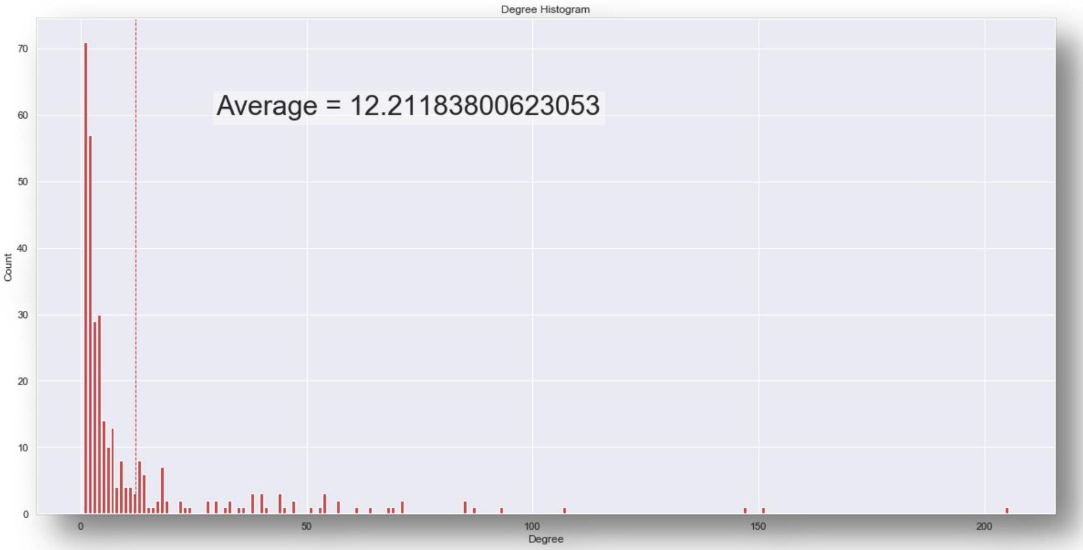
Communities are aggregated to create super communities and the relationships between these super nodes are weighted as a sum of previous links. (Self-loops represent the previous relationships now hidden in the super node.)

Steps 1 and 2 repeat in passes until there is no further increase in modularity or a set number of iterations have occurred.

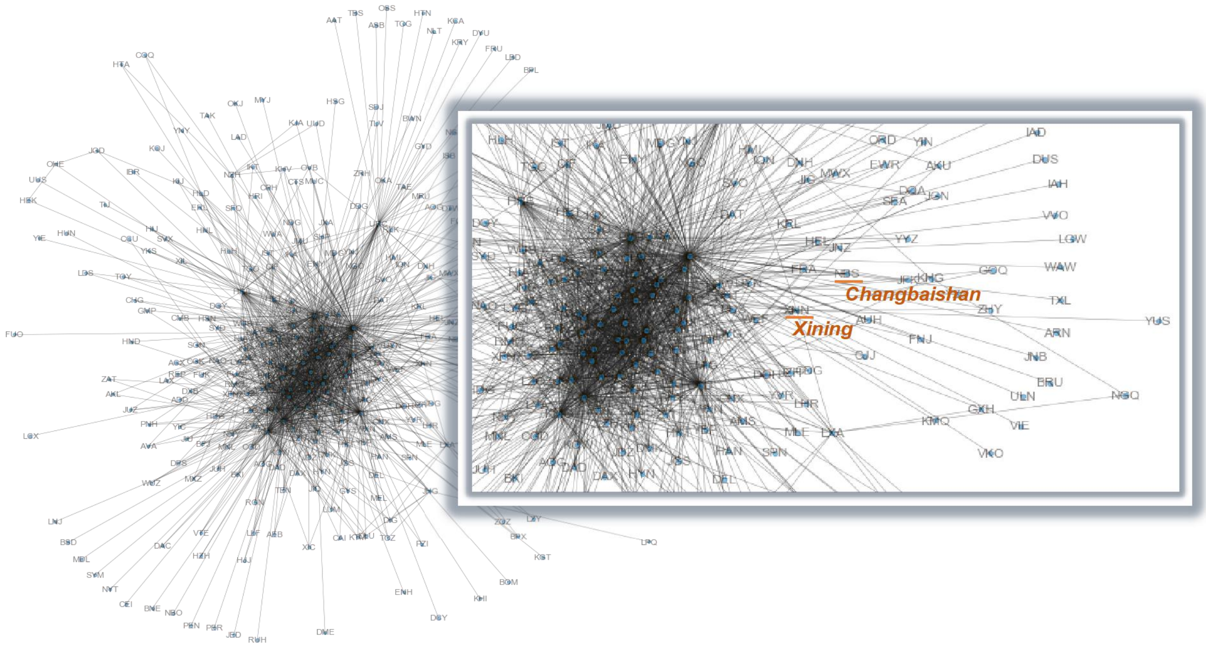
2. China Overview

2.1 Network measures

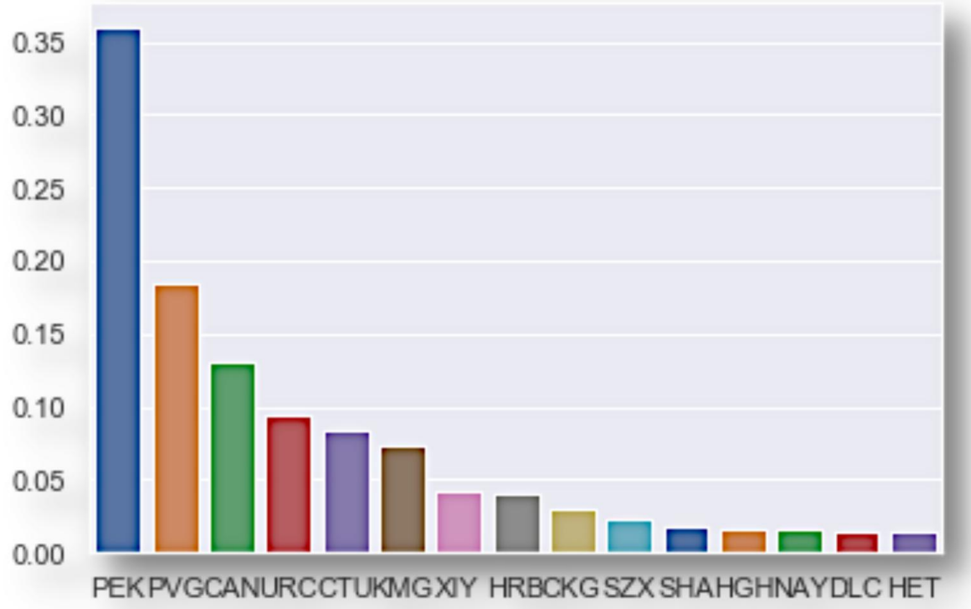
In this part, we will do data exploration to have an overview of network in China. Firstly, we made the graph of degree distribution. The dotted line represents the average degree of network. What's more, the distribution basically follows the power law.



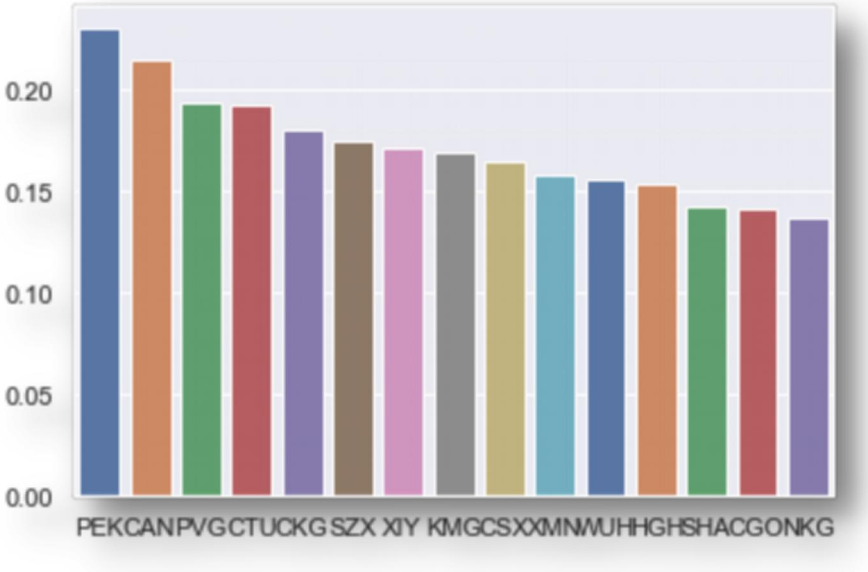
Secondly, the picture below shows the graph connectivity. The 3-letter names are abbreviations of airports in the word. The density of graph represents how busy an airport is. So for some airports in the first-tier city in China, they will locate in the center of graph, which means that they have high connectivity with other airports than the points in edge.



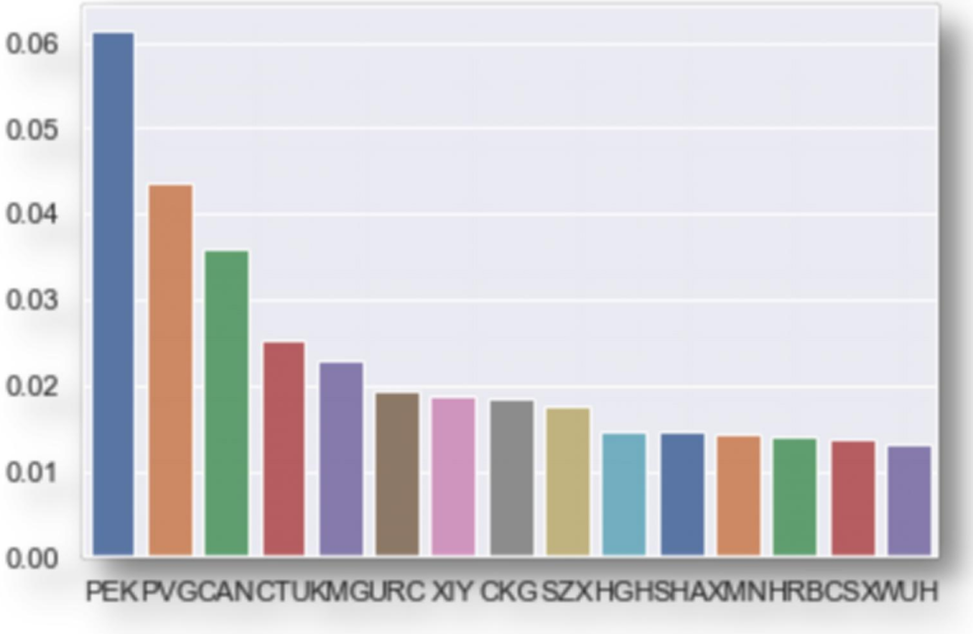
We can see that our graph is not connected. In fact, there are 1 component in total. We get the largest connected component for the continuation. It contains 321 nodes. The diameter of the graph is the length of the longest shortest path between any pair of nodes.
The diameter of our graph is 5. And we also obtain the average clustering coefficient. It’s about 0.61912.
Thirdly, we did some research on centrality. Centrality is a measure of importance of nodes/edges in a network. Betweenness centrality measures how much nodes/edges are part of shortest paths. In this graph, we built the bar plot to show the top 15 airports in means of betweenness.



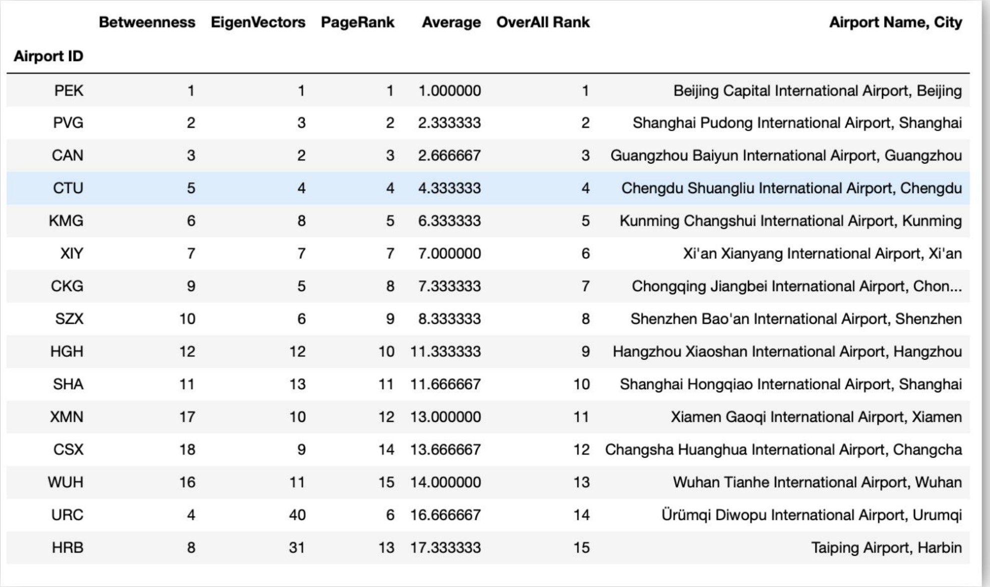
Next, the graph below is about eigenvalue centrality. Eigenvalue centrality measures how much nodes are connected to nodes with high betweenness centrality. This bar plot shows the top 15 airports in means of eigenvalue.



What’s more, we also included the pagerank algorithm. PageRank is a variant of eigenvalue centrality. So the graph below shows the top 15 airports in means of pagerank.



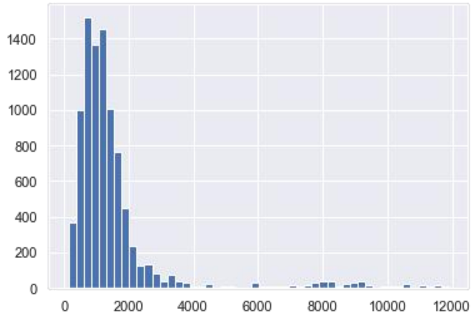
To sum up, we made a comparison. In this form, we can see that the PageRank is closer to the overall rank, which according to the average of these 3 ranks. The top five airports are in Beijing, Shanghai, Guangzhou, Chengdu and Kunming.

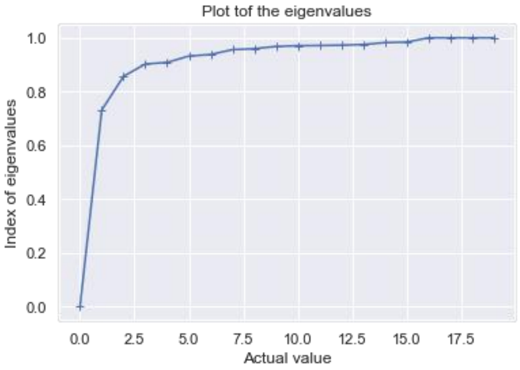


	Betweenness	EigenVectors	PageRank	Average	OverAll Rank	Airport Name, City
Airport ID						
PEK	1	1	1	1.000000	1	Beijing Capital International Airport, Beijing
PVG	2	3	2	2.333333	2	Shanghai Pudong International Airport, Shanghai
CAN	3	2	3	2.666667	3	Guangzhou Baiyun International Airport, Guangzhou
CTU	5	4	4	4.333333	4	Chengdu Shuangliu International Airport, Chengdu
KMG	6	8	5	6.333333	5	Kunming Changshui International Airport, Kunming
XIY	7	7	7	7.000000	6	Xi'an Xianyang International Airport, Xi'an
CKG	9	5	8	7.333333	7	Chongqing Jiangbei International Airport, Chon...
SZX	10	6	9	8.333333	8	Shenzhen Bao'an International Airport, Shenzhen
HGH	12	12	10	11.333333	9	Hangzhou Xiaoshan International Airport, Hangzhou
SHA	11	13	11	11.666667	10	Shanghai Hongqiao International Airport, Shanghai
XMN	17	10	12	13.000000	11	Xiamen Gaoqi International Airport, Xiamen
CSX	18	9	14	13.666667	12	Changsha Huanghua International Airport, Changcha
WUH	16	11	15	14.000000	13	Wuhan Tianhe International Airport, Wuhan
URC	4	40	6	16.666667	14	Ürümqi Diwopu International Airport, Urumqi
HRB	8	31	13	17.333333	15	Taiping Airport, Harbin

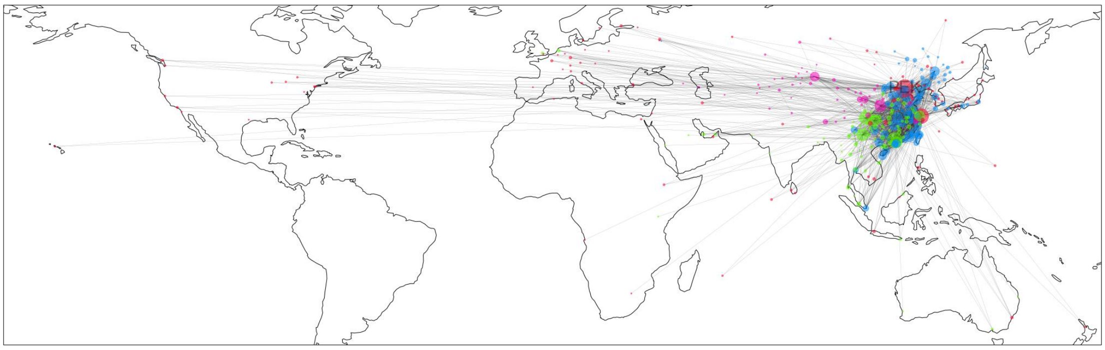
2.2 Applied algorithms

Finally, looking around the whole world range, we made the graph of flight distance distribution. The distribution of distances follows a power law with most fly being relatively short distances. When using spectral clustering as a mean to find communities, the plot below can help to find the number of communities. we can see such a gap only between the first two eigenvalues, suggesting that indeed there are only 2 communities.

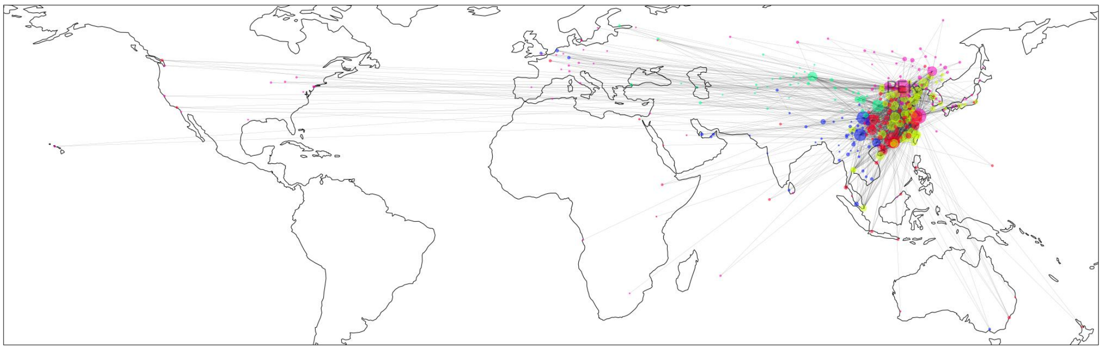




Then we applied the greedy modularity maximization to improve the result. It seems that by using the modularity, we can get better results. We see for example that the east and west of China are detected as different communities. This means that for each of these, there are way more connections inside them, compared to the number connections that leads outside of them.



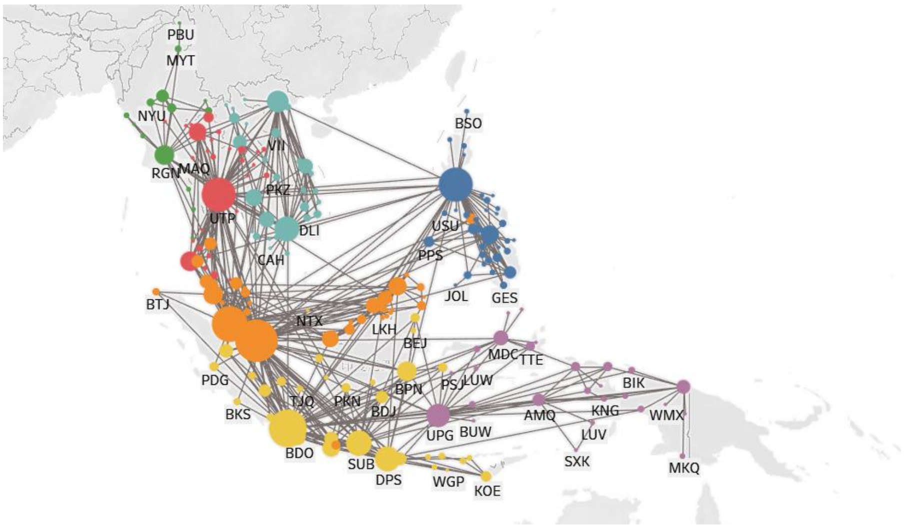
And we also tried the Louvain algorithm to get the maximum of modularity. From the output we can see the hierarchy of community more specific.

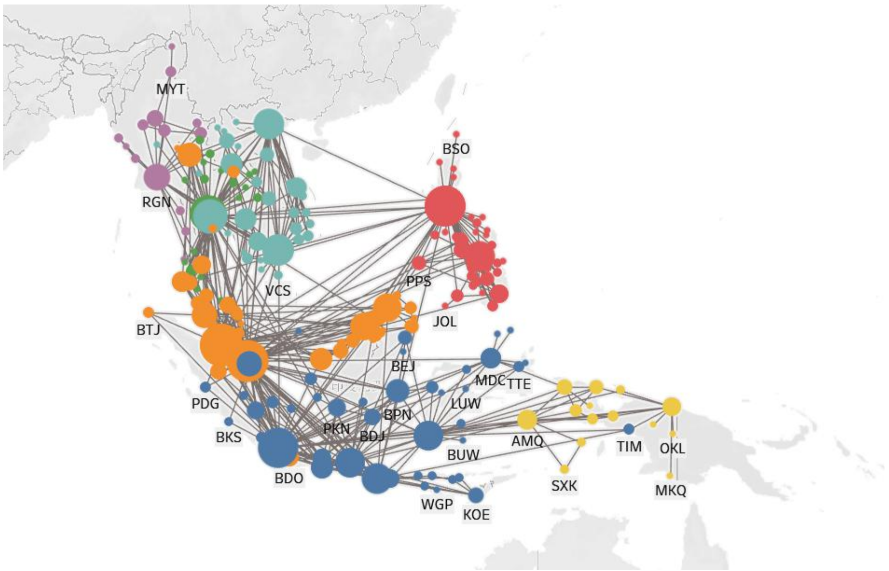


3. ASEAN Overview

3.1 Communities within Southeast Asia

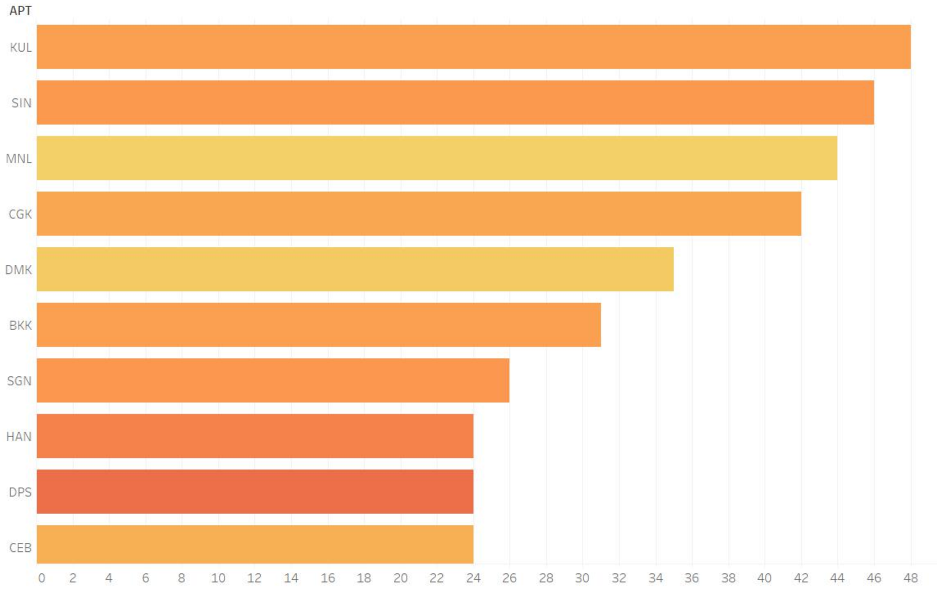
The Association of Southeast Asian Nations is a regional intergovernmental organization comprising ten countries in Southeast Asia. ASEAN countries have many economic zones (industrial parks, eco-industrial parks, special economic zones, technology parks, and innovation districts) (see reference for comprehensive list from 2015). In 2018, eight of the ASEAN members are among the world's outperforming economies, with positive long-term prospect for the region. ASEAN's Secretariat projects that the regional body will grow to become the world's fourth largest economy by 2030. Therefore, analyzing the aviation industry of ASEAN countries is quite meaningful, since with the rapid growth of the economy in ASEAN, the airlines will also keep increasing robustly.
Community categories in ASEAN with two different algorithms.



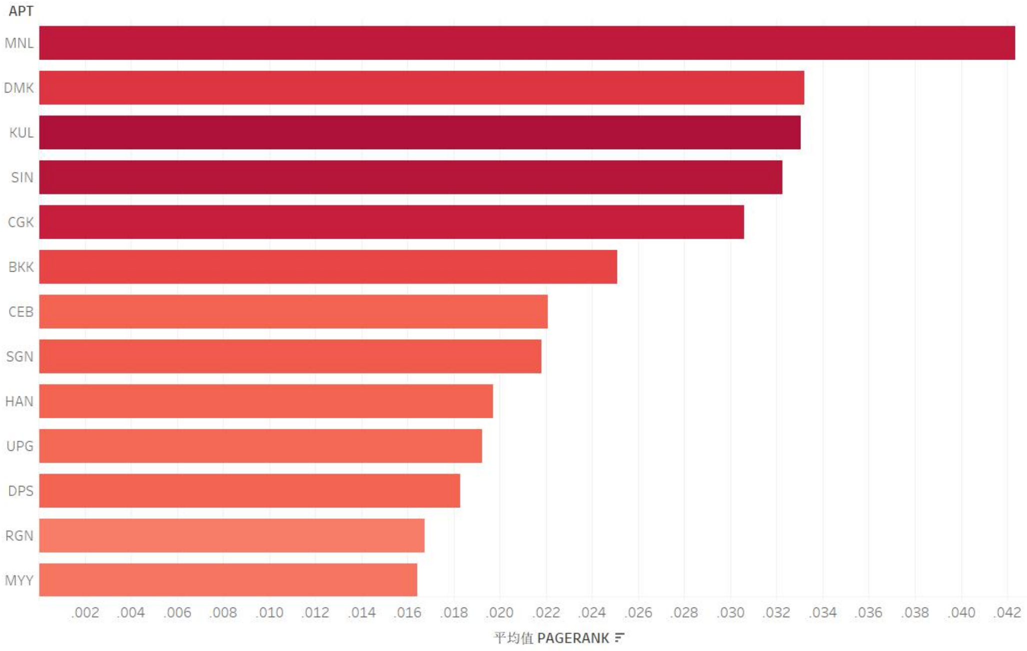


The first figure shows the airlines between different ASEAN countries, and the size of the bubbles represent how many flights within ASEAN the airport holds. We can easily infer from the figure that the largest airport according to the number of flights is Singapore Changi Airport, and Soekarno-Hatta International Airport in Jakarta takes the second place. The second figure is only slightly different from the first one. The first one used Louvain algorithm to detect communities, while the second one deployed Greedy Modularity Maximization. The results showed a similar pattern. The community of the Indonesia in second picture is classified to be the largest one in ASEAN. Besides, the size of the bubbles also was given different meanings in the second graph. The size represents the degree of different airports. With this measurement, the Kuala Lumpur International Airport ranks top among the ASEAN airports with a total connection of 48 other ASEAN airports, and Singapore Changi Airport ranks the second with 46 connections, which is probably because Singapore only connect to the most important airports in ASEAN, while the airport of KL has many domestic airlines, which are not important destinations for Singapore. The graph below used another dimension to show the degrees of different airports in ASEAN.

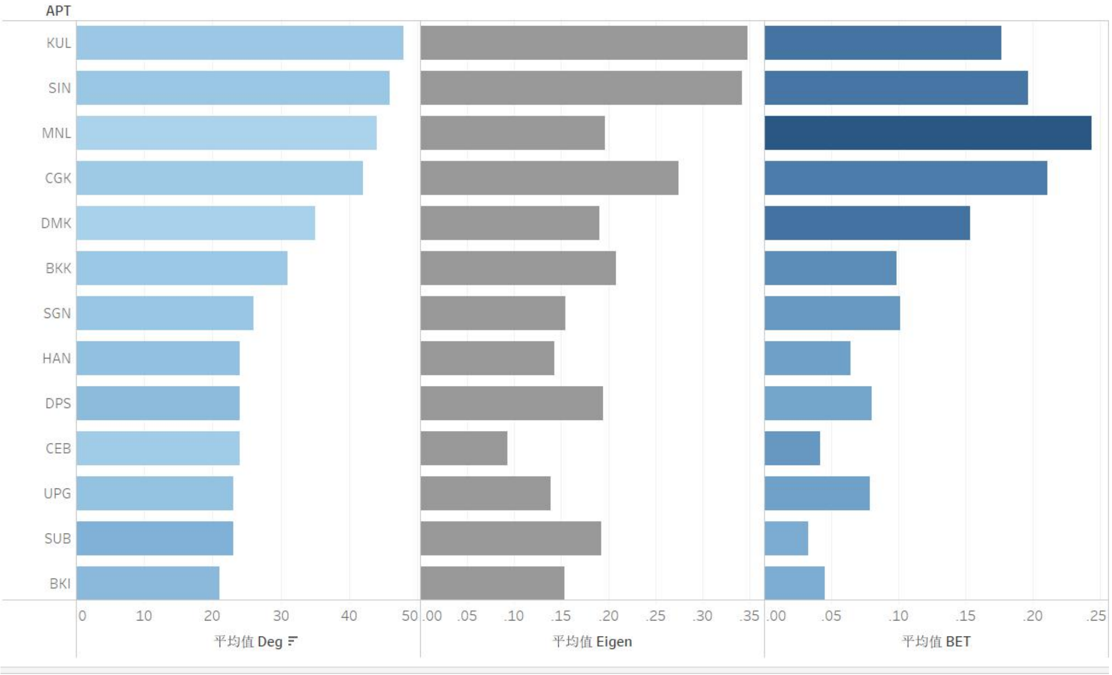
3.2 Network Statistics and measurements



This graph shows the rank of PageRank score of the ASEAN airports. What is quite surprising is that the Ninoy Aquino International Airport in Manila takes the first place. Why is Manila’s PageRank so high? After a careful scrutiny, we found out that it’s actually determined by the geographical environment. Philippine is a country with thousands of islands. (Over 7000). Therefore, it has quite many small airports. So, the Manila airport both has connection with large airports within ASEAN, it also has quite a lot domestic airlines, which might only are connected to the capital airport of Philippine.

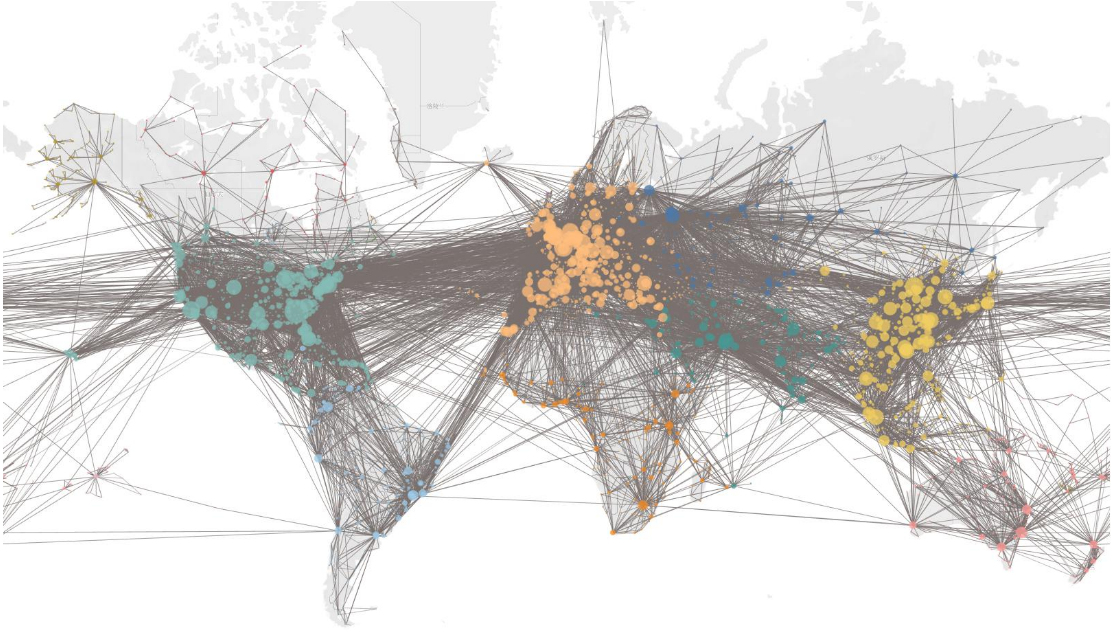


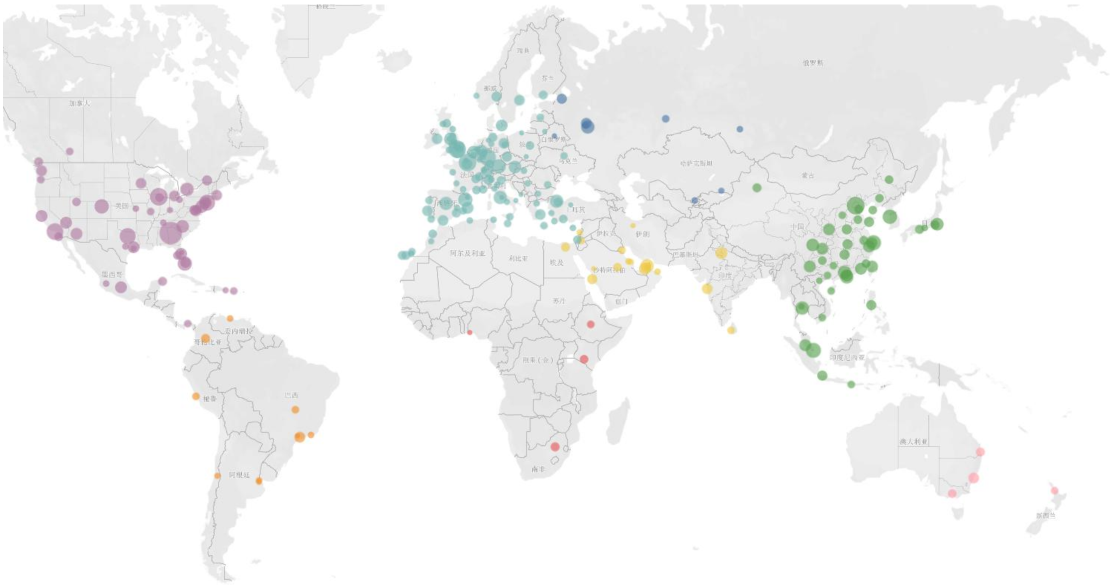
The following graph shows the ranks of ASEAN airports in 3 different metrics, which namely are degree, eigenvector, and betweenness. The first graph used color to show the clustering coefficient of different countries, and the third one used color to represent the PageRank of different countries. It’s interesting to note that these three metrics showed a similar pattern with each other. And betweenness and PageRank showed a high similarity. However, the clustering coefficient performs differently. The higher the degree is, the lower the clustering coefficient is.

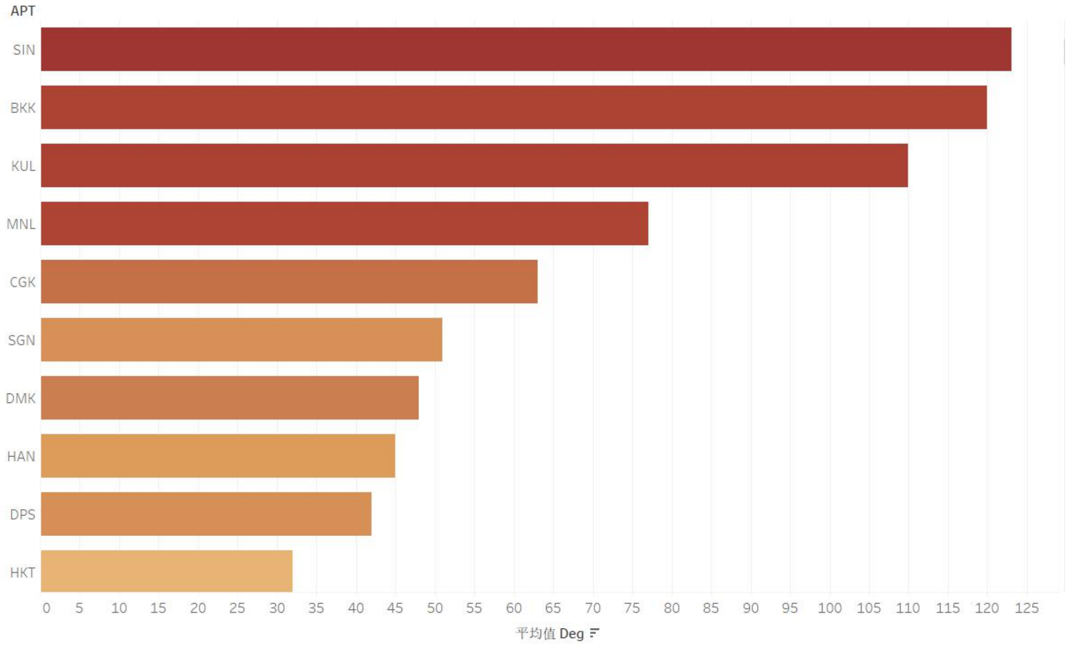


3.3 ASEAN countries from a global perspective

The following two map showed the communities across the world using two different algorithms.



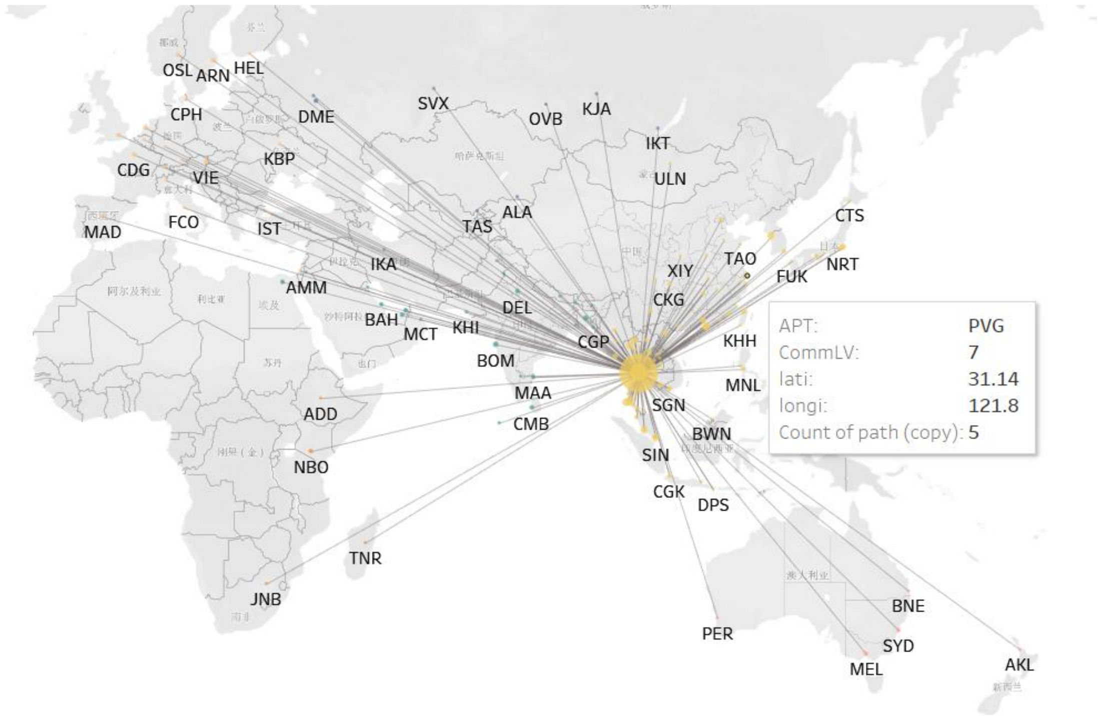




By analyzing the global airlines and flights, we constructed the above graph. It basically shows the degree of different ASEAN airport based on their connections with global airports. This time, the Changi Airport win the first place again. It has connect 123 global airports, and Suvarnabhumi Airport in Bangkok takes the second place by connecting 120 international airports, which partly show how globalized these Asian countries are.

3.4 A closer scrutiny of the international tourist destination:

Bangkok-The pearl of SEA



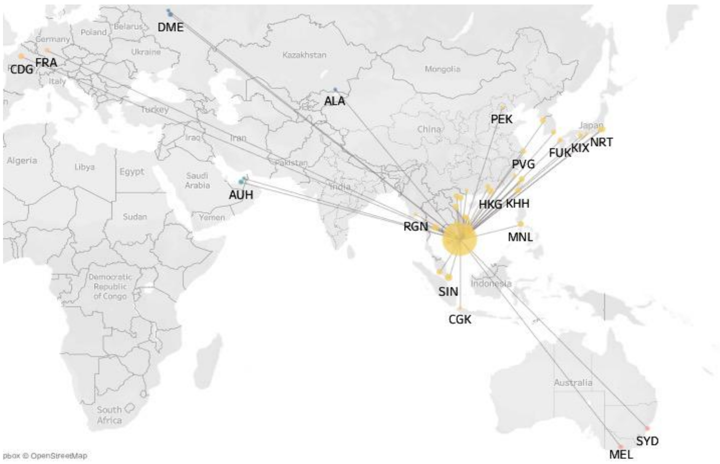
Now, let’s take a closer scrutiny of the two airports in Bangkok, which are DMK Don Mueang International Airport and BKK Suvarnabhumi Airport. These two airports has many airlines globally, mainly to east Asia and west Europe, which shows that Thailand is a very popular destination for Chinese people and Europeans. However, what is quite surprising is that Bangkok has no direct flight to North America and South America, which we thought it is might because the dataset is kind of outdated. After some simple search using google, we found out that it is true that Bangkok has no direct flight to the States. Therefore, we think with the rapid growth of Asian economy, the need of direct flight to the US will be increasing in the long term. We highly recommend that the Airlines companies develop some flights from Bangkok to the States or other countries in North America or South America.

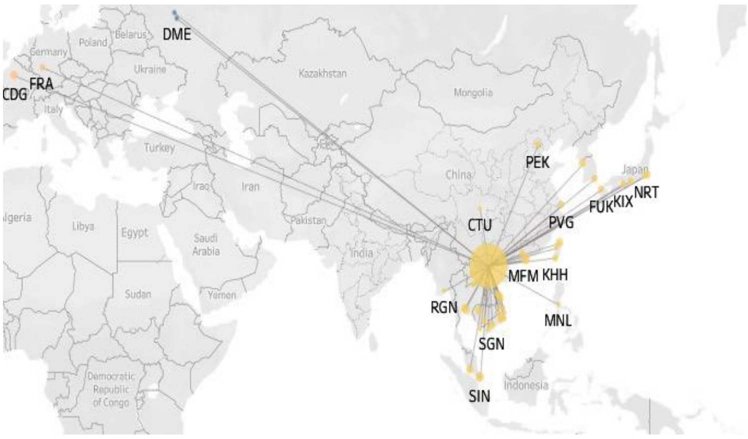
4. A Rising Star in ASEAN---Vietnam

In 2019, the trade war between China and America is one of most important political events in the world. Although the economy growth is decreased due to the tension of that geopolitical event, many famous agencies reported there is a winner from this event that is Vietnam. According to the reports from World Bank, Vietnam now is one of the most dynamic emerging countries in East Asia region. However, Vietnam is still far away from a developed country so that it’s sufficient to consider it as a rising star in Southeast Asia Community. Therefore, Hanoi International Airport(HAN) and Ho Chi Minh City international airport(SGN) are 2 representatives of Vietnam.

4.1 Analysis of Vietnam(HAN and SGN) from routes dataset

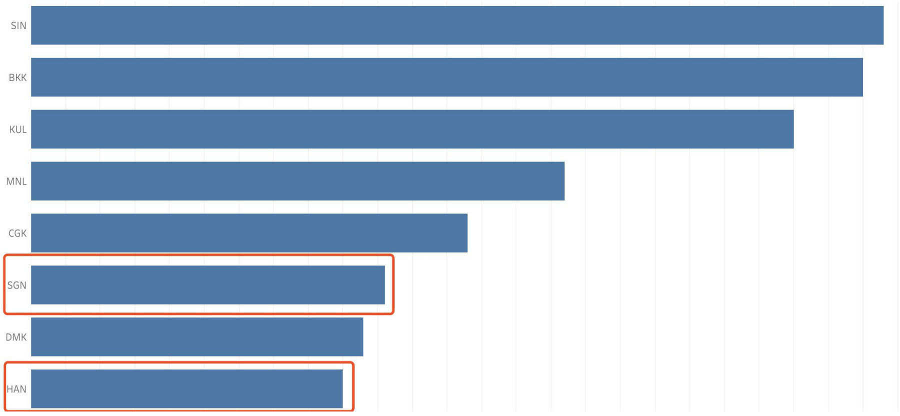
4.1.1 Flights Routes



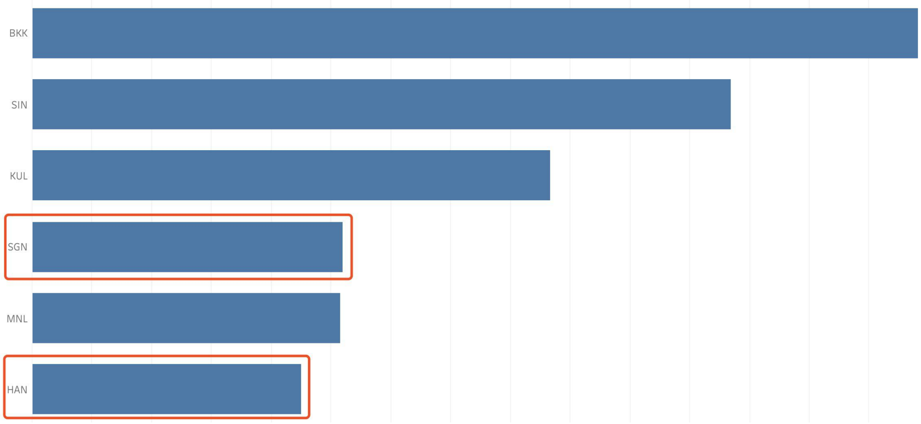


The routs.dat file was updated 3 years ago. We can see that there are not many direct flights from Vietnam to other countries. Except ASEAN countries, both Hanoi can Ho Chi Minh City can only go to limited countries in Europe (France, Germany and Russia). Mainly flights are distributed in East Asia (China, Japan and Korea). Comparing against the complex flights network in Bangkok, Vietnam still has a huge gap to fill with.

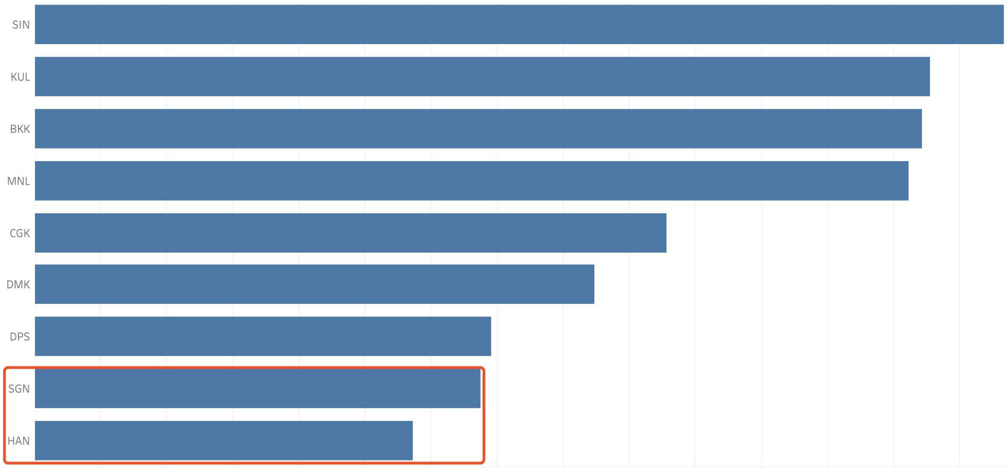
4.1.2 Network Statistics

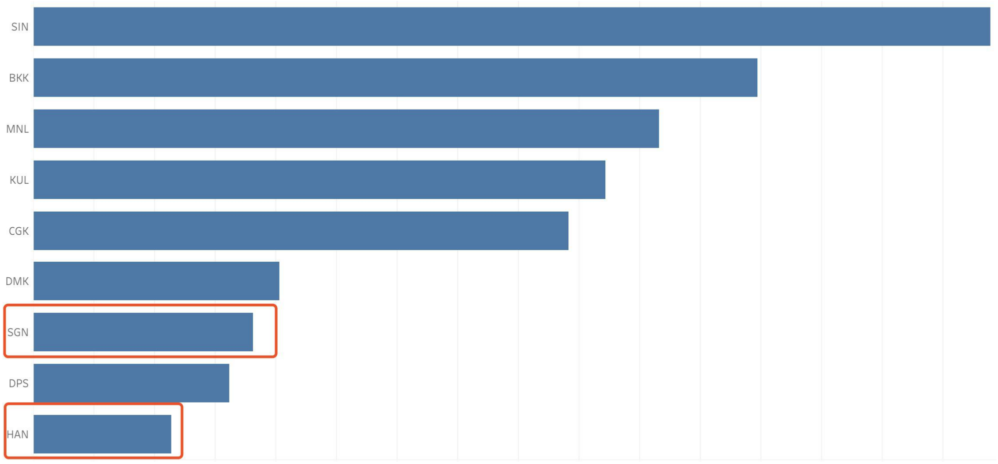


This is the degree distribution in ASEAN countries. We can see that both SGN and HAN are only half of the Singapore and Bangkok.



This is the hub score distribution in ASEAN countries. The comprehensive result is quite similar to the degree distribution.

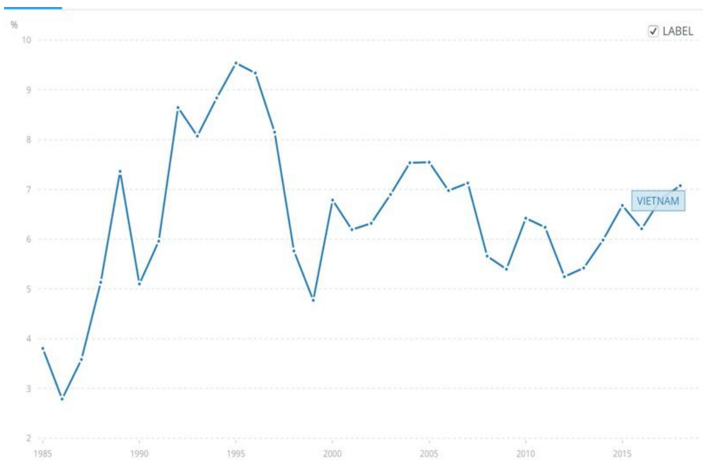




From the PageRank score and Betweenness Centrality, we can find that HAN plays quiet badly in the betweenness centrality. To sum up, both HAN and SGN have much potential to realize which is the premise of a rising star.

4.2 Economic Analysis and Prediction of Vietnam

4.2.1 Economy Growth Curve



After beginning “Revolution” activity in 1986, Vietnam’s economic growth is quite obvious. Over past 30 tears, the average growth rate is nearly about 7 percent. The government estimated that Vietnam also gained 7.1% in 2019.

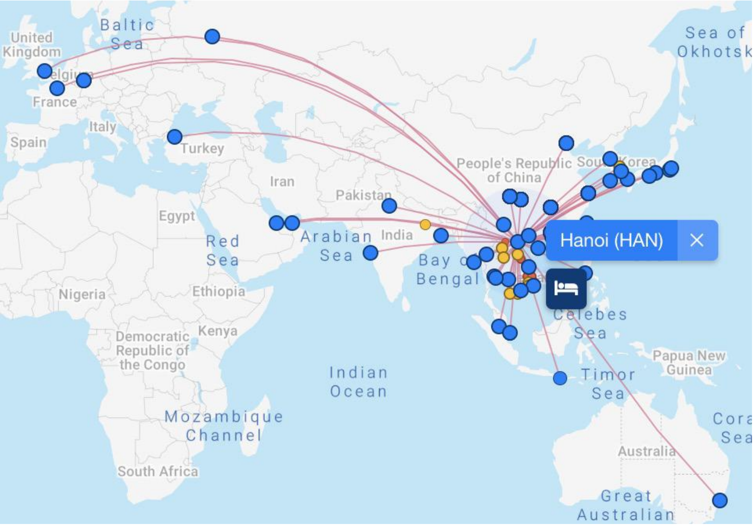
4.2.2 Economic related reports and comments

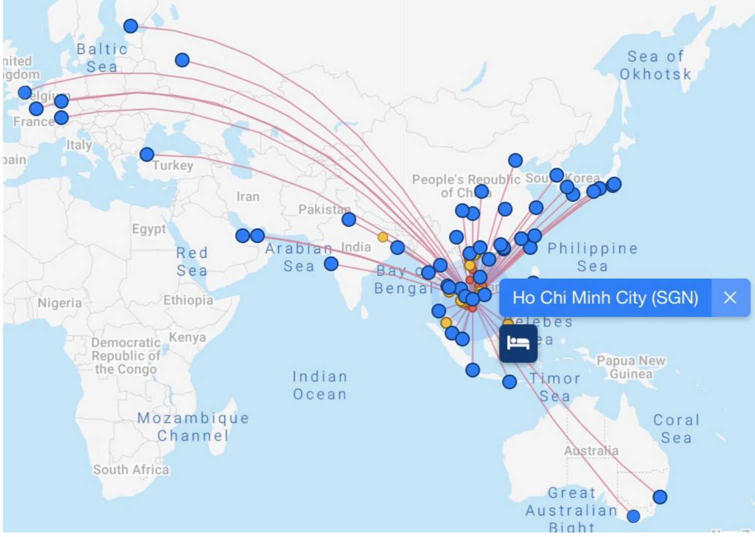
HANOI, Oct. 29 (Xinhua) --- Vietnam welcomed roughly 4.2 million Chinese visitors in the first 10 months of this year, or 32.6 percent of the total international arrivals, posting a year-on-year rise of 28.8 percent, the Vietnam National Administration of Tourism said on Monday.

Forbes --- Vietnam Is Becoming The Big Winner In The China Trade Wars: Sourcing Journal, an online industry newsletter, reports that footwear imports from Vietnam are up 11.3% year-to-date and the country's share of the American market is now just over 26 percent. That still trails China's nearly 50% share but marks a significant shift in a product classification that once was overwhelmingly dominated by China.

According to the 2 news above, we can see that some foreign investments are made in Vietnam. Also, the tourists from China skyrocketed a lot. Those facts implies that there will be and must be more direct flights from Vietnam to some hubs in the world for the convenience of tourism and industry development. Last but not least, GDP growth is a key driver of air travel demand. We predict that more flights will be opened to China and other hubs.

4.2.3 Latest Data and Insights





This is the latest flights network of HAN and SGN. We can see that the network are denser than 3-4 years ago. More flights to China and the farthest flight is to London (LHR), which is also new. Those new flights to famous hubs will increase the PageRank Score of HAN and SGN, which implies Vietnam plays a more important role in ASEAN. Also, the network analysis is more helpful in decision making or business intelligence when combined with some background knowledge and information.