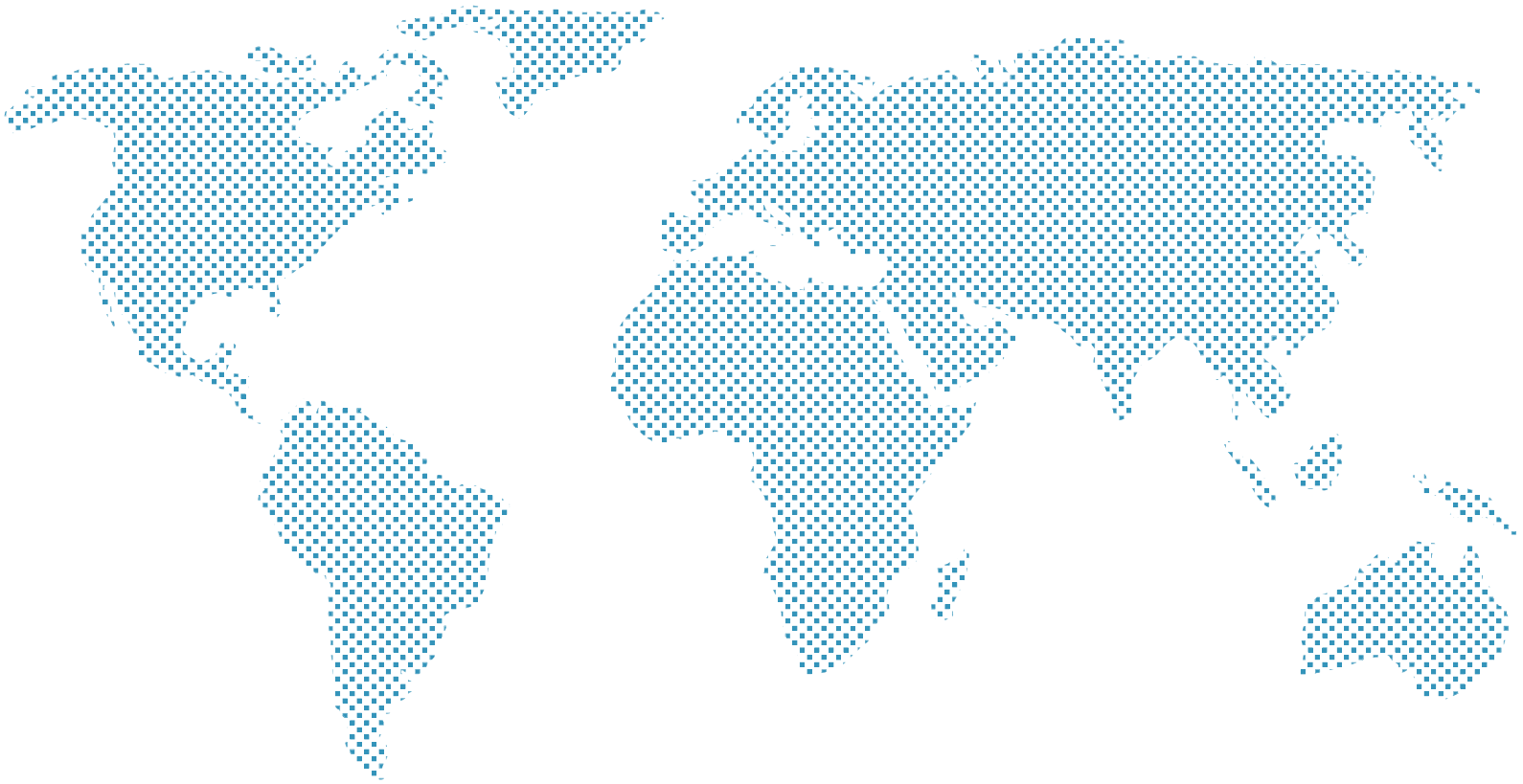
AIR ROUTES NETWORK ANALYSIS

MH8351 Web Analytics Project Report



Chan Ka Wei Sherice	(G1902324A)
Koh Chin Nam Andrew	(G1901964B)
Lim Guowei	(G1901873A)
Lee Sin Tat	(G1800658J)

1. INTRODUCTION

1.1. Background

According to the International Air Transport Association (IATA), there were 3.8 billion air passengers in 2016. Present trends in air transport suggest that passenger numbers could double to 8.2 billion in 2037. The strong growth in air travel presents a need for industry players to improve on air connectivity by generating new routes and build more infrastructure such as airports.

The IATA's Director General and CEO, Alexandre de Juniac also observes a geographical reshuffling of world air traffic to the East. The Asia-Pacific region is expected to drive the biggest growth in air travel, with China eventually displacing the United States as the world's largest aviation market in the mid-2020s.

The air transportation system can be represented as a network. Understanding the network of airlines would be useful for guiding developments in the aviation industry and for obtaining business insights.

1.2. Objectives

This report provides an analysis of flight routes and airlines. The objectives of this study are listed as follows:

- To identify geographic regions by analysing flight route networks
- To determine key players (airport hubs) in the geographic regions identified
- To understand the network behavior of airlines

2. METHODOLOGY

2.1. Data Description

Two data sets were obtained from https://openflights.org: airports.dat and routes.dat. Tables 1 and 2 below show the descriptions of variables.

Variable Name	**Description of Variable**
Airport 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
ICAO	4-letter ICAO code
Latitude	Latitude in decimal degrees
Longitude	Longitude in decimal degrees
Altitude	Altitude in feet
Timezone	Hours offset from UTC
DST	Daylight savings time
Tz database time zone	Timezone in Olson format
Type	Type of the airport
Source	Source of the observed data

Table 1. Description of variables in airports.dat file

Variable Name	**Description of Variable**
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	Indicates if the flight is a codeshare
Stops	Number of stops on this flight
Equipment	3-letter codes for plane type(s) generally used on this flight

Table 2. Description of variables in routes.dat file

2.2. Approach

The analyses in this project were executed using Python programming language. Numpy and Pandas packages were mainly used for data preprocessing and manipulation. Graph analyses such as network measures calculation, community detection and key player determination were performed using NetworkX package. In addition to the community detection algorithms from NetworkX, spectral clustering from Scikit-learn package was used as well. Visualisation of network was plotted using Matplotlib Basemap Toolkit.

The experimental approach for this project is illustrated in ***Figure 1***. After the datasets were downloaded and cleaned, the properties of the network such as degree distribution and clustering coefficients were computed. Various community detection algorithms such as Clauset-Newman-Moore modularity maximization, label propagation, Louvain best partition and spectral clustering were compared based on the coverage and performance scores. Subsequently, key players existing in each region partitioned by the down-selected community detection algorithm were determined based on various centrality metrics.

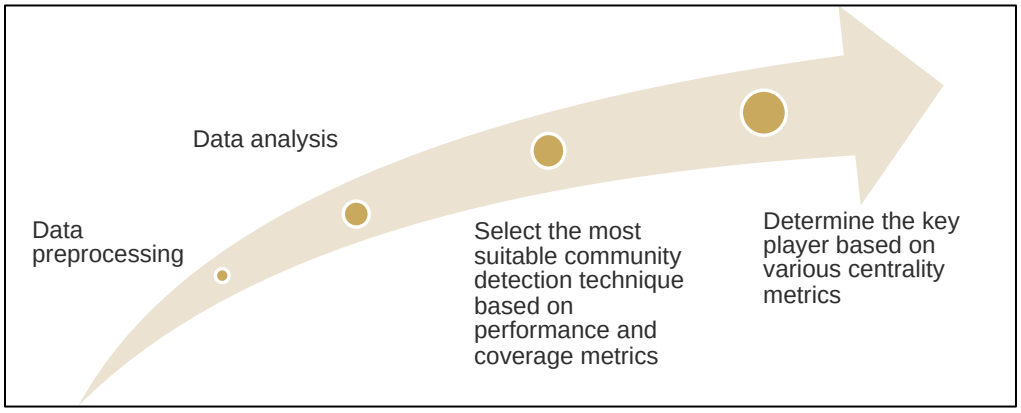


Figure 1. Experimental approach for flight routes network analysis

2.3. Algorithms and Metrics

The algorithms and metrics used for analyses are described in Tables 3 and 4.

Algorithm	**Description**
Clauset-Newman-Moore modularity maximization	The algorithm initially associates each node with a community. Then, it repeatedly combines the communities of which the union produces the highest increase to the modularity of the community structure.
Label propagation	Every node is initialised with a unique label. Iteratively, labels are reassigned such that each node takes the majority label of its neighbours.
Louvain best partition	The method first looks for "small" communities by optimising modularity locally, then aggregates nodes belonging to same community and build a new network. Repeat the process iteratively until maximum modularity is obtained.
Spectral clustering	The method makes use of eigenvalues of the adjacency matrix of the graph to perform dimensionality reduction before clustering in fewer dimensions.
K-means clustering	Every cluster is identified by a centre and at the beginning start with k arbitrary disjoint sets. Iteratively, calculate the centre of the partition and modify these partitions adding closest nodes.
Strongly connected components	The algorithm checks if there is a path between all pairs of vertices in a directed graph. A strongly connected component graph is a maximal subgraph.

Table 3. Descriptions of algorithms used.

Metric	**Description**
Coverage	Fraction of covered edges over total number of edges
Performance	Fraction of correctly classified pairs of nodes
Degree centrality	The number of links incident upon a node
Betweenness centrality	This quantifies the number of times a node acts as a bridge along the shortest path between two other nodes, and
VoteRank	This metric computes a ranking of the nodes in the graph based on a voting scheme. With VoteRank, all nodes vote for each neighbour and the node with the highest score is elected iteratively. The voting ability of neighbors of elected nodes will be decreased in subsequent turn.

Table 4. Descriptions of metrics used.

2.4. *Assumptions*

The following assumptions were made in order to perform analyses using different algorithms.

Assumption	**Algorithm used**
The graph is undirected.	Clauset-Newman-Moore modularity maximization Label propagation Louvain best partition Spectral clustering K-means clustering
The graph is directed.	Strongly connected component
The edges of the graph are non-weighted.	Clauset-Newman-Moore modularity maximization Label propagation Louvain best partition K-means clustering Strongly connected component

Table 5. Assumptions for network analysis.

2.5. *Definition*

In the analysis of flight route network, an airport is represented as a node and the route between two airports is defined as an edge.

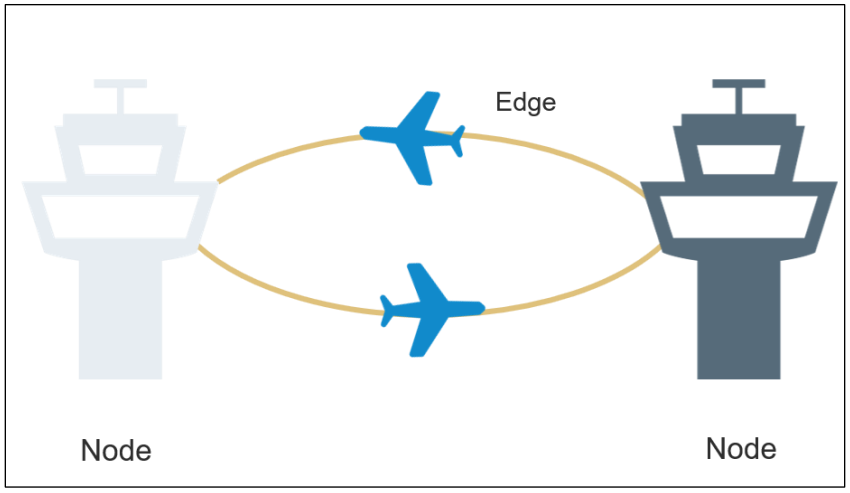


Figure 2. Basic representation of flight route network.

3. RESULTS

3.1. Network measures

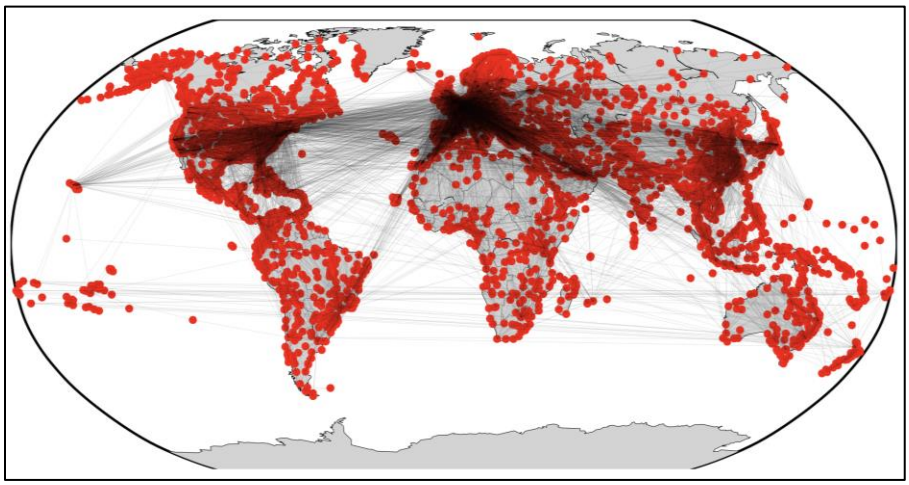


Figure 3. Visualisation of flight routes dataset.

The airports(vertices) and flight routes (edges) are plotted into a graph and superimposed on Earth map using as shown in Figure 3. The original dataset contains 3214 nodes and 18,858 edges in total. The density and average clustering coefficient are calculated to be 0.365% and 0.492 respectively. The degree distribution plotted as shown in Figure 4 indicates the characteristic of a power law distribution which suggests a scale-free network. This conforms to our understanding of global flight routes network which should consist of several air hubs located in different regions.

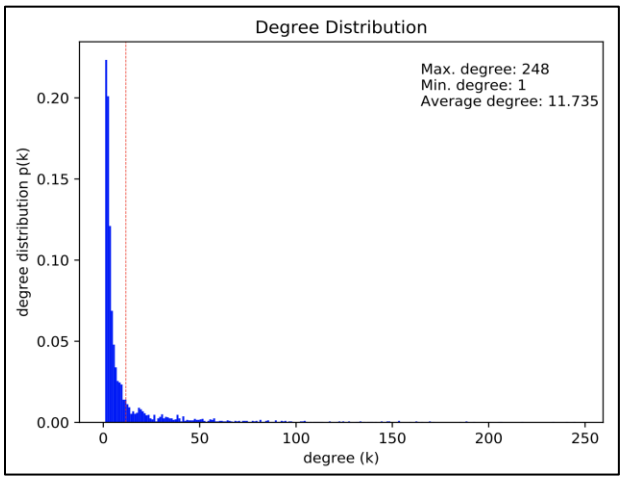


Figure 4. Degree distribution of flight routes dataset.

3.2. Community detection

Three community detection algorithms were first used in attempt to identify geographic regions based on flight routes data, namely Clauset-Newman-Moore modularity maximization (CNM), Louvain best partition and label propagation. Top seven largest subgraphs generated from each algorithm are plotted from Figures 5 to 7***Figure*** . The communities generated by all three algorithms are satisfactory as the global regions are defined clearly. The good quality of the clustering is expressed by the coverage and performance scores which are very close to the value of 1 as shown in Table 6.

Another frequently used algorithm which does not yet exist in Python Networkx library, spectral clustering, was also attempted for community detection. The algorithm consists of three main steps: (1) compute a similarity graph, (2) project the graph data into a low-dimensional space or eigenvector embedding map, and finally (3) K-Means on the vector to create cluster or communities. We first performed the full spectral clustering procedures on the non-weighted and weighted (based on distance) graph. The results underperformed as shown in Figures 8 and 9, as the algorithm recognized mainly one large community. This was likely due to a combination of random initialization in the final K-Means step within the Spectral Clustering algorithm and the presence of strong inter-continental connections/edges. We intervened in the final K-mean clustering stage by laying some ground truth for initialization[1]. Key players (airport hubs) were pre-identified for each region (stated in the dataset) through some of the centrality measures as discussed in Section 3.3. As a result, we observed a marked improvement in the visualization in Figures 10 and 11 and clustering scores in Table 6. However, the clustering performance and coverage were not better than the previous autonomous algorithm and were significantly dependent upon the sufficiency and accuracy of our ground truth.

Since label propagation produced adequate regions partitioning and exhibited relatively high-quality scores, the communities generated by this algorithm were further analyzed to determine their key players.

[1] *Initialization method for spectral clustering was referenced from https://github.com/lchenbb/NTDS2019_Project*

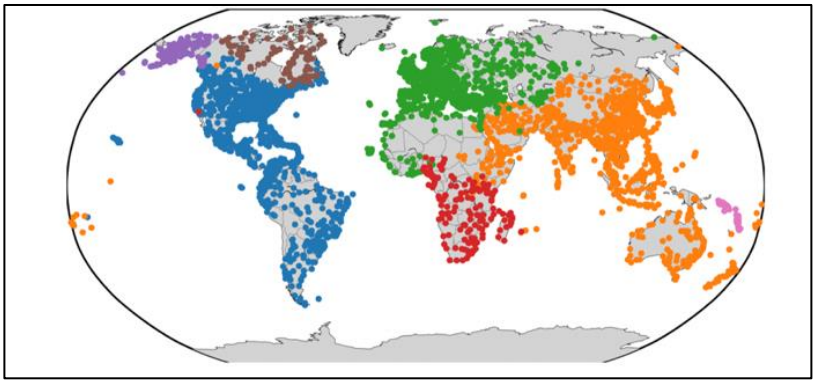


Figure 5. Communities generated using Clauset-Newman-Moore modularity maximization

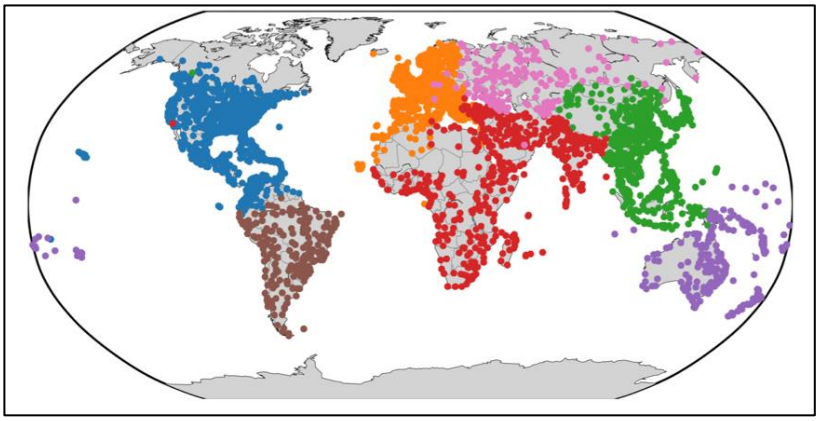


Figure 6. Communities generated using Louvain best partition

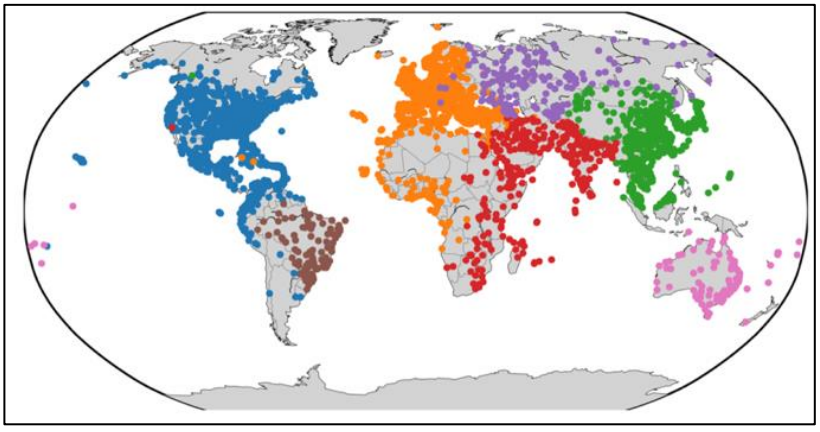


Figure 7. Communities generated using label propagation

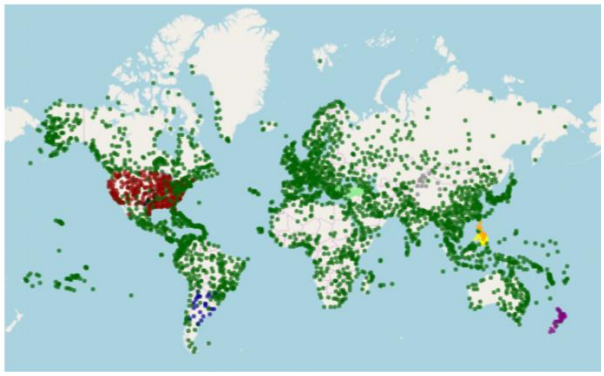


Figure 8. Communities generated using non-weight spectral clustering

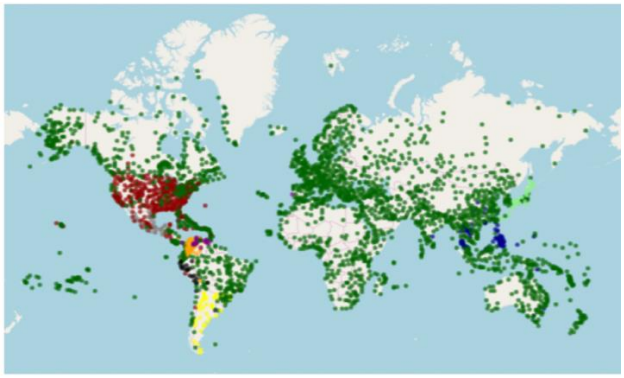


Figure 9. Communities generated using weight spectral clustering

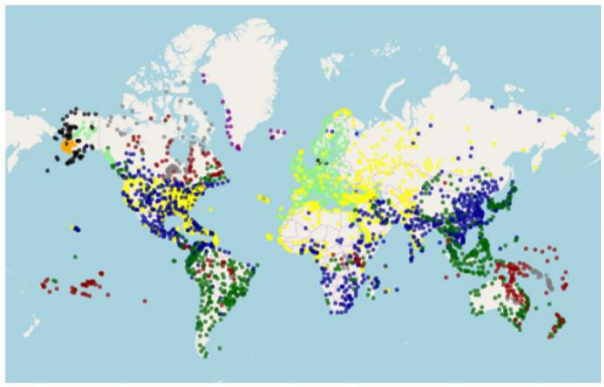


Figure 10. Communities generated semi-supervised non-weighted spectral clustering

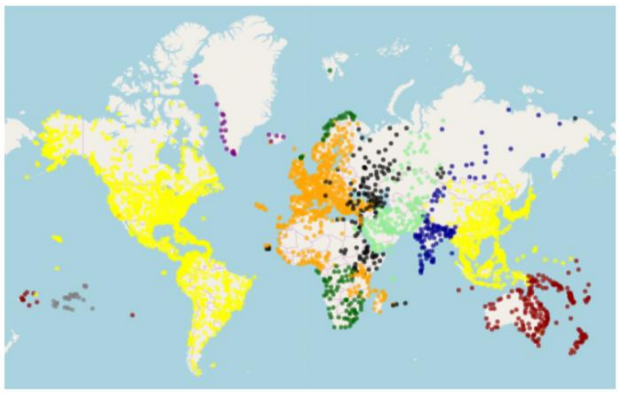


Figure 11. Communities generated using semi-supervised weighted spectral clustering

Algorithm	**Performance**	**Coverage**
Clauset-Newman-Moore modularity maximization	0.796	0.904
Label propagation	0.892	0.852
Louvain best partition	0.866	0.867
Spectral clustering (non-weighted)	0.363	0.792
Spectral clustering (weighted)	0.225	0.912
K-means clustering (non-weighted)	0.807	0.767
K-means clustering (weighted)	0.663	0.847

Table 6. Performance and coverage comparisons for community detection algorithms

3.3. Analysis of key players (airport hubs)

Key player is defined as the most important node in a social network and for this project, the most important airport hub for each community generated by label propagation will be identified using the following centrality metrics:

- Degree Centrality
- Betweenness Centrality
- Voterank

The following results are observed:

- The key player for blue region is Hartsfield Jackson Atlanta International Airport.

blue region

	Airport ID	degree_centrality	Name	City	Country
0	3682	0.28	Hartsfield Jackson Atlanta International Airport	Atlanta	United States
1	3830	0.254	Chicago O'Hare International Airport	Chicago	United States
2	3670	0.253	Dallas Fort Worth International Airport	Dallas-Fort Worth	United States
3	3751	0.234	Denver International Airport	Denver	United States
4	3550	0.221	George Bush Intercontinental Houston Airport	Houston	United States

	Airport ID	betweenness_centrality	Name	City	Country
0	3682	0.104	Hartsfield Jackson Atlanta International Airport	Atlanta	United States
1	3751	0.102	Denver International Airport	Denver	United States
2	3670	0.098	Dallas Fort Worth International Airport	Dallas-Fort Worth	United States
3	3830	0.09	Chicago O'Hare International Airport	Chicago	United States
4	3550	0.074	George Bush Intercontinental Houston Airport	Houston	United States

	Airport ID	voterank_centrality	Name	City	Country
0	3682	1	Hartsfield Jackson Atlanta International Airport	Atlanta	United States
1	3830	2	Chicago O'Hare International Airport	Chicago	United States
2	3670	3	Dallas Fort Worth International Airport	Dallas-Fort Worth	United States
3	3751	4	Denver International Airport	Denver	United States
4	3550	5	George Bush Intercontinental Houston Airport	Houston	United States

Figure 12. Top 5 airports in the blue region for various centrality measures.

- The key player for orange region is Amsterdam Airport Schiphol.

orange region					
	Airport ID	degree_centrality	Name	City	Country
0	580	0.276	Amsterdam Airport Schiphol	Amsterdam	Netherlands
1	1382	0.254	Charles de Gaulle International Airport	Paris	France
2	346	0.246	Munich Airport	Munich	Germany
3	548	0.244	London Stansted Airport	London	United Kingdom
4	1701	0.242	Atatürk International Airport	Istanbul	Turkey
	Airport ID	betweenness_centrality	Name	City	Country
0	1701	0.097	Atatürk International Airport	Istanbul	Turkey
1	1382	0.069	Charles de Gaulle International Airport	Paris	France
2	644	0.055	Oslo Lufthavn	Oslo	Norway
3	3941	0.055	Eleftherios Venizelos International Airport	Athens	Greece
4	580	0.054	Amsterdam Airport Schiphol	Amsterdam	Netherlands
	Airport ID	voterank_centrality	Name	City	Country
0	580	1	Amsterdam Airport Schiphol	Amsterdam	Netherlands
1	1382	2	Charles de Gaulle International Airport	Paris	France
2	548	3	London Stansted Airport	London	United Kingdom
3	1701	4	Atatürk International Airport	Istanbul	Turkey
4	346	5	Munich Airport	Munich	Germany

Figure 13. Top 5 airports in the orange region for various centrality measures.

- The key player for green region is Beijing Capital International Airport.

green region					
	Airport ID	degree_centrality	Name	City	Country
0	3364	0.375	Beijing Capital International Airport	Beijing	China
1	3370	0.321	Guangzhou Baiyun International Airport	Guangzhou	China
2	3406	0.305	Shanghai Pudong International Airport	Shanghai	China
3	3395	0.264	Chengdu Shuangliu International Airport	Chengdu	China
4	3382	0.237	Kunming Changshui International Airport	Kunming	China
	Airport ID	betweenness_centrality	Name	City	Country
0	3364	0.153	Beijing Capital International Airport	Beijing	China
1	2359	0.12	Tokyo Haneda International Airport	Tokyo	Japan
2	3406	0.084	Shanghai Pudong International Airport	Shanghai	China
3	3157	0.076	Don Mueang International Airport	Bangkok	Thailand
4	3370	0.076	Guangzhou Baiyun International Airport	Guangzhou	China
	Airport ID	voterank_centrality	Name	City	Country
0	3364	1	Beijing Capital International Airport	Beijing	China
1	3370	2	Guangzhou Baiyun International Airport	Guangzhou	China
2	3406	3	Shanghai Pudong International Airport	Shanghai	China
3	3395	4	Chengdu Shuangliu International Airport	Chengdu	China
4	3382	5	Kunming Changshui International Airport	Kunming	China

Figure 14. Top 5 airports in the green region for various centrality measures.

- The key player for red region is Dubai International Airport

red region					
	Airport ID	degree_centrality	Name	City	Country
0	2188	0.268	Dubai International Airport	Dubai	United Arab Emirates
1	2072	0.261	King Abdulaziz International Airport	Jeddah	Saudi Arabia
2	3093	0.225	Indira Gandhi International Airport	Delhi	India
3	2997	0.221	Chhatrapati Shivaji International Airport	Mumbai	India
4	2191	0.204	Sharjah International Airport	Sharjah	United Arab Emirates
	Airport ID	betweenness_centrality	Name	City	Country
0	2072	0.19	King Abdulaziz International Airport	Jeddah	Saudi Arabia
1	2188	0.183	Dubai International Airport	Dubai	United Arab Emirates
2	2997	0.137	Chhatrapati Shivaji International Airport	Mumbai	India
3	3093	0.133	Indira Gandhi International Airport	Delhi	India
4	813	0.129	OR Tambo International Airport	Johannesburg	South Africa
	Airport ID	voterank_centrality	Name	City	Country
0	2188	1	Dubai International Airport	Dubai	United Arab Emirates
1	2072	2	King Abdulaziz International Airport	Jeddah	Saudi Arabia
2	3093	3	Indira Gandhi International Airport	Delhi	India
3	2997	4	Chhatrapati Shivaji International Airport	Mumbai	India
4	2191	5	Sharjah International Airport	Sharjah	United Arab Emirates

Figure 15. Top 5 airports in the red region for various centrality measures.

- The key player for purple region is Domodedovo International Airport

purple region					
	Airport ID	degree_centrality	Name	City	Country
0	4029	0.722	Domodedovo International Airport	Moscow	Russia
1	2948	0.431	Pulkovo Airport	St. Petersburg	Russia
2	2975	0.312	Koltsovo Airport	Yekaterinburg	Russia
3	4078	0.243	Tolmachevo Airport	Novosibirsk	Russia
4	2988	0.236	Vnukovo International Airport	Moscow	Russia
	Airport ID	betweenness_centrality	Name	City	Country
0	4029	0.528	Domodedovo International Airport	Moscow	Russia
1	2948	0.183	Pulkovo Airport	St. Petersburg	Russia
2	2988	0.101	Vnukovo International Airport	Moscow	Russia
3	2975	0.095	Koltsovo Airport	Yekaterinburg	Russia
4	2923	0.054	Yakutsk Airport	Yakutsk	Russia
	Airport ID	voterank_centrality	Name	City	Country
0	4029	1	Domodedovo International Airport	Moscow	Russia
1	2948	2	Pulkovo Airport	St. Petersburg	Russia
2	2975	3	Koltsovo Airport	Yekaterinburg	Russia
3	2988	4	Vnukovo International Airport	Moscow	Russia
4	4078	5	Tolmachevo Airport	Novosibirsk	Russia

Figure 16. Top 5 airports in the purple region for various centrality measures.

- The key player for brown region is Viracopos International Airport

brown region					
	Airport ID	degree_centrality	Name	City	Country
0	2578	0.521	Viracopos International Airport	Campinas	Brazil
1	2564	0.427	Guarulhos - Governador André Franco Montoro In	Sao Paulo	Brazil
2	2531	0.396	Presidente Juscelino Kubistschek International Air	Brasilia	Brazil
3	2537	0.323	Tancredo Neves International Airport	Belo Horizonte	Brazil
4	2618	0.271	Congonhas Airport	Sao Paulo	Brazil
	Airport ID	betweenness_centrality	Name	City	Country
0	2578	0.291	Viracopos International Airport	Campinas	Brazil
1	2531	0.206	Presidente Juscelino Kubistschek International Air	Brasilia	Brazil
2	2564	0.141	Guarulhos - Governador André Franco Montoro In	Sao Paulo	Brazil
3	2537	0.123	Tancredo Neves International Airport	Belo Horizonte	Brazil
4	2548	0.111	Marechal Rondon Airport	Cuiaba	Brazil
	Airport ID	voterank_centrality	Name	City	Country
0	2578	1	Viracopos International Airport	Campinas	Brazil
1	2564	2	Guarulhos - Governador André Franco Montoro In	Sao Paulo	Brazil
2	2531	3	Presidente Juscelino Kubistschek International Air	Brasilia	Brazil
3	2537	4	Tancredo Neves International Airport	Belo Horizonte	Brazil
4	2618	5	Congonhas Airport	Sao Paulo	Brazil

Figure 17. Top 5 airports in the brown region for various centrality measures.

- The key player for pink region is Sydney Kingsford Smith International Airport.

pink region					
	Airport ID	degree_centrality	Name	City	Country
0	3361	0.548	Sydney Kingsford Smith International Airport	Sydney	Australia
1	3320	0.376	Brisbane International Airport	Brisbane	Australia
2	3339	0.344	Melbourne International Airport	Melbourne	Australia
3	3341	0.215	Adelaide International Airport	Adelaide	Australia
4	3322	0.194	Cairns International Airport	Cairns	Australia
	Airport ID	betweenness_centrality	Name	City	Country
0	3361	0.462	Sydney Kingsford Smith International Airport	Sydney	Australia
1	3320	0.227	Brisbane International Airport	Brisbane	Australia
2	3341	0.163	Adelaide International Airport	Adelaide	Australia
3	3339	0.157	Melbourne International Airport	Melbourne	Australia
4	3351	0.121	Perth International Airport	Perth	Australia
	Airport ID	voterank_centrality	Name	City	Country
0	3361	1	Sydney Kingsford Smith International Airport	Sydney	Australia
1	3320	2	Brisbane International Airport	Brisbane	Australia
2	3339	3	Melbourne International Airport	Melbourne	Australia
3	3341	4	Adelaide International Airport	Adelaide	Australia
4	3322	5	Cairns International Airport	Cairns	Australia

Figure 18. Top 5 airports in the pink region for various centrality measures.

Hence we identified the key airport hubs for each of the seven largest communities generated by label propagation as follows:

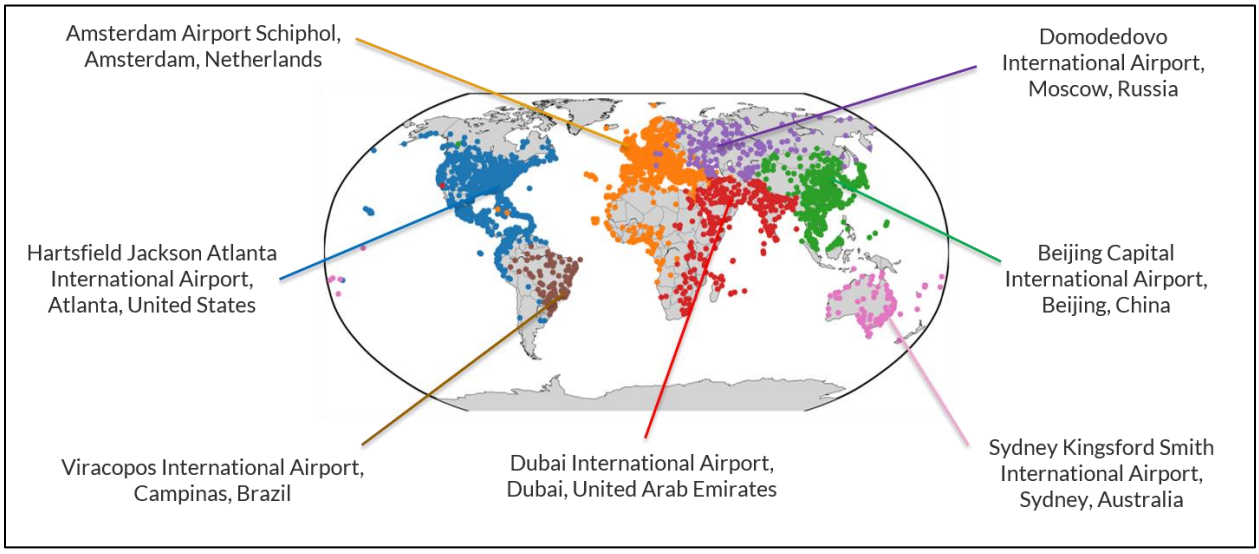


Figure 29. Key airports for each community.

3.4. Analysis of airlines' interaction with airport communities

This section provides analysis into how airlines' network behaves across (inter-community) and within (intra-community) the airport communities that were detected using the "Label Propagation" algorithm proposed in our earlier section. We performed the detection on a graph weighted based on the geodesics distance between each node and obtained the following results using Python:

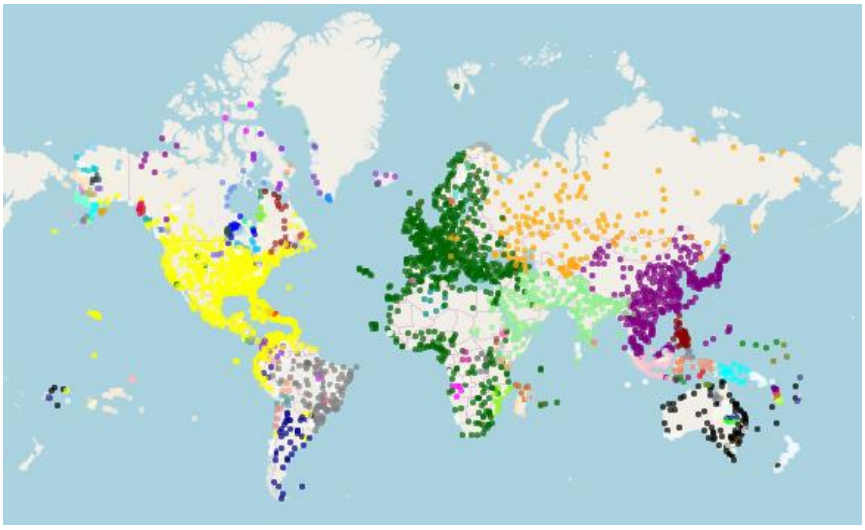


Figure 20. Detected communities using "Label Propagation" algorithm on weighted graph.

The results here were comparable to the results obtained from "Label Propagation" on unweighted graph in the earlier section by virtue of the above visualization and clustering quality measures. A clustering performance score of 0.886 and coverage score of 0.853 was obtained versus 0.892 and 0.852 achieved for the unweighted graph. The following analysis were performed using the detected communities:

3.4.1 Visualization of Airlines Network with Airport Communities

The flight networks of some popular international and domestic airlines were projected upon the detected communities on the world map as shown in the figures below using Python. The highlights from the visualization were as follows:

- Connections within domestic airlines network are generally denser with more observable hubs and authorities as compared to the international airlines network;
- International airlines network tends to span across multiple communities (see multiple coloured nodes in the diagrams below), while the domestic flights network is mostly confined to the same-coloured community nodes.

These observations were consistent with general flight experiences. We expect airlines to generally focus on either connecting across several regions or within a specific geographic market. This validated the quality of our community detection algorithm.

<u>International Airlines' Networks</u>

Figure 21. Singapore Airlines' flight network and its visited (multi-coloured) communities

International Airlines' Networks (cont'd)

Figure 22. United Emirates Airlines' flight network and its visited (multi-coloured) communities

Domestic Airlines' Networks

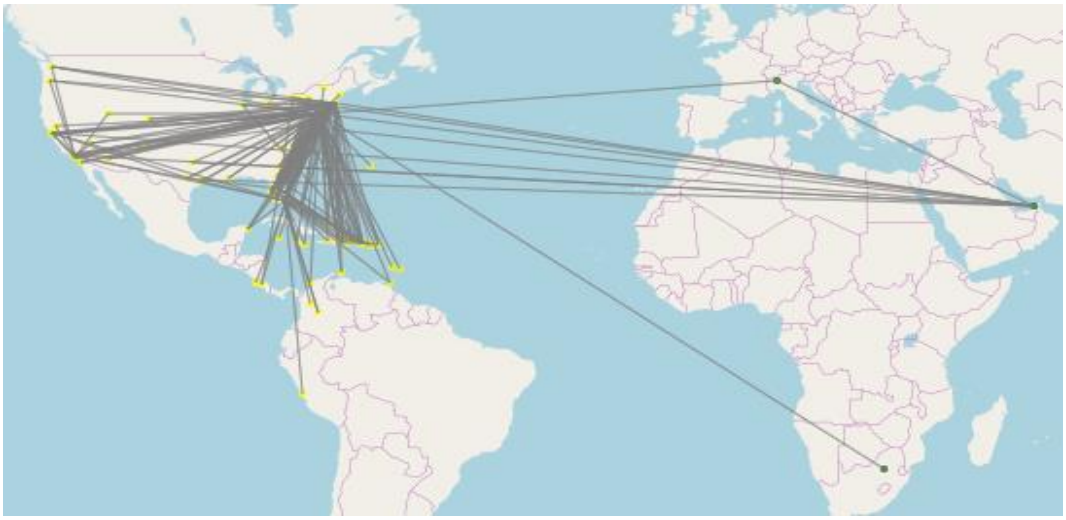


Figure 23. JetBlue (US) Airlines' flight network within a dominant (mostly one-colour) community

Domestic Airlines' Networks (cont'd)

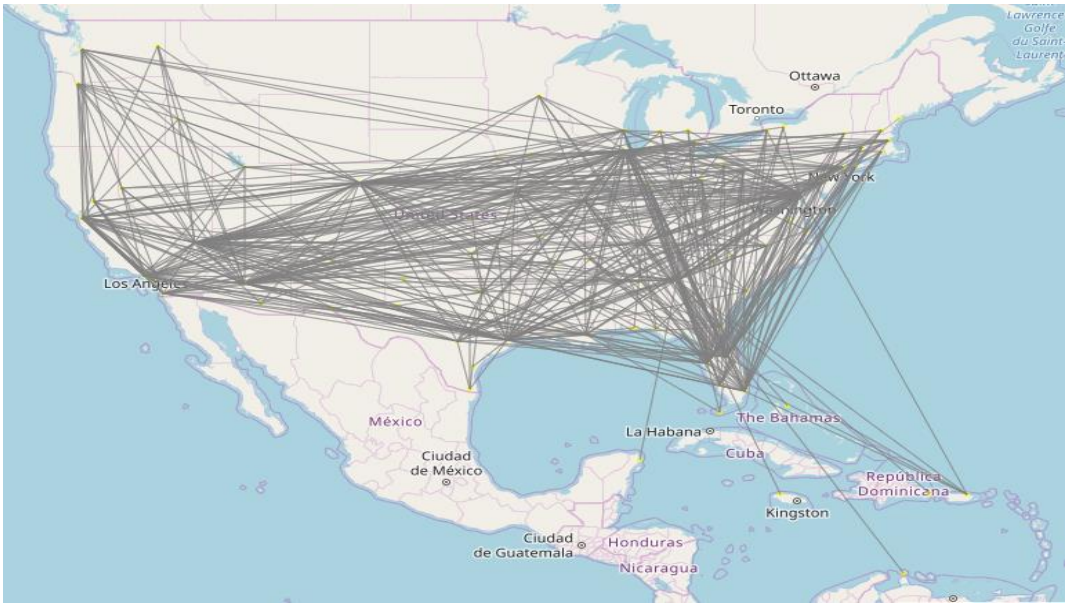


Figure 24. South West (US) Airlines' flight network within a (only one-colour) community

3.4.2 Detection of key airline competitions based on inter-community interaction

Airlines can also efficiently search for closest competitors through evaluating the similarity in both parties' flight network, as represented by their respective visited community sets which was defined earlier using the "Label Propagation" algorithm. Differences in community portfolio could also be identified for market opportunities.

We performed such an analysis for Singapore Airlines ("SQ"), using a Jaccard Coefficient to determine its community portfolio's similarity with all the other airlines using Python. We defined **Jaccard Similarity** as follows: *(Number of similar or **INTERSECT** community sets) / (Total number of community sets visited by SQ, **UNION** with number of sets visited by specific competitor).* The top competitors of SQ were identified as follows:

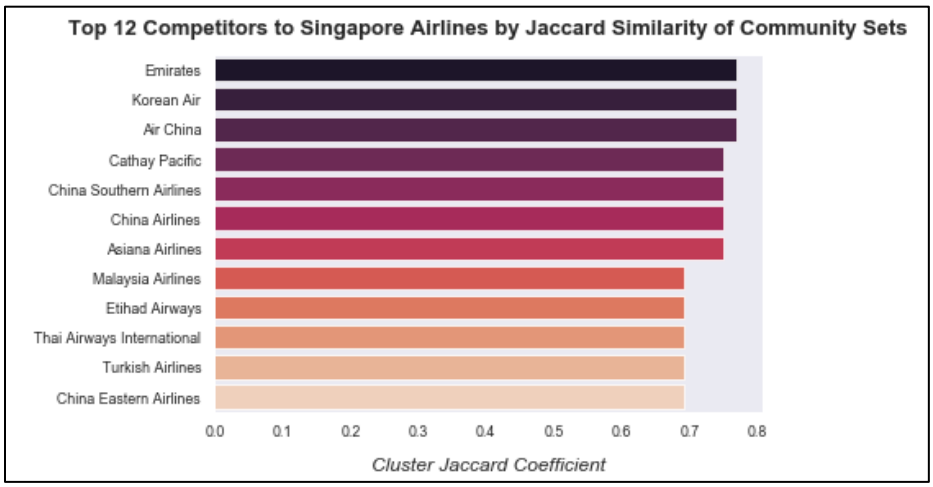


Figure 25. Search for top 12 competitors of SQ via **Jaccard Similarity** of community sets using Python

The above identification of top competitors was relatively consistent with third-party online media sources. The extract of our python script output below shows the community sets available in competitor, but missing from SQ, and vice versa, for potential business exploration:

Name	Unique_airports	No_of_airports	Communities	Communities_no	Competitor_Cluster_Advantage	Subject_Cluster_Advantage	Cluster_Jaccard
Singapore Airlines	[1107, 3885, 3341, 3316, 2006, 2994, 580, 1218...	102	[{4105, 16, 8209, 4130, 4160, 4161, 4162, 4166...	12	[]	[]	1.000000
Air China	[1107, 3370, 3364, 2006, 3406, 6404, 3395, 737...	186	[{4105, 16, 8209, 4130, 4160, 4161, 4162, 4166...	11	[{6372, 6373, 6374, 7558, 6795, 7470, 3380}]	[{9888, 9889, 3250, 3886}, {3240, 3241, 3243, ...	0.769231
Emirates	[253, 248, 2188, 1107, 3341, 2006, 3320, 3339,...	135	[{4105, 16, 8209, 4130, 4160, 4161, 4162, 4166...	11	[{6289, 4217, 6337, 6238}]	[{9888, 9889, 3250, 3886}, {3240, 3241, 3243, ...	0.769231
Korean Air	[2006, 3930, 580, 2340, 3682, 2560, 3797, 2650...	118	[{2006, 2007, 2009, 2011, 2012, 2014, 2015, 20...	11	[{6372, 6373, 6374, 7558, 6795, 7470, 3380}]	[{9888, 9889, 3250, 3886}, {3240, 3241, 3243, ...	0.769231
Asiana Airlines	[1107, 4059, 2006, 2279, 2908, 3930, 3682, 379...	105	[{4105, 16, 8209, 4130, 4160, 4161, 4162, 4166...	9	[]	[{2560, 2562, 2564, 2569, 2570, 2572, 2575, 25...	0.750000

Figure 26. Extract of SQ's competitor list with community differences from our Python script. Red boxed section represents the differences in community sets and their component "airport_ids"

Airline management could also consider airlines with the least Jaccard Similarity in community sets and hence the least conflict of interest for flight partnership.

3.4.3 Detection of key airline competitions based on intra-community interaction

Domestic airline's network is usually confined within a community and hence would likely evaluate key competitors within a community and the integrity or resilience of its flight network. Similarly, citizens from large countries such as US and China would often fly domestically or within a community and hence would want to know the airline with the best connectivity, so that they can best benefit from the frequent travel rewards.

For best analysis, we selected a **large community (yellow-coloured), America**, from our earlier algorithm. See mapping of full domestic routes within the community below:

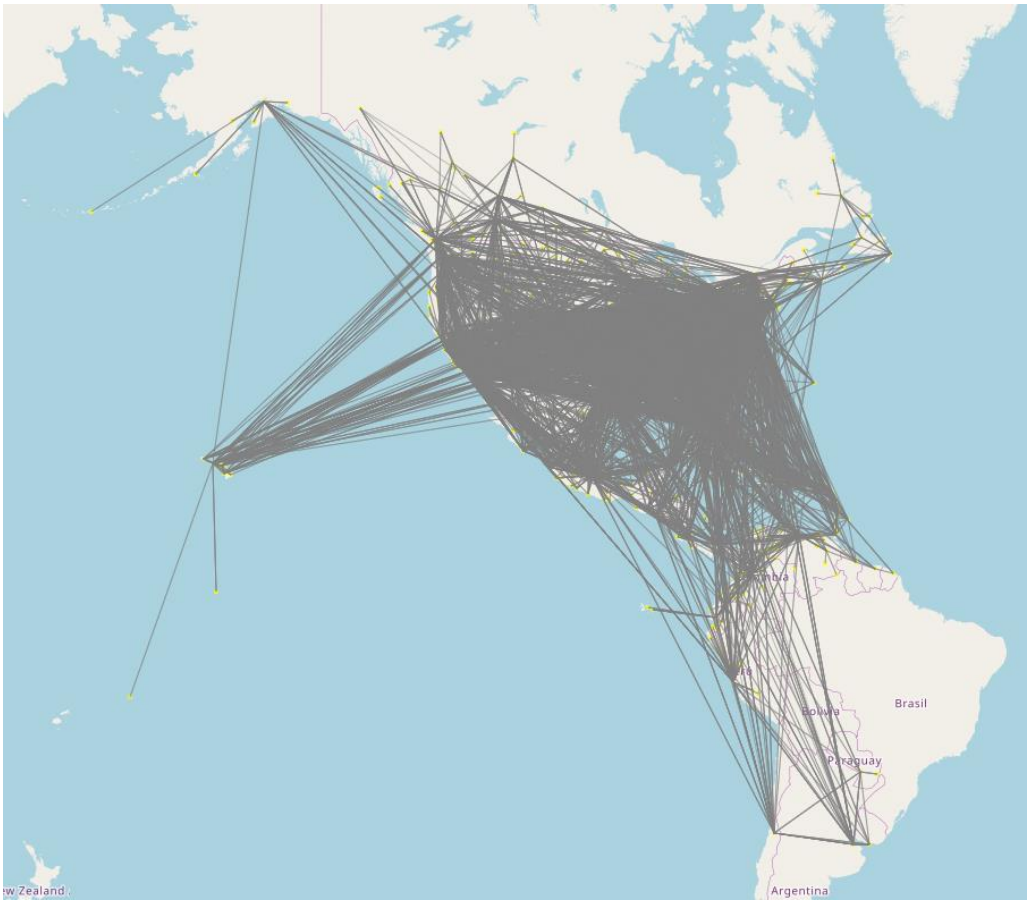


Figure 27. Map of all existing airlines' routes within the yellow-coloured community or "America" detected by Label Propagation algorithm. This domestic graph excludes routes going out and into this community.

To evaluate the connectivity of airline network with respect to the concerned community, we studied two key measures and combined them for collective use:

1) Strongly Connected Component (SCC) Weighted Average Score

Objective:
To measure the number of visited airports, adjusted by the risk of disconnection within a directed airline network in that community (Example: there are no return trips back to the originating airport).

Measurement and rationale:
The number of nodes in each SCC component is weighted averaged relative to the entire airline network. See example of measurement below:

Example: There are 4 SCC components within the airline network {99 airports, 1 airport, 1 airport, 1 airport}. There is a total of 102 airports in the network. The weighted average score is calculated as [(99/102)*99 + (1/102)*1 + (1/102)*1 + (1/102)*1] = **96.118.**

The intention is to penalize multiple SCC components with almost equal number of nodes/airports. An airline's SCC configuration, such as {30 airports, 50 airports, 22 airports} with the same number of visited airports, would have a worser score of **38.08**. Such configuration may be undesirable as the airline risk losing return trip revenue from customer travelling from any one component to another. The equally large size of each component further amplifies such likelihood and losses. While code-sharing with partner airlines could resolve this, it may be risky to depend on third parties for servicing such major chokepoint between two large components/market.

Key Procedures in Python:
- We first extract only the routes of each airlines in the target community (here: America)
- Build a directed weighted (distance) graph network for each airline
- Measure the SCC weighted average score for each airline as described above

Key observations:
The top 12 airlines with the highest SCC weighted average scores were identified as follows. This result is also reasonably consistent with our perception of key market players in that community,

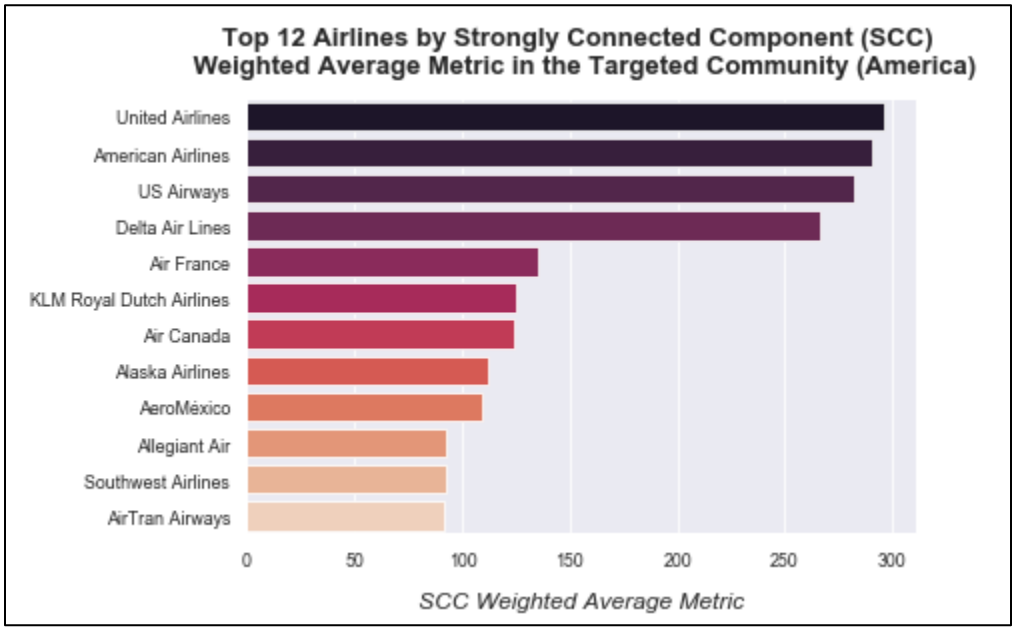


Figure 28. Top 12 Airlines by SCC Weighted Average Metric in Community (America)

Other observations:
Successful players such as Singapore Airlines (International) and US Airways (US Domestic) may have small number of SCC components or at least one big SCC component with other insignificant ones. Through python, we noted that their SCC configuration is {99, 1, 1, 1} and {333, 1, 1, 1, 1 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1} respectively with the smaller loose ends (i.e. 1, 1, 1…) possibly resolved through code-sharing. This is consistent with our rationale for the formulating the SCC Weighted Average Metric.

While this SCC score consider the number of airports and strongly connectiveness, it does not consider the extent of route coverage which will be addressed by the GED score below.

2) Graph Edit Distance (GED) Similarity Score

Objective:
To measure the similarity of the airline's network (consisting of only routes within that community) with the community's flight graph (consisting of all airlines' routes in that community).

Measurement and rationale:
This indicates to management the extent of the airline's connectivity relative to the prevailing market offer (routes) in that community. It represents convenience for the customer. Using GED, we measure the cost of transforming the airline's domestic network into the community

(America) network. The higher the cost, the less similar and hence relatively less connected is the airline's network with respect to the community.

Key Procedures in Python:

- We first extract only the routes of each airlines in the target community (here: America)
- Build a directed weighted (distance) graph network for each airline
- Run an optimized version of GED or **Gmatch4py** from *https://github.com/Jacobe2169/GMatch4py.git*. **Gmatch4py** runs with both Python and Cython, provide C-like performance to this algorithm. *The GED algorithms in Networkx library for Python were too slow for our purpose of comparing many large graphs.*
- We extract the similarity values for ranking as noted below.

Key observations:

There are changes to the top players as compared to the previous SCC metric as this measure takes the similarity of routes into consideration, not only the nodes or airport.

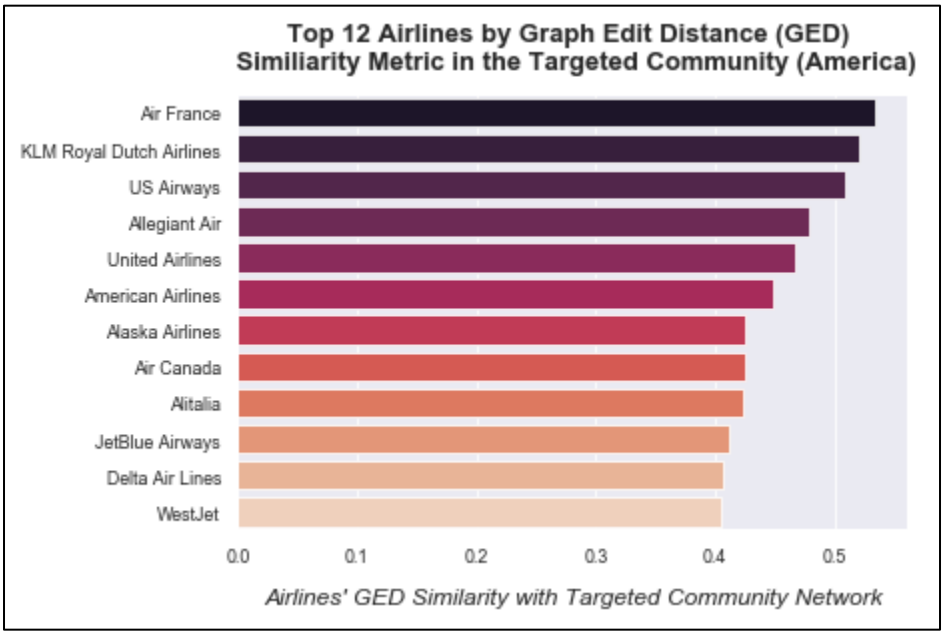


Figure 29. Top 12 Airlines by GED Similarity with the Community (America)

3) Correlation between SCC Weighted Average Score and GED Similarity Metric

While there is similarity in content measured by both the SCC and GED score (both consider the number of nodes or airports), there are some aspects not measured by either one (i.e. SCC penalizes heavily for any disconnection while GED looks at the routes). These variances give purpose for combining both metrics for more complete evaluation. A near perfect correlation of both metric results is undesirable as the combination will not result in diversification or improvement in evaluation.

We expect a generally positive correlation as more nodes (measured by both SCC and GED) comes with more routes (measured by GED). Various noticeable divergences, due to some measurement differences mentioned above, can be leveraged through combining both metrics. This was proven in our correlation analysis below:

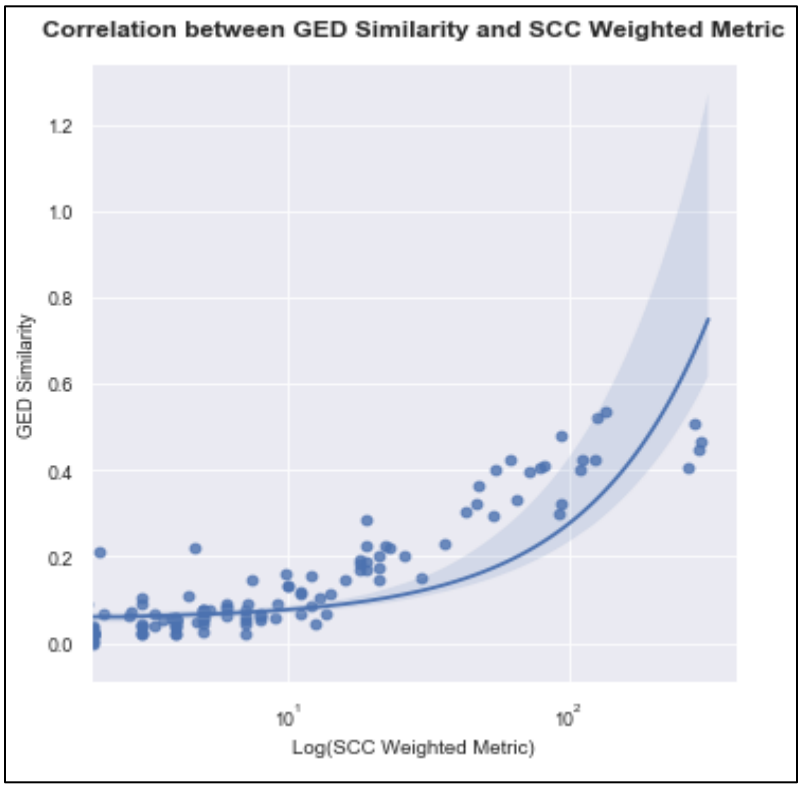


Figure 30. Correlation between GED and SCC results

4) Combination of SCC Weighted Average Score and GED Similarity Metric

Objectives:
Combine both scores for more comprehensive evaluation.

Measurement and rationale:
Both scores are combined for a specific airline in the following manner so that differences in measurement unit will not affect final score:

(SCC Score of the Airline / Maximum SCC Score achieved by all airlines in that Community) + (GED Similarity Score of the Airline / Maximum GED Score achieved by all airlines in that Community)

Key procedures:
We computed the balanced scores for each airline through Python and ranked them as shown in results below.

Key observations:

Airlines are ranked from best (US Airways) to lower scored ones (WestJet) based on the combined score:

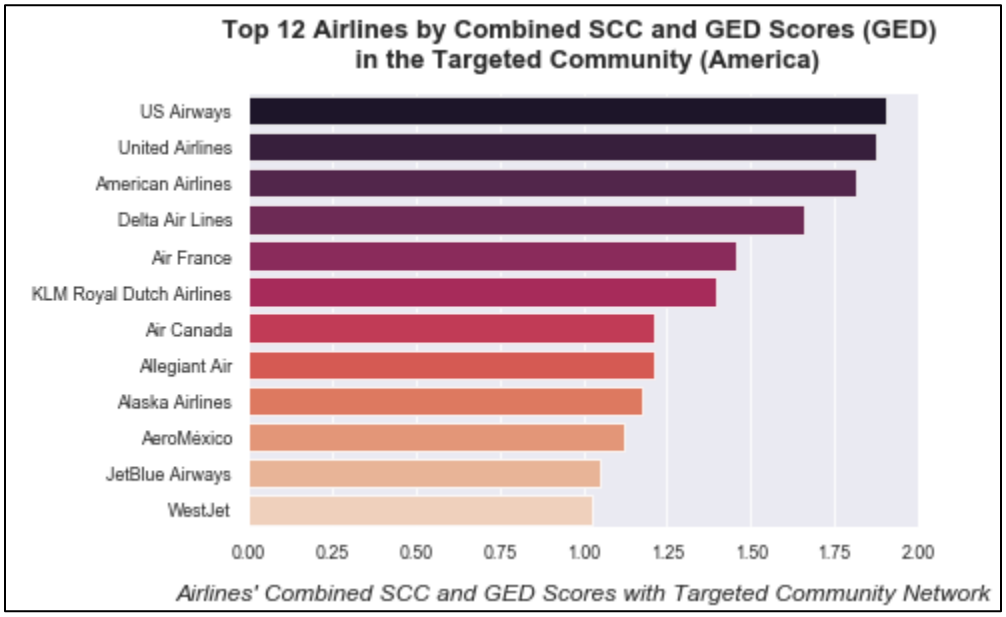


Figure 31. Combined GED and SCC score results (Top 12 Airlines in America)

Below shows the breakdown by SCC and GED scores for the above top 12 airlines:

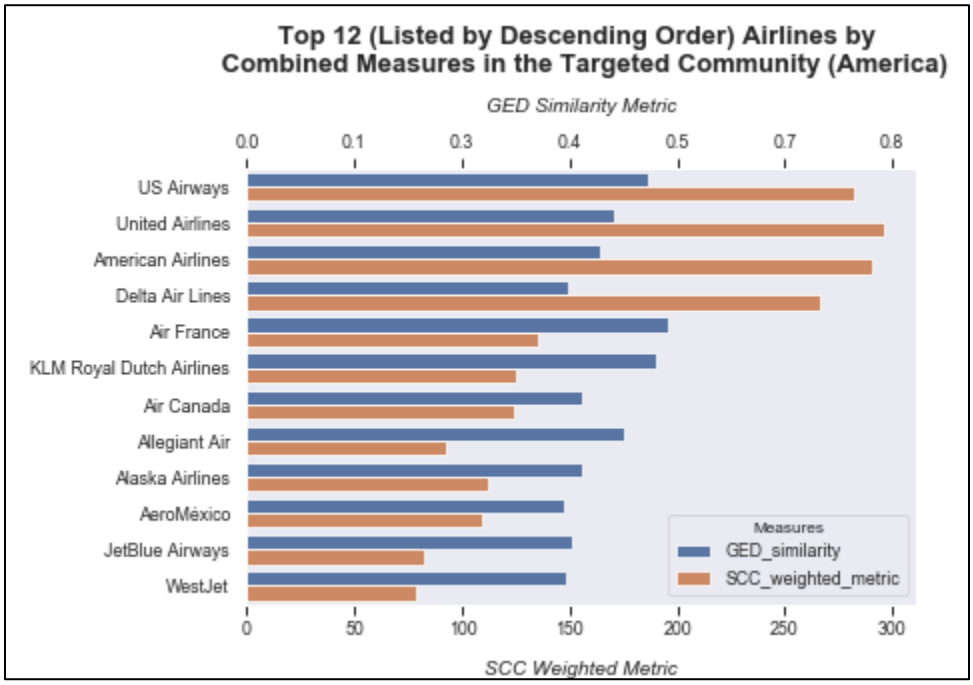


Figure 32. Breakdown of GED and SCC score results (Top 12 Airlines in America)

To appreciate the differences between the 1st ranked airline and the 12th ranked airline in the above list and to validate our scoring quality, both networks can be visualized and compared on the community maps as shown below:

Figure 33. Domestic Network of US Airways (1st place) in Community (America)

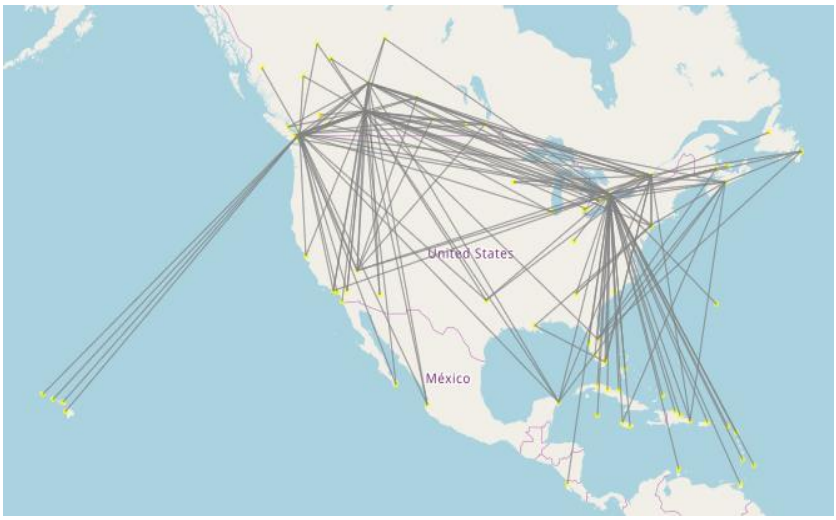


Figure 34. Domestic Network of WestJet (12th place) in Community (America)

The 1st ranked airline, US Airways, network was certainly more connected and wide-spanned in the community (America) as compared to the 12th ranked airline, WestJet.

4. RECOMMENDATIONS

Improvement to the existing regions clustering can be made by finding ways to merge neighboring communities. Using communities generated via label propagation illustrated in Figure as an example, neighboring communities adjacent to brown region may be merged to form a more defining cluster for the Africa continent. Clique merging algorithms together with using airport to airport distance as network weights can be explored for this purpose. Hence, subgraphs produced will involve more datasets and may be more representative of global regions. Key player extracted from these subgraphs would be more comprehensive as well.

Obtaining more detailed code-sharing routes data, along with parent-subsidiary relationship (for example, Tiger Airways is a subsidiary of Singapore Airlines and will focus on regional and international flight respectively), will allow a more accurate review of an airline's flight network communities. Otherwise, the disconnection among the strongly connected components of the airline network, as detected in our analysis, could not be fully addressed in this report.

Flight route network could be dynamic over time, changing over travel seasons and this may affect the identification of key airport and airline players. Temporal network analysis can be conducted over long historical period to review trends of changing centrality and network growth. This could be useful for airline management planning of routes and projection of capital investments. Unfortunately, temporal data was not present in the dataset.

5. CONCLUSION

In this report, we explore the world flight routing network to identify potential communities among airports and their key airport players. We also studied the various airlines' interaction with our identified airport communities from an inter-community and intra-community perspective.

Firstly, we noted that the flight graph follows a scale-free network, that consist of several air hubs located in different regions.

We then used several well-known autonomous community detection algorithms, namely Clauset-Newman-Moore modularity maximization (CNM), Louvain best partition and Label Propagation to identify geographic regions based on flight routes data. We also attempted spectral clustering method and semi-supervised method through establishing ground truths for the k-means procedure within the spectral clustering method. Ultimately, the Label Propagation algorithm provided the best quality clustering scores and visualization of communities.

Key airport players for each community detected by the Label Propagation were then identified through various popular centrality measures: Degree Centrality, Betweenness Centrality and VoteRank.

We then investigated into how businesses can exploit the airline network interaction with our established communities to identify the degree of competition among airlines. The simplest similarity measure, Jaccard Coefficient, was used to compare similarity between community sets visited by each airline, as means to evaluate competitiveness. Apart from inter-community interaction, we also looked specifically at airlines' connectivity at intra-community level. We applied various connectivity measures: (1) Strongly Connected Component Weighted Average Score to identify disconnection in airline's network within the community, (2) Graph-Edit Distance to measure similarity of the airline's domestic network versus the community network, and lastly, integrated both measures to identify some of the best connected airlines within the community.

6. APPENDIX

Clauset-Newman-Moore Modularity maximization							
sub_graph	**0**	**1**	**2**	**3**	**4**	**5**	**6**
Number of nodes	900	874	719	145	131	92	43
Number of edges	4686	4367	6876	292	192	133	69
Density (%)	1.158	1.145	2.664	2.797	2.255	3.177	7.641
Average clustering coefficient	0.5	0.539	0.465	0.487	0.44	0.297	0.526
Performance	0.796						
Coverage	0.904						

Label propogation							
sub_graph	0	1	2	3	4	5	6
Number of nodes	697	628	372	281	145	97	94
Number of edges	4091	6215	2473	1096	535	333	208
Density (%)	1.687	3.157	3.584	2.786	5.125	7.152	4.759
Average clustering coefficient	0.488	0.455	0.551	0.527	0.515	0.432	0.483
Performance	0.892						
Coverage	0.852						

Louvain best partition							
sub_graph	**0**	**1**	**2**	**3**	**4**	**5**	**6**
Number of nodes	733	511	490	463	238	211	159
Number of edges	4086	5323	2830	1745	494	575	676
Density (%)	1.523	4.085	2.362	1.632	1.752	2.595	5.382
Average clustering coefficient	0.509	0.44	0.531	0.549	0.495	0.432	0.505
Performance	0.866						
Coverage	0.867						

Network measures from community detection algorithms