MSBD5008 Introduction to Social Computing

Air Transportation Network Analysis: A GNN-based Approach for Airport Classification and Route Pattern Clustering

LAKHANI Sunil Harsh	LI Ka Ho	TSANG Kai Ho	SIMPSON, Patrick
20910249	20922151	20905476	20820028
hslakhani@connect.ust.hk	khlibg@connect.ust.hk	khtsangak@connect.ust.hk	psimpson@connect.ust.hk

Abstract

In this project, we analyze flight route networks and address the potential application of graph neural networks (GNNs) in performing airport classification and clustering. Our dataset consists of publicly available data on airports and flight routes, which we transform into airports as nodes and flight routes as edges. Along with performing the rudimentary tasks of network visualization and analysis, we perform two major tasks which can be applied to airport or airline operational optimization. Firstly, we identify top global airports based on a mix of centrality measures alongside further visualization, and secondly, conduct a meta-analysis on the airline's sub-graphs which in principle can be leveraged to perform competitor analysis. We then implement node clustering, where we analyze the similarity between airports based on their route connections portrayed, which enables us to recommend routes for airports based on their centrality measures. Lastly, we perform node classification by categorizing airports based on their size according to the number of runways they consist of. For node classification, we employ a 3-layered GNN, achieving an accuracy of over 80%. For node clustering, we leverage a 2-layered GNN as well as K-means. Our study showcases the potential of graph-based approaches in understanding airport networks and provides insights for future research in link prediction for new airports.

1. Introduction

Air transportation is an essential aspect of modern society, providing economic benefits and connecting people across the globe. Understanding airport networks and flight routes is crucial for the aviation industry, tourism boards, and travel agencies to improve operations and tailor their strategies. Enhancement and optimization of these networks can also help reduce air transport's carbon footprint. In this project, we perform network analysis methods to understand airport networks and leverage graph neural networks (GNNs) to perform airport classification and clustering.

We leveraged publicly available datasets to curate a dataset containing 3425 nodes and 37595 edges along with features including geographical coordinates and number of runways, which helps us classify airports based on their size (small, medium, or large). Performing node classification enables us to compare the operational efficiency and profitability of airports and offers insights for market targeting strategies. For node clustering, we investigate similarities between airports based on their route connections. By clustering airports with similar route patterns, we can gain insights into the relationships between airports and compare centrality measures across clusters, after which, we are able to recommend routes for airports to optimize their operational efficiency.

To perform the classification and clustering tasks, we leverage graph neural networks, which are well-suited for complex network analysis. Specifically, we employ a 3-layered GNN for node classification, achieving an accuracy of over 80%, and a 2-layered GNN as well as K-means for clustering. Our results demonstrate the potential of graph-based approaches for understanding airport networks and provide a foundation for future research in link prediction for new airports, temporal analysis, and the investigation of multimodal transportation networks.

2. Data

2.1. Datasets

We initially obtained a dataset for flight routes on Kaggle. In order to incorporate more information about each airline, we found two more datasets from OurAirports. The data used for this project is all publicly available and is summarized in Table 1.

Table 1: Dataset Summary

Dataset	URL	No. of entities	No. of attributes
Routes	https://www.kaggle.com/datasets/open-flights/flight-route-database	62,825	9
Airports	https://ourairports.com/data/	75,006	18
Runways	https://ourairports.com/data/	44,723	21

2.2. Data Preprocessing

To prepare a clean dataset for our project, we performed exploratory data analysis to expose erroneous and missing data and used data manipulation to remove the problematic entities as well as redundant attributes. Key attributes used in our project are listed in Table 2.

Table 2: Description of key attributes

Attribute	Description
source_airport_iata	Source airport code given by IATA
source_airport_id	Source airport ID
destination_airport_iata	Destination airport code given by IATA
destination_airport_id	Destination airport ID
type	Airport size {large, medium, small}
latitude_deg	Latitude of the airport
longitutde_deg	Longitude of the airport

Since the airport IDs are not numbered consecutively and some of them are missing, we renumbered the airport IDs from 0 to 3424 for ease of performing network analysis and for a less sparse adjacency matrix. Furthermore, we combined multiple datasets to create the desired dataset used for specific tasks.

2.3. Exploratory Data Analysis

Ultimately our dataset is transformed into a graph, so the initial exploratory data analysis (EDA) is limited to outlier detection, attribute name cleaning and providing an early understanding of the distribution of the data. One important outcome of the EDA was identifying and merging many niche airport types (including heliports, seaplane ports and baloonports) into the ‘small airport’ type.

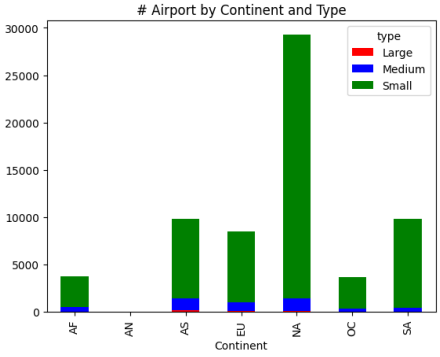

Figure 1: Airport Type Histogram

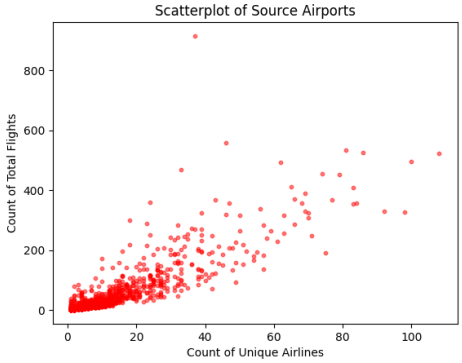

Figure 2: Source Airport Scatterplot

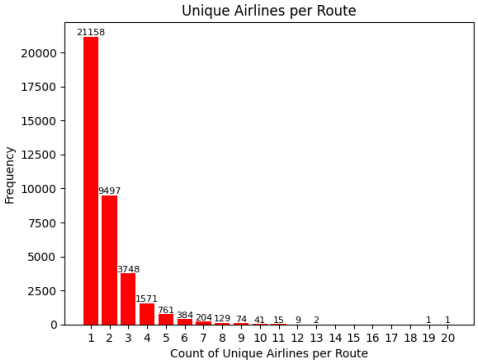

Figure 3: Unique Airlines per Route

The charts above are an example of some of the analysis conducted. They show us the distribution of airports by region and type (Figure 1), clustering of airports based on activity (Figure 2) and the dispersion of airlines across routes (Figure 3). We conducted further in-depth exploration of the data once it has been converted to a graph in the next section of this report.

3. Network Analysis

3.1. Degree Distribution

Considering the importance of the direction of an air transportation network (ATN), we treat our data as a directed graph, allowing us to distinguish between incoming and outgoing airlines. In particular, a flight flying from an airport to another airport does not necessarily imply that there is a route to return, directly or indirectly. For this reason, in-degree and out-degree of our network was studied, respectively.

Our ATN is composed of 3,425 nodes and 37,595 unique edges, where a node indicates an airport and an edge indicates a route. The average degree of the network $\bar{k}$ is defined as the sum of the average in-degree $\overline{k^{in}}$ and the average out-degree $\overline{k^{out}}$ of the network. In particular,

$$\overline{k^{in}} = \frac{E}{N} = \overline{k^{out}}$$

where E and N denotes the number of edges and the number of nodes, respectively. Hence, we found that $\overline{k^{in}}$ and $\overline{k^{out}}$ are 10.98 and $\bar{k}$is 21.95.

To further study the properties of a directed graph, we identified the airports that are sources, i.e., $k_i^{in} = 0$ and sinks, i.e. $k_i^{out} = 0$ in Table 3 and Table 4, respectively. Seven airports belong to sources while 16 airports belong to sinks.

In addition, we identified the airports with the highest in-degree and out-degree to determine which airports have the most airlines heading to and departing from them. The top five airports for both in-degree and out-degree are listed in Table 5, we noticed that they are the same airports. This observation suggests that these airports are probably the hubs of the air transportation network. Busy airports is a key consideration in our project, as previous studies have found that the top 25

airports in the world accounted for more than half of global air traffic, while the vast majority of airports had very few flights [1].

Table 3: Airports are sources

IATA	Airport Name
LJA	Lodja Airport
IUE	Niue International Airport
STZ	Santa Terezinha Airport
SXX	São Félix do Xingu airport
PTJ	Portland Airport
VDA	Ovda Airport
MSW	Massawa International Airport

Table 4: Airports are sinks

IATA	Airport Name
KZB	Zachar Bay Seaplane Base
KYK	Karluk Airport
KPR	Port Williams Seaplane Base
CZJ	Corazón de Jesús Airport
SPI	Abraham Lincoln Capital Airport
FMI	Kalemie Airport
TUA	Lieutenant Colonel Luis A. Mantilla International Airport
QFX	Igaliku Heliport
KZI	Kozani State Airport Filippos
ORX	Oriximiná Airport
BVS	Breves Airport
MTE	Monte Alegre Airport
DLZ	Dalanzadgad Airport
UII	Utila Airport
CMP	Campo Alegre
BSS	Balsas Airport

Table 5: Top five airports for in-degree and out-degree

Rank	Airport Name	In-degree	Out-degree
1	Frankfurt Airport	238	239
2	Charles de Gaulle International Airport	233	237
3	Amsterdam Airport Schiphol	231	232
4	Istanbul Airport	230	227
5	Hartsfield Jackson Atlanta International Airport	216	217

The degree distribution is a simple metric that provides insight into a network's structure. To better understand the structure of our network, we plotted in-degree and out-degree distributions on both linear-linear and log-log scales in Figure 1. As most airports have relatively small degrees, we found that a log-log scale was more appropriate as almost all points in the linear-linear plots lie on the x and y-axes. The heavy-tailed distribution of both in and out-degrees is evident, with the slope of the log-log plots indicating the presence of a power-law in the degree distribution of our network.

To estimate the power-law exponent α, we plotted the Complementary Cumulative Distribution Function (CCDF) for the in and out-degrees on a log-log scale. By minimizing the Kolmogorov-Smirnov distance between our empirical data and the fitted power-law [2], we found the minimum value of the power-law distribution x_{min} to be equal to 2 and the exponents for in-degree and out-degree to be 1.875 and 1.878, respectively.

The typical range of scaling exponents for power-law distributions in networks is between 2 and 3. However, the exponents we found fall outside of this range. The smaller scaling exponent indicates that the proportion of moderately connected nodes in our network may be larger than in typical scale-free networks, resulting in a slower decay rate in the degree distribution of our network. One possible explanation for this deviation is that the air transportation network is a hybrid of different types of networks, each with its own degree distribution. For example, regional

hubs may have a high degree of connectivity within their respective regions, but relatively few connections to other regions.

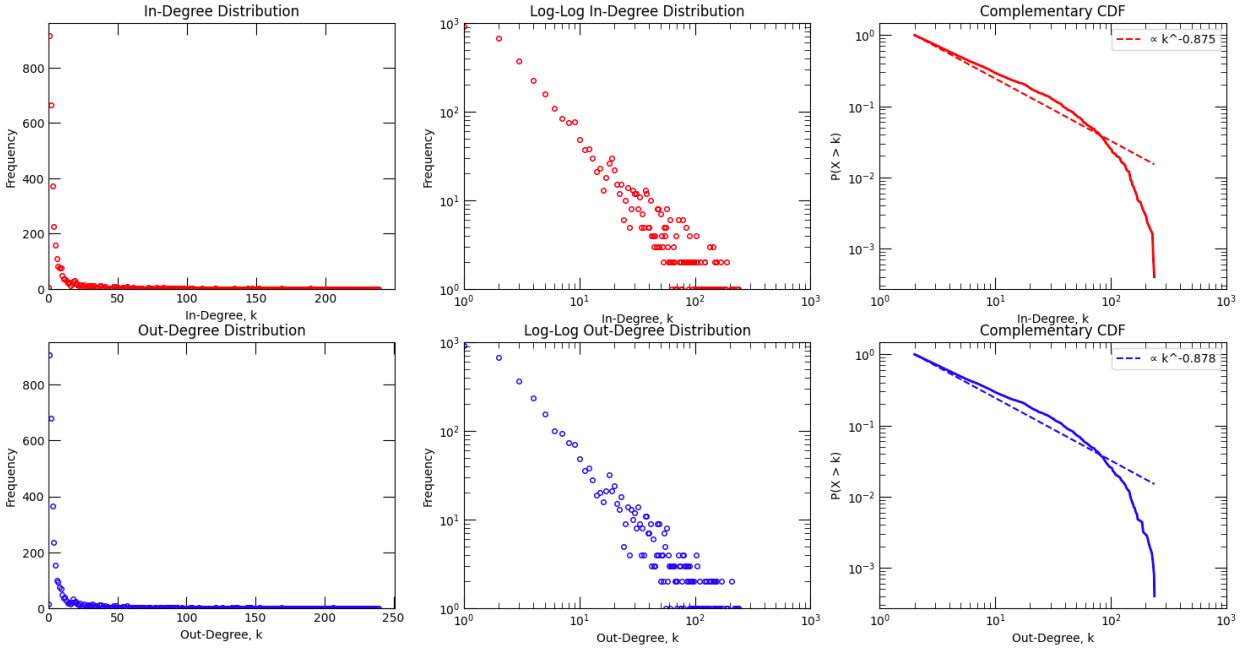

Figure 4: In- and out-degree distribution with a power-law

3.2. Network Diameter & Average Path Length

The diameter of a network is the maximum shortest path between any pair of nodes in the network, and it provides a measure of the connectivity of a network. A smaller diameter means that the network is better connected and easier to traverse.

To calculate the diameter of our network, we defined h_{ij} as the shortest path from node i to node j. The diameter of the network is then given by

$$diameter = max(h_{ij})$$

Additionally, we computed the average shortest path length, which is given by

$$\bar{h} = \frac{1}{2E} \sum_{i,j \neq i} h_{ij}$$

Figure 3 illustrates that the diameter of our network is only 14, indicating that any two airports can be reached within 14 hops. Furthermore, the majority of airports can be reached within four hops. These findings suggest that the air transportation network is well-connected, making it easy to travel between different destinations.

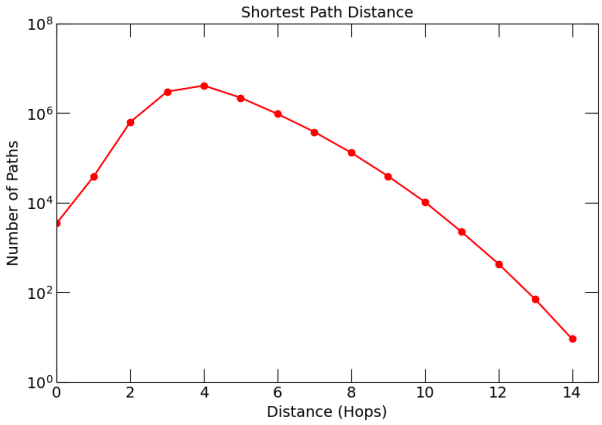

Figure 5: Shortest path distance in ATN

3.3. Connected Components

We analyzed the connectivity of our network by plotting the sizes of its strongly connected components (SCCs) and weakly connected components (WCCs) in Figure 4. We found 44 SCCs, of which 35 were singletons, consistent with the presence of sources and sinks identified in Section 4.1. The largest SCC contained 3,354 nodes, while the rest had at most 8 nodes. For the WCCs, we observed only 8 components, as a weakly connected directed graph is essentially an undirected graph with more accessible nodes. The largest WCC also contained the majority of nodes. Table 6 provides additional details on these results.

Table 6: Distribution of the size of the connected components

Size of SCCs	Frequency	Size of WCCs	Frequency
1	35	2	3
2	3	4	3
4	3	10	1
8	1	3397	1
10	1		
3354	1		

Our analysis reveals the existence of a giant component in ATN, represented by the dominating SCC. This giant component likely comprises the busiest airports with many flight connections, making it a hub. Conversely, the smaller SCCs could represent airports with low traffic or those that are geographically isolated from the main network.

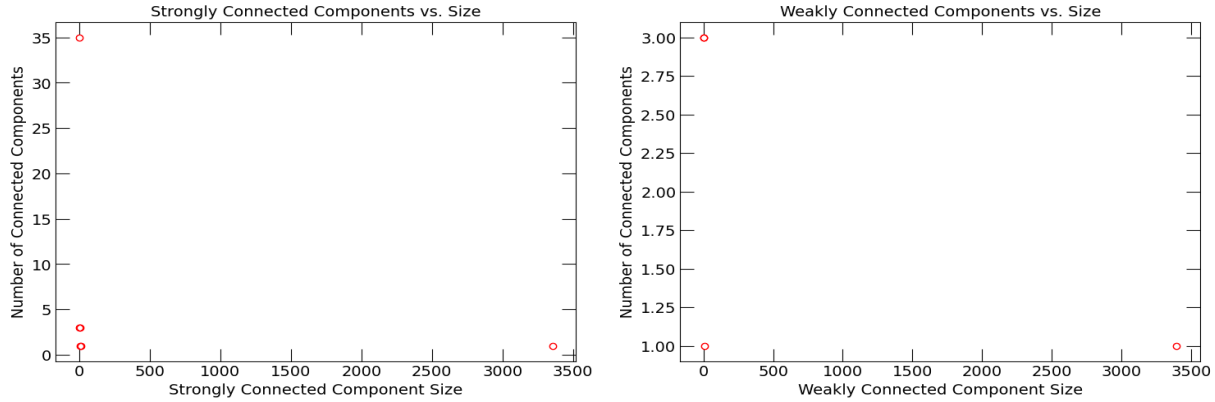

Figure 6: Strongly and weakly connected components of ATN

3.4. Clustering Coefficient

Clustering coefficient is a crucial metric of the extent to which nodes in a network tend to form groups or communities [3], offering insights into the structure and density of the communities. In our study, we analyzed the degree to which airports tend to be connected to one another by computing their clustering coefficient.

The local clustering coefficient C_i of a node in a directed graph measures how connected its neighbors are, as introduced by Watts and Strogatz in 1998 [3]. This coefficient can be calculated

$$C_i = \frac{\left|e_{jk} : v_j, v_k \in N_i, e_{jk} \in E\right|}{k_i(k_i - 1)}$$

where k_i is the number of neighbors a node i has, and N_i is the set of its neighbours. In a directed graph, e_{jk} does not necessarily imply e_{kj}, so the maximum number of edges a node i can have is $k_i(k_i - 1)$, instead of $\frac{k_i(k_i-1)}{2}$ as in an undirected graph.

Upon computing the local clustering coefficient for each airport in our network, we observed that more than one-fifth of the airports (708 nodes) have a clustering coefficient of 1, indicating that all their adjacent airports are fully connected and can communicate with each other directly. Furthermore, more than one-fourth of the airports (984 nodes) have a clustering coefficient of 0, indicating that none of their adjacent airports are connected to each other. This underlines the importance of airports with clustering coefficients of 0 as they act as intermediaries for communication between airports that are not directly connected.

Additionally, we investigated the global behavior of the clustering coefficient and plotted the relationship between the clustering coefficient and the degree of our network in Figure 5. We observed that while the clustering coefficient decreases with the size of the network, our network exhibits high clustering at low degrees and low clustering at high degrees. To quantify this finding, we calculated the global clustering coefficient $\bar{C}$ by averaging the local clustering coefficient of all airports [3],

$$\bar{C} = \frac{1}{n}\sum_{i=1}^{n} C_i$$

obtaining 0.469.

The detection of a high average clustering coefficient and a short average path length in Section 3.2 suggests that the air transportation network exhibits small-world characteristics.

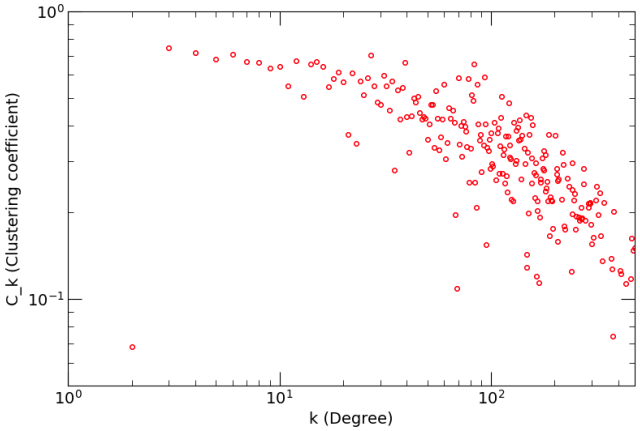

Figure 7: Clustering coefficient against degree of ATN

3.5. Centrality

To determine the most significant airports in our network, we evaluated five different centrality measures, namely degree centrality, betweenness centrality, closeness centrality, eigenvector centrality, and PageRank centrality. As our network is directed, we utilized the notion of prestige in computing the centrality scores. We present the top ten airports based on each centrality measure in Table 7 to facilitate comparison and analysis. By using multiple centrality measures, we can gain a more comprehensive understanding of the most important airports in the air transportation network, as each measure emphasizes different aspects of network centrality, which is discussed in the following sections.

3.5.1. Degree Centrality

Degree centrality measures the number of connections a node has in an undirected graph. However, in directed graphs such as our air transport network, we use in-degree centrality and out-degree centrality to capture the prestige and popularity of nodes [4]. In-degree centrality measures the total number of incoming connections a node receives from other nodes, while out-degree centrality measures the number of outgoing connections a node sends to other nodes. The equations for in-degree and out-degree centrality are given by

$$Centrality_{inDegree,i} = \frac{1}{N-1}\sum_{i=1}^{N-1} k_i^{in}$$

$$Centrality_{outDegree,i} = \frac{1}{N-1}\sum_{i=1}^{N-1} k_i^{out}$$

In the context of our network, an airport with high in-degree centrality receives numerous flights, whereas an airport with high out-degree centrality sends numerous flights. An airport with both high in-degree and out-degree centrality can be seen as a busy hub.

3.5.2. Betweenness Centrality

Betweenness centrality measures the frequency of a node lying on the shortest path between any two nodes in the graph [4]. Given that our network is a directed graph, betweenness prestige was determined by considering geodesics. The computation of betweenness prestige is given by

$$Centrality_{betweenness,i} = \frac{\sum_{j,k} g_{jk}(i)/g_{jk}}{(N-1)(N-2)}$$

where g_{jk} is the paths between j and k that pass through i.

In the context of our network, the significance of an airport's betweenness centrality is reflective of its "controlling power" of the airport over the most efficient path between any two airports. This measure also underlines the importance of an airport's presence, making the pairs of airports accessible. An airport with high betweenness centrality plays a crucial role in the efficient movement of passengers and goods.

3.5.3. Closeness Centrality

Closeness centrality measures how easily a node can reach other nodes in the graph and is computed by the reciprocal of the average shortest path length between a node and all other nodes [5]. Similarly, due to our directed network, we used closeness prestige instead, which only considers nodes within the influence range I_i of the target node i. The formula for closeness centrality is given by

$$Centrality_{closeness,i} = \left(\frac{\sum_j d(i,j))}{|I_i|}\right)^{-1}$$

In the context of our network, a high closeness centrality score for an airport indicates that it is easily accessible to other airports and may not require route optimization. This suggests that the airport be strategically located and could potentially serve as a hub for connecting flights.

3.5.4. Eigenvector Centrality

Eigenvector centrality measures a node's importance based on the centrality of its neighbors. In other words, a node's importance increases if it has connections to other important nodes [4]. It is proven that eigenvector centrality is proportional to the leading eigenvector of the adjacency matrix of a network and can be computed iteratively using the following equation

$$x(t) = A^t x(0)$$

In the context of our network, eigenvector centrality is an interesting measure as it not only indicates the influence of an airport, but also suggests the potential for its neighbors to become important. High eigenvector centrality of an airport can help boost the economic development of less prosperous airports by forming an alliance with the influential airport.

3.5.5. PageRank Centrality

PageRank centrality, considered a variant of the eigenvector centrality, also quantifies the influence of a node based on its connections and contacts [4]. Unlike eigenvector centrality, PageRank centrality is appropriate for directed graphs, especially directed acyclic graphs, since a

node with zero in-degree can still have a non-zero contribution to the importance of its neighbors [5]. However, it is important to note that recent research suggests that eigenvector centrality can achieve a comparable result to PageRank for directed graphs and even outperform it in terms of time complexity [6]. To determine which measure is better suited for our analysis, we computed the PageRank centrality score using the following equation [5]

$$x = \alpha \frac{A}{k^{out}} x + (1 - \alpha) \frac{1}{N}$$

In Table 7, it can be observed that the top 10 airports with the highest eigenvector centrality are distinct from those with the highest PageRank centrality. This finding is inconsistent with the results presented in [6], where the authors discovered that the rankings of web vertices obtained by eigenvector centrality and PageRank are practically the same. In Section 4.1, we however found that only a small proportion (0.2%) of the airports belong to a source, so any negative effect of using the eigenvector centrality may be negligible. Given the different structure and connectivity of various networks, we decided to retain both measures as features of our network.

Table 7: Top-10 airports for each centrality measure

In-degree	Out-degree	Betweenness	Closeness	Eigenvector	PageRank
Frankfurt Airport (0.069509)	Frankfurt Airport (0.069801)	Ted Stevens Anchorage International Airport (0.070204)	Frankfurt Airport (0.392389)	Amsterdam Airport Schiphol (0.165909)	Hartsfield Jackson Atlanta International Airport (0.004666)
Charles de Gaulle International Airport (0.068049)	Charles de Gaulle International Airport (0.069217)	Los Angeles International Airport (0.066164)	Charles de Gaulle International Airport (0.389993)	Frankfurt Airport (0.165748)	Istanbul Airport (0.004388)
Amsterdam Airport Schiphol (0.067465)	Amsterdam Airport Schiphol (0.067757)	Charles de Gaulle International Airport (0.061703)	London Heathrow Airport (0.388306)	Charles de Gaulle International Airport (0.159224)	Chicago O'Hare International Airport (0.004275)
Istanbul Airport (0.067173)	Istanbul Airport (0.066297)	Dubai International Airport (0.059350)	Dubai International Airport (0.384084)	Munich Airport (0.148957)	Denver International Airport (0.004251)
Hartsfield Jackson Atlanta International Airport (0.063084)	Hartsfield Jackson Atlanta International Airport (0.063376)	Frankfurt Airport (0.051000)	Amsterdam Airport Schiphol (0.382932)	London Heathrow Airport (0.137050)	Dallas Fort Worth International Airport (0.004180)
Beijing Capital International Airport (0.060164)	Beijing Capital International Airport (0.060164)	Beijing Capital International Airport (0.049167)	Los Angeles International Airport (0.379042)	Leonardo da Vinci International (0.135864)	Domodedovo International Airport (0.004110)
Chicago O'Hare International Airport (0.059287)	Chicago O'Hare International Airport (0.060164)	Chicago O'Hare International Airport (0.047430)	John F Kennedy International Airport (0.377277)	Istanbul Airport (0.129644)	Charles de Gaulle International Airport (0.003941)
Munich Airport (0.055199)	Munich Airport (0.055783)	Seattle-Tacoma International Airport (0.045268)	Lester B. Pearson International Airport (0.372411)	Josep Tarradellas Barcelona-El Prat Airport (0.129428)	Frankfurt Airport (0.003830)
Domodedovo International Airport (0.055199)	Domodedovo International Airport (0.055199)	Amsterdam Airport Schiphol (0.042658)	Istanbul Airport (0.371370)	Zurich Airport (0.126192)	Beijing Capital International Airport (0.003811)
Dallas Fort Worth International Airport (0.054030)	Dubai International Airport (0.054907)	Lester B. Pearson International Airport (0.042527)	Chicago O'Hare International Airport (0.370376)	Adolfo Suárez Madrid–Barajas Airport (0.123239)	Dubai International Airport (0.003645)

4. Visualization

4.1. Flight Route Network

Network visualization is done via geographical plots, which are initially used to display the entire network, but in later analysis can also be leveraged to show sub-graphs and their relationship within the entire network. The latitude and longitude attribute for each airport was utilized to plot these graphs.

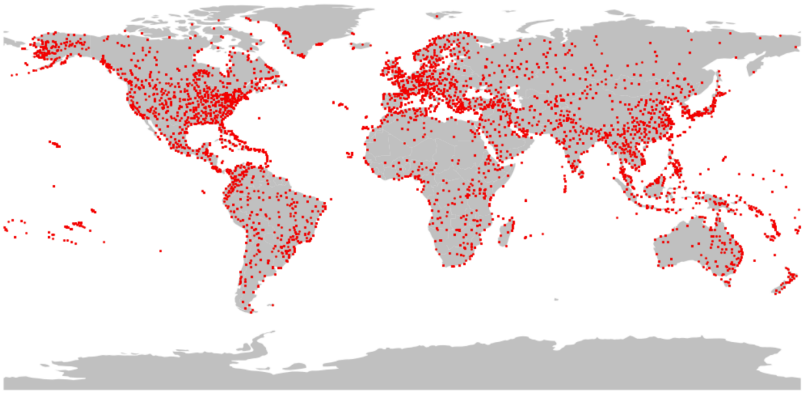

Figure 8: Geo plot of all airport locations

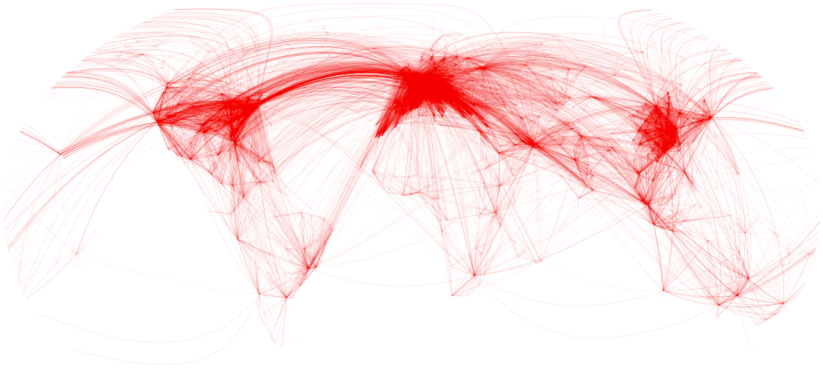

Figure 9: Geo plot of all flight routes

4.2. Small Clusters

There entire global network as a total of 8 connected components:

Size	# Components
2	3
4	3
10	1
3397	1

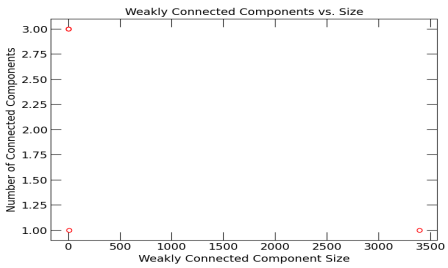

Figure 10: Small Connected Components

We plotted weakly connected components that are of size 10 or less to investigate further:

This is a weakly connected component zoomed in. We suspect that the reason for these 7 small weakly connected components could be because they are located in remote areas and are therefore not very active. Another reason could be that they are extremely small in size such as heliports or private runways, which could explain why they are used less frequently and not by that many passengers which is why they are not connected to the giant connected component.

5. Airline Applications

Flight route and airport data, even in a public domain form such as the model in the scope of this project can be hugely valuable for airline operations and strategy planning. This section aims to explore the possibility of optimization of airlines via an applied network analysis problem.

5.1. Identifying Hubs

Firstly, using the directed graph as previously constructed, we perform an analysis on all global airports with a view to uncover the statistically most important airports in the world as described as 'hubs'. Centrality methods including Degree Centrality, Betweenness Centrality and PageRank methods are employed to define the top airports. For demonstration purposes, we limit the results to the top 10 global airports (or 0.29% of the total in our data set).

Table 8: Top 10 Airport Hubs

IATA	Name	Degree Centrality	Betweenness Centrality	PageRank
CDG	Charles de Gaulle International Airport	0.1373	0.0617	0.0039
FRA	Frankfurt Airport	0.1393	0.0510	0.0038
IST	Istanbul Airport	0.1335	0.0412	0.0044
ORD	Chicago O'Hare International Airport	0.1195	0.0474	0.0043
PEK	Beijing Capital International Airport	0.1203	0.0492	0.0038
AMS	Amsterdam Airport Schiphol	0.1352	0.0427	0.0036
ATL	Hartsfield Jackson Atlanta International Airport	0.1265	0.0294	0.0047
DXB	Dubai International Airport	0.1081	0.0594	0.0036
DME	Domodedovo International Airport	0.1104	0.0294	0.0041
LAX	Los Angeles International Airport	0.0867	0.0662	0.0033

The top 10 list of hubs is compiled via a summation of the ranks for each airport in each centrality measure, then ranking this result in ascending order. The resulting hubs do have a feature in common in that they are all large airports. Each global region is also represented in this list, with 6 located in Europe (including Moscow & Istanbul). From this list, we are able to see that these

highly connected airports are not all connected in the same ways, as some of the hubs have vastly different metrics across the measures.

Nodes with high degree are important in a variety of contexts. For example, in a communication network, the nodes with high degree may be the busiest email addresses or the servers that route the most traffic [7]. This analogy can be directly translated to our air traffic network, whereby an airport with high degree centrality would suggest that it has a large number of connections to other airports in the network.

Betweenness centrality measures how often a node (in this case, an airport) serves as a bridge or intermediary between other nodes in the network and is a well-established measure for quantifying the influence of vertices on the connectivity of a graph. The betweenness centrality of a vertex v is defined as the number of shortest paths between all pairs of vertices in the graph that pass-through v [8]. If an airport has low betweenness centrality, it suggests that it is not as critical in connecting different destinations within the network.

On the other hand, PageRank is a measure of the importance or influence of a node based on the number and quality of links (in this case, flights) that connect to it. PageRank uses an iterative algorithm to compute these scores, with the idea that a page's importance can be approximated by the sum of the importance of the pages that link to it [9]. Again, we can apply an analogy of the PageRank web model with our flight route network. PageRank can highlight the importance of an airport as a hub based on the importance of the airports adjacent to it also.

The most interesting hub of the top 10 results here is ATL (Hartsfield Jackson Atlanta International Airport), in Georgia, USA. This airport has the lowest betweenness centrality amongst its peers in this top 10 list, however has the highest PageRank. Anecdotal, the reasoning behind this would be twofold; strong market position due to direct connections to many other popular destinations and limited connecting traffic whereby the airport has low betweenness centrality as it does not serve as a critical connecting point, however dies have many direct flights connecting it to other directions. In viewing the ATL case specifically to see if these assumptions hold true, we inspect the airport geographically. The airport is located in the southeast of the USA, which hosts 2 other hubs within this top 10 list. ATL is connected to these other airports (ORD, LAX), which contributes to a high PageRank score as there are two other airport hubs with many edges to ATL. This also contributes to a high betweenness centrality in the context of the overall network, but a metric that is lowest when compared to its other peers due to the other two hubs appearing in a higher proportion of global shortest paths. The other contributing factor to low betweenness centrality is that many routes in and out of ATL do not connect to other hubs, but smaller airports that may have less routes outside of the reciprocal one with ATL.

We plot the edges from these nodes to visualize the coverage granted by these hubs globally. Figure 11 shows the in and out edges from the hub nodes in red, with the remaining routes in the graph plotted in the background in gray:

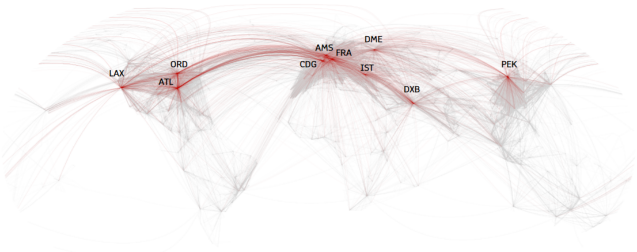

Figure 11: Top 10 Airport Hubs geo-plotted

Further insight can be gained here as we can see that ATL has many edges to Central and South America when compared to the other two hubs in the USA. These flights are the contributors to low edge betweenness as many of these could be vacation destinations or small airports, which have limited connection to the rest of South America, which we see has a faint imprint in the overall global flight route network when compared to the Likes of North America, Europe and North Asia.

5.2. Airline Reachability

Secondly, from a competitive standpoint, the reachability of an airline's routes can drive a competitive advantage. Here we take the top 25 airlines by # unique routes and assess the reachability of their network. To explore this idea, a meta-analysis is conducted of each airline's network where a directional sub-graph is constructed for each airline with nodes and edges that are routes operated by that airline. We are then able to perform analysis on these individual subgraphs and compare the results to view the similarities and disparities between the top airlines in the world. Of the 25 sub-graphs created, only 5 were strongly connected. For the purpose of the analysis, any metric that is dependent on a strongly connected component, we took the largest strongly connected component as the network. For the airlines in scope, the strongly connected component that made up the lowest proportion of the overall airline's network was 93.5% (KL airline). Therefore, we are confident that these strongly connected components are a suitable representation where needed.

Beginning with initial analysis into the shortest paths and average path lengths extracted from each sub-graph, the results show that the largest airlines exhibit similar path lengths, typically with 2 or 3 hops being the most common shortest path within their network. This also mirrors the overall flight route shortest path distance as in Figure 12.

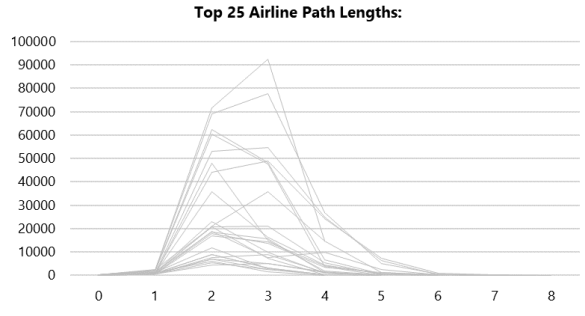

Figure 12: Top 25 Airline Path Lengths

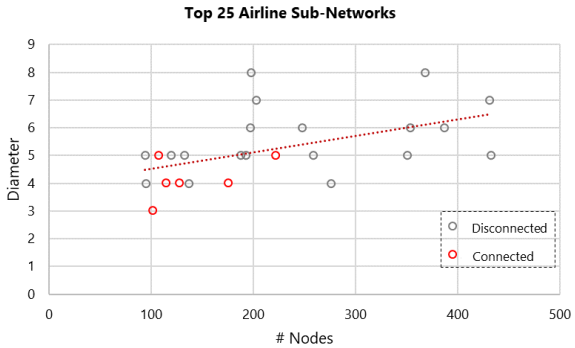

Figure 13: Top 25 Airline Sub-Networks

As expected, the diameter of the networks also correlates with the number of nodes within the networks. This is because on average as the number of nodes increases, the number of possible paths between any two nodes also increases.

In plotting the top airlines' sub networks, we begin to see some interesting outliers:

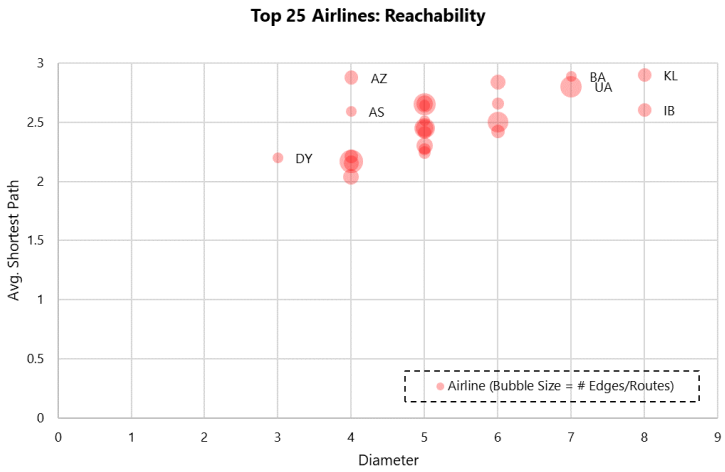

Figure 14: Top 25 Airlines: Reachability

- **AZ** (Alitalia) has a high average, shortest path length relative to its network's diameter and number of nodes. This could suggest increased fuel costs, flight delays and network efficiency due to the network complexity.
- **DY** (Norwegian Shuttle) has a low network diameter when compared to the other large airline peers. This airline likely has good connectivity and operational efficiency due in part to the geographic location of its main hub and the fact that much of its network is within Norway.
- **UA** (United) has the second largest number nodes of all airliners, however the diameter and avg. shortest path. The nature of the beast here, in that this airline covers so many locations domestically and internationally. The number of nodes is also a likely reflection of turnover, although in looking at competition with a similar number of airports - AA (American), we see that American are able to command a 5% smaller average shortest path and 28% smaller network diameter.
- **IB** (Iberia) has a low average, shortest path length relative to its high network's diameter and the number of nodes. Spanish airline which serves many far flung islands.

Further analysis into these outliers, versus the top 25 peer average. Using degree centrality to discern the hub for each carrier, then assessing the significance of its centrality. There is an observable pattern with average shortest path and the centrality of the airline's hub (table). Plotting the top 25 airlines shows a correlation between the degree centrality of the hub and the average shortest path (chart).

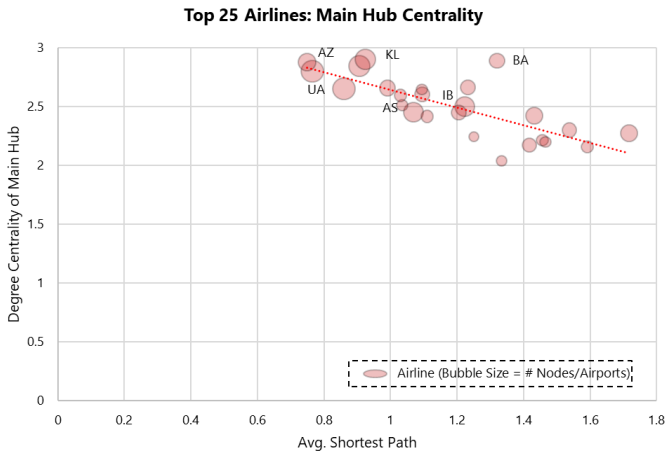

Figure 15: Top 25 Airlines: Main Hub Centrality

Therefore, airlines that may look to reduce operational complexity can look to reduce flights outside of their hub, route existing flights through their hub or generate more flights via their hub. Here the evidence supports intrinsic logic around operations airline planning. Furthermore, a large diameter in a network may not necessarily be a negative as a larger diameter in global transport could suggest a diverse network which may be beneficial for passengers seeking a wide variety of destinations or layovers.

Through this meta-analysis of airline sub-graphs, we have shown that this analysis can be extremely useful in competitor analysis, benchmarking, and operational optimization for airlines. As discussed, merely viewing an airline's metrics against the whole network's or the combined top airline's may be of limited value as the geo-locational component of the flight route network lends itself to certain limitations to the airlines. Therefore, a more specific derived comparison between airlines against specifically chosen peers can yield actionable insights.

6. Methodology

6.1. Node Classification

In this section, we perform node classification to classify the size of an airport with regard to centrality measures (including In-degree Centrality, Out-degree Centrality, Betweenness Centrality, Closeness Centrality, and Eigenvector Centrality). The airport size can be large, medium, or small.

Our objective is twofold. Firstly, we aimed to perform benchmarking and analysis by comparing the operational efficiency and profitability of airports. The classification can identify areas for improvement and inspiration as well as understand the best practices and policies that can be shared across the industry. Secondly, we aimed to perform optimization and marketing by helping airlines, travel agencies, and tourism boards target specific markets. One example is that airlines may select larger airports that are better equipped to handle larger aircrafts and more passengers. Moreover, travel agencies may promote smaller airports as a more convenient and hassle-free alternative to larger airports.

To train our model, we divided our dataset into 60% for training, 20% for validation, and 20% for testing. Prior to feeding the nodes and features into the classifier, we applied three GraphSAGE convolutional layers [9] with ReLU as activation functions. Next, we introduced three linear layers with ReLU as an activation function between the layers to classify the airports into three categories: large, medium, or small. The optimizer used was Adam with a learning rate of 0.01, while the loss function was cross-entropy loss.

We trained the network for 500 epochs. It started with a loss of 1.106, validation accuracy of 0.123, and test accuracy of 0.123. With cross-entropy loss, it reached optimal performance around 100 epochs, and remained stable with little variations from 100 to 500 epochs. At the end, the loss decreased to 0.432, and the best validation and testing accuracy were 0.830 and 0.808 respectively.

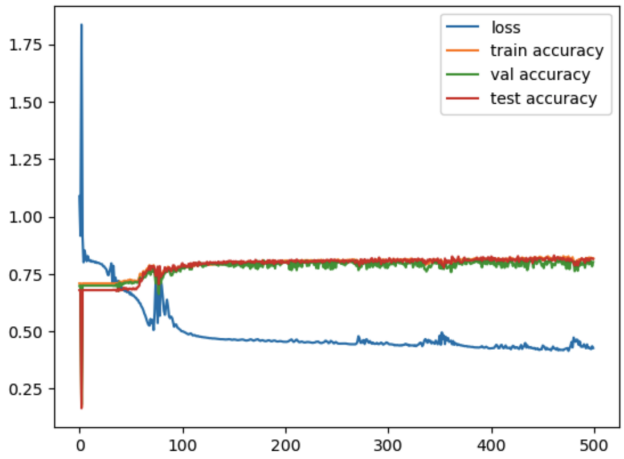

Figure 15: Top 25 Airlines: Main Hub Centrality

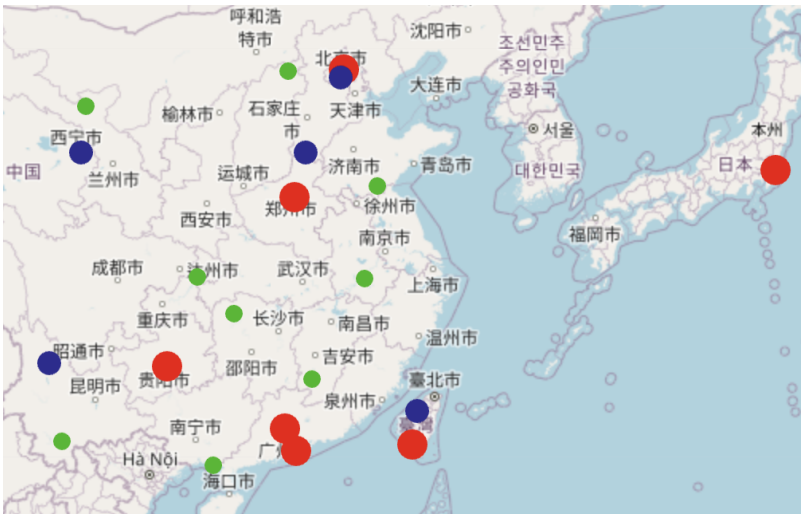

Figure 16: Top 25 Airlines: Sampled airports with airport size
Red: Large Airport, Blue: Medium Airport, Green: Small Airport

6.2. Node Clustering

We applied node clustering to airports in our network based on their route patterns as reflected in the weighted and non-symmetric adjacency matrix (as shown in Figure 3). The goal of this

clustering is to gain insights into route recommendation and optimization, and operational efficiency improvements for airlines, airports, and passengers, by analyzing and comparing various network features, including clustering coefficient and centrality measures outlined in Section 3.5. For example, if a cluster has a low degree centrality score but high betweenness score, we can suggest adding new routes to connect the underserved but high-demand airports to other airports with high visitor traffic.

To achieve this, we employed a 2-layered Graph Neural Network (GNN) with two GraphSAGE convolutional layers [9] to create node embeddings that capture the network structural information. We then used these embeddings to determine the similarities between airports based on their route patterns, and applied the K-means algorithm to create clusters from the embeddings, providing a partition of the airports according to their route similarity. The elbow curve in Figure 17 shows that the optimal number of clusters k should be between 2 and 3. We visualized the resulting clusters for both $k = 2$ and $k = 3$ scenarios in Figure 18.

To determine the best k for our air transportation network, we analyzed each cluster's characteristics by computing their average clustering coefficient and various centrality measures (including In-degree Centrality, Out-degree Centrality, Betweenness Centrality, Closeness Centrality, Eigenvector Centrality, and PageRank Centrality) across all clusters for both $k = 2$ and $k = 3$ scenarios, as shown in Figure 19 and Figure 20. We found that the second cluster (label 1) is dominant for almost every centrality measure among all three clusters. However, the other two clusters (label 0 and 2) have very close centrality scores and average clustering coefficient, suggesting that they be merged. Thus, the optimal number of clusters was determined to be 2.

In Figure 20, we also observe that the second cluster (label 1) is dominant for almost every centrality measure. The clustering coefficient values and closeness centrality scores remain consistently high across both clusters. As discussed in Section 3, this result suggests strong connectivity patterns and interconnectedness among the nodes' neighbors in our network, discouraging us from making route recommendations.

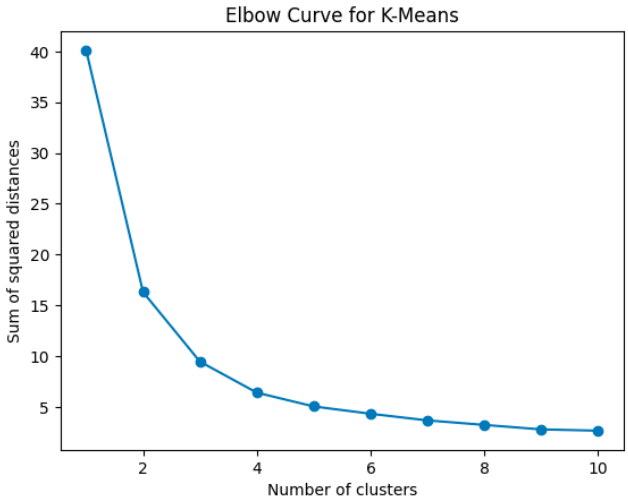

Figure 17: Elbow Curve for K-Means

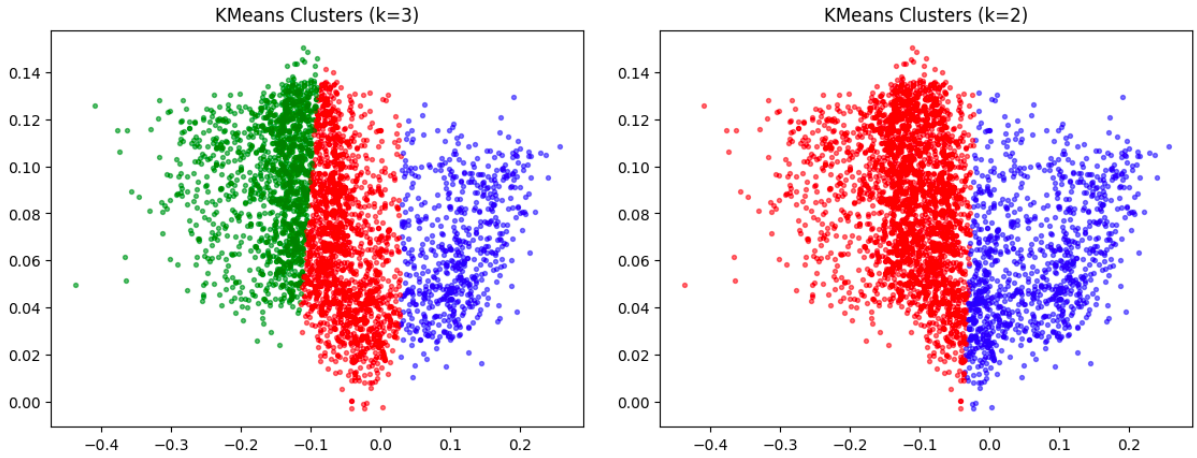

Figure 18: K-means clustering outputs

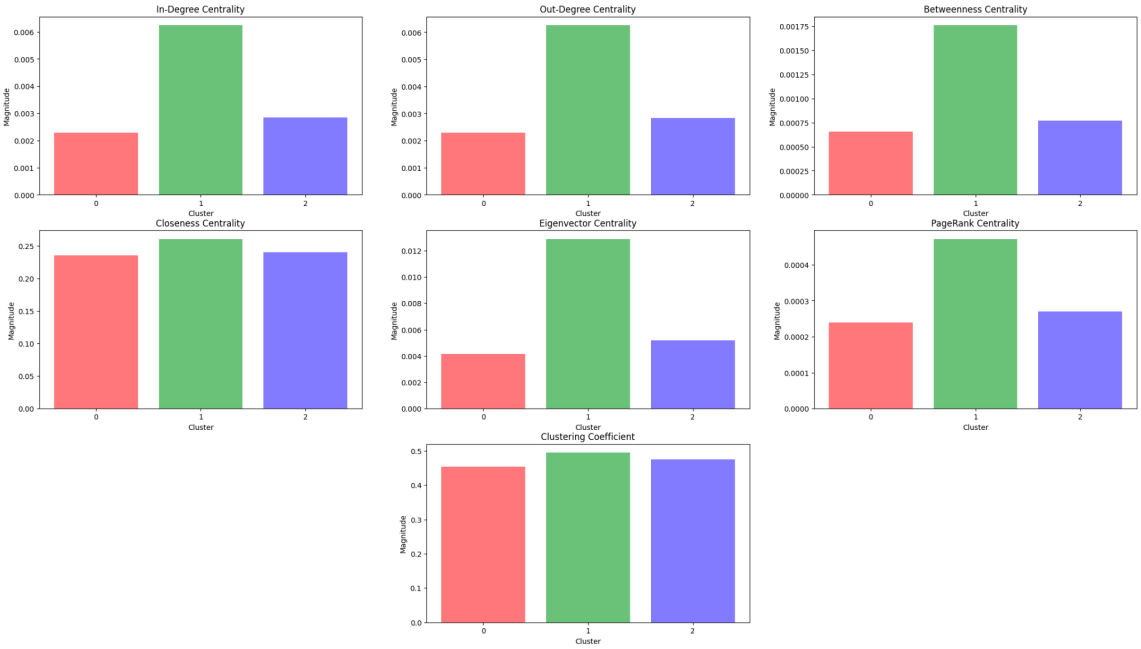

Figure 19: K-means clustering outputs for 3 clusters

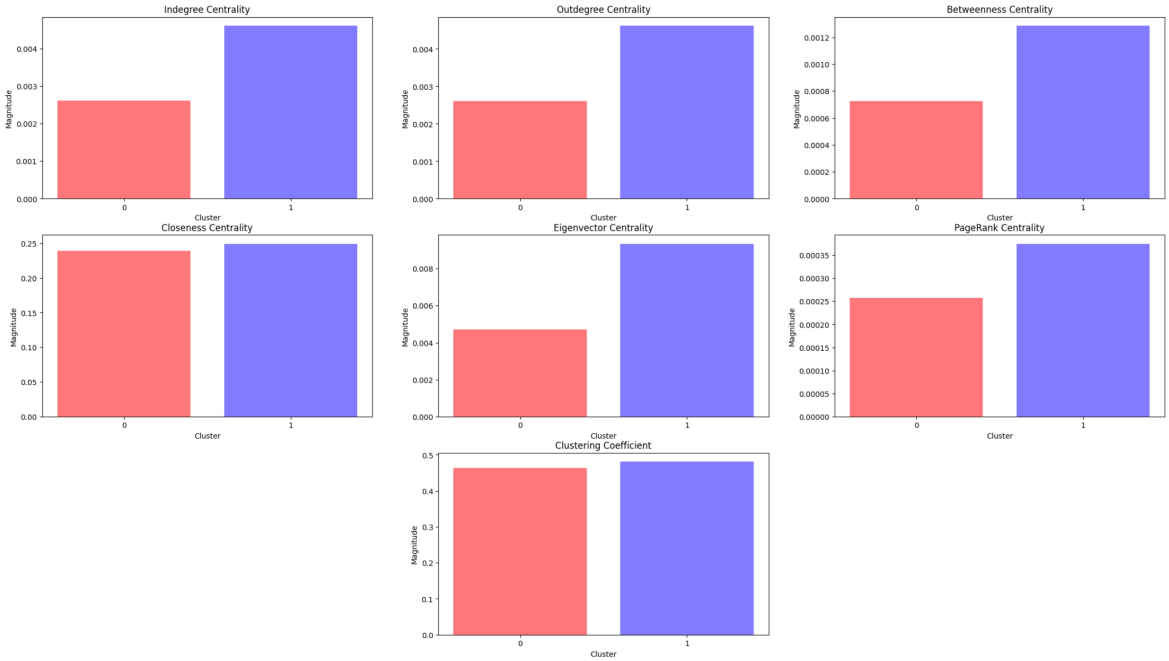

Figure 20: K-means clustering outputs for 2 clusters

7. Conclusion

In this project, we successfully applied graph neural networks to analyze airport networks and flight routes, focusing on node classification and clustering tasks. Our approach allowed us to classify airports based on their size and cluster them according to their route similarities, providing valuable insights for route optimization and operational efficiency improvements.

The results of our analysis indicate that GNNs are well-suited for complex network analysis tasks, such as those encountered in the analysis of airport networks. By comparing various centrality measures across clusters, we were able to identify patterns and trends that could inform the decision-making process for airlines, airports, and passengers.

In conclusion, this project demonstrates the potential of graph-based approaches for understanding airport networks and provides a foundation for future work in link prediction, temporal analysis, and the investigation of multimodal transportation networks. Our findings have the potential to drive innovations in the aviation industry and contribute to more efficient, sustainable, and interconnected transportation systems.

8. Future Work

Building on the promising results of this project, there are several avenues for future research that can further improve our understanding of airport networks and their underlying patterns. Some potential directions for future work include:

1. Link Prediction: Incorporating link prediction tasks to forecast the establishment of new routes between airports or the termination of existing ones. This can help airlines and airports optimize their route planning and resource allocation strategies. To achieve this,

we would require additional attributes including location, size as well as the demand to use the new airport.

2. Temporal Analysis: Extending the current analysis to incorporate temporal information, allowing for the investigation of the evolution of airport networks over time. This would enable the identification of emerging trends and potential growth opportunities in the aviation industry.
3. Multimodal Transportation Networks: Expanding the analysis to encompass other modes of transportation, such as railways and highways, and investigating the interplay between these networks. This could provide valuable insights into the broader transportation ecosystem and help inform integrated transportation planning.
4. Node Feature Engineering: Exploring additional node features that can enhance the classification and clustering tasks, such as passenger traffic, airport facilities, or socio-economic factors.
5. Alternative Clustering Techniques: Investigating other clustering algorithms or GNN architectures to improve the quality of the clusters and potentially uncover more nuanced patterns in the airport network data.

Reference

[1] Barabási, A.L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., and Vicsek, T. (2003). Evolution of the social network of scientific collaborations. Nature 425, 560-563. DOI: 10.1038/nature02070.

[2] J. Alstott, E. Bullmore, and D. Plenz, “Powerlaw: A python package for analysis of heavy-tailed distributions,” *arXiv.org*, 31-Jan-2014.

[2] D. J. Watts and S. H. Strogatz, “Collective dynamics of 'small-world' networks,” *Nature News*.

[3] S. Mohammed, E. Ramy, T. Matti, et al., “How well centrality measures capture student achievement in computer-supported collaborative learning? – A systematic review and meta-analysis,” *Elsevier Ltd.*, *Educational Research Review*, 35(2020), 100437.

[4] H. Liao, M. Mariani, M. Medo, Y. Zhang & M. Zhou (2017), “Ranking in evolving complex networks,” *Physics Reports*, 689(April), 1–54.

[5] S. C. Suvarna, M. Srivastava, B. Jaganathan, and P. Shukla, “PageRank algorithm using eigenvector centrality- new approach,” *arXiv.org*.

[6] Easley, D., & Kleinberg, J. (2010). Networks, Crowds, and Markets: Reasoning About a Highly Connected World. Cambridge University Press.

[7] Brandes, U. (2001). A faster algorithm for betweenness centrality. Journal of Mathematical Sociology, 25(2), 163-177.

[8] Langville, A.N., & Meyer, C.D. (2011). Google's PageRank and Beyond: The Science of Search Engine Rankings. Princeton University Press.

[9] William L. H., Rex Y. and Jure L. (2018). “Inductive Representation Learning on Large Graphs.” arXiv.org.