University of Bologna

Social Network Analysis

Open Flights

Analysis of OpenFlight's Airplane Routes Network

Brajucha Filippo [0001130613]

Leonardo Dessì [0001141270]

Gianmarco Gabrielli [0001133622]

Simone Rinaldi [0001140193]

Contents

1 Introduction

This project falls under the specific field of **Transport Network Analysis**, an interdisciplinary research area that applies Network Science to the study of transportation systems. Transport Network Analysis uses mathematical and computational tools derived from graph theory to analyze, to model and to understand the flow of people, goods, and information through complex transportation networks.

In this project, we apply Transport Network Analysis to the study of the global air transport system. Specifically, we will analyze the network of airports and the air routes that connect them to identify the most relevant and strategic airports on a global level and subsequently test the hypothesis that these airports are critical to the overall connectivity of the network.

2 Problem and Motivation

The air transportation network is one of the most complex and vital networks in modern society, enabling global connectivity and supporting economic growth. Despite its importance, this network is vulnerable to disruptions caused by natural disasters, geopolitical events, technical failures, and pandemics. These disruptions can have significant cascading effects, leading to delays, economic losses, and reduced global mobility. Therefore, understanding the structure of the airline network and identifying critical nodes (airports) is crucial to ensuring the stability and resilience of air travel.

This project aims to analyze the global network of airports and air routes. The main goal is to **identify the most relevant and strategic airports in terms of network**.

- Understanding the structure and dynamics of global air transport: Identifying major hubs and their influence on global connectivity can provide valuable information on how air traffic spreads around the world. This analysis can contribute to scientific research in the field of network analysis applied to air transport.
- Optimize route planning and air traffic management: Knowledge of the most strategic airports can help airlines optimize their routes, improving air transport efficiency and reducing costs.
- Improving flight network resilience to disruptions: Understanding the role of hubs in global connectivity enables assessment of the network's robustness to unforeseen events such as airport closures or route cancellations. This knowledge can be used to develop strategies that minimize the impact of such disruptions on the global air transportation system.

The main contributions of the project include:

- Identification of the most globally relevant airports: Using various measures of centrality such as *Degree Centrality*, *Eigenvector Centrality*, *PageRank*, etc., the

project aims to identify airports that play a key role in the global air transport network.

- Verification of the hypothesis that airports with high betweenness centrality are critical to global connectivity: Through robustness simulations, the project will evaluate the impact of removing these airports on network connectivity. Connectivity, fragmentation, and compactness will be used to measure the impact.
- The identification of regional airport clusters: Using clustering algorithms such as the community *louvain algorithm*, the project will identify groups of strongly interconnected airports within specific geographic areas.

3 Dataset

3.1 Dataset Description: OpenFlights Air Route Network

The dataset used for this project was obtained from the OpenFlights website (and github), a public resource that provides information on airports and air routes around the world. The data were downloaded in digital format and are freely accessible from both the website and the associated GitHub repository. OpenFlights data is divided into two databases:

- **Airport database**: contains information on more than 10,000 airports, train stations, and ferry terminals worldwide. Each entry includes details such as airport ID, name, city, country, IATA and ICAO codes, geographic coordinates, and data source.
- **Route database**: contains information on 67,663 routes among 1,578 airports, operated by 548 airlines (data as of June 2014). Each entry includes airline code, airline ID, departure airport, departure airport ID, destination airport, destination airport ID, and type of aircraft used.

Data manipulation and analysis tools were used to analyze and construct the airport network graph. These tools included the Python programming language, which offers libraries specifically designed for network analysis, such as NetworkX.

The network graph used airports as nodes and direct air routes as edges. The edges were weighted according to the number of existing flights or connections between airports to reflect the intensity of air traffic.

3.1.1 Format and Structure

The process of converting the OpenFlights database into a networkX graph for network analysis is as follows:

- **Data import**: Airport and route data were imported into Python using Pandas library, which allows tabular data to be handled efficiently.
- **Node creation**: Each airport in the database was converted into a graph node. Relevant information about the airport, such as name, city and country, was associated as attributes to the corresponding node.
- **Creation of edges**: For each air route in the database, an edge was created between the nodes corresponding to the departure and arrival airports. The presence of an edge thus indicates the existence of a direct air link between two airports.
- **Edge weighting**: To reflect the intensity of air traffic between airports, the arcs were weighted according to the number of flights or connections present between the two connected airports. This information was extracted from the route database and assigned as an attribute to the corresponding edge.
- **Distance calculation**: using the coordinates in the airport database, distances in km between each airport were calculated and added as an attribute of the edges.
- **Symmetrization**: in order to create an undirected graph, a summation of the weights of the edges passing through the same nodes was performed.

The result of this process is a weighted, undirected graph, where the nodes represent the airports and the edges represent the direct air routes between them. The weight of each edge indicates the intensity of air traffic between the connected airports.

The structure of the graph is shown in figure 1. In this figure are shown only the nodes with degree greater or equal than 10, and the dimension of the node represents the betweenness centrality of the node.

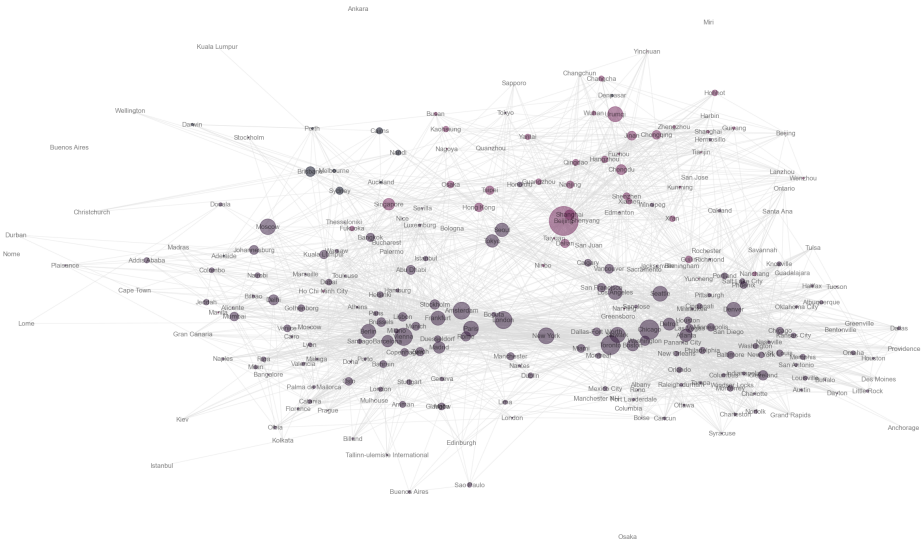


Figure 1: Structure of the main node graph

In the figure 2 the nodes are placed in the world map using their coordinates, and the arcs represent the links between airports.

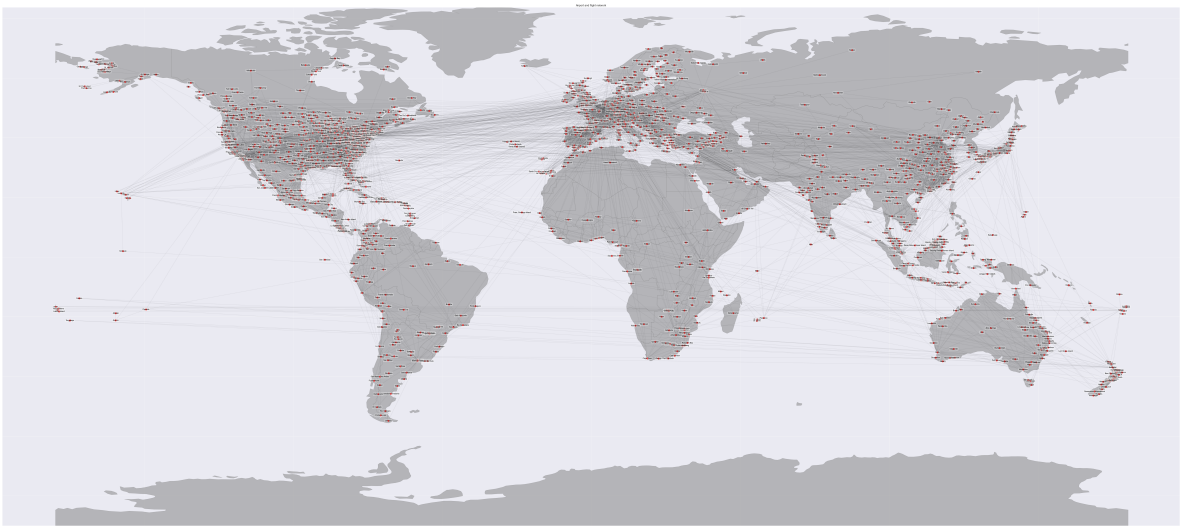


Figure 2: Graph structure displayed in the world map

4 Validity and Reliability

Our analysis is built on the reputable `OpenFlights` dataset and standard network metrics, ensuring both validity and reliability.

To maintain validity, we:

- Use established centrality measures (e.g., degree, betweenness, eigenvector) that effectively capture airport connectivity.
- Construct a weighted, undirected graph that faithfully represents global air routes.

Reliability is ensured by:

- **Data Reliability**: The OpenFlights dataset, sourced from a reputable public repository and maintained on GitHub, is extensively used in academic research. We have taken care to clean and preprocess the data to minimize errors and inconsistencies. Where possible, data points were cross-verified with available documentation to ensure accuracy.
- **Measurement Reliability**: Network metrics were computed using the established algorithms provided by the NetworkX Python library. These algorithms have been thoroughly tested in many studies, ensuring that our computations are robust. Repeating the analysis with the same dataset consistently yields identical results, confirming the reliability of the measurement process.
- **Replicability**: We have provided a detailed description of our methodology, from data import and preprocessing to network construction and metric computation. The use of open-source tools such as Python, Pandas, and NetworkX ensures that other researchers can replicate our analysis under similar conditions. This transparency in our workflow enhances the replicability of the study.
- **Sensitivity Analysis**: To further assess reliability, we conducted robustness tests by removing the top hub nodes and evaluating the impact on network fragmentation and compactness. The observed changes were consistent with theoretical expectations, demonstrating that the network's behavior under stress aligns with our computed metrics. Such sensitivity analysis lends additional support to the reliability of our results.

Despite the rigorous approach, certain limitations must be acknowledged. The dataset represents a static snapshot of the air transportation network and may not capture temporal dynamics or recent changes in air traffic. Additionally, the decision to symmetrize the graph, while useful for the analysis, may result in the loss of directional information that could be significant in other contexts.

5 Measures and Results

For the metrics graph we use the `IATA/FAA` value id, in order to distinguish every airport. It's important to say that there could be more than 1 airport in the same city.

5.1 Measures

This dataset is ideal for exploring various network science topics, including:

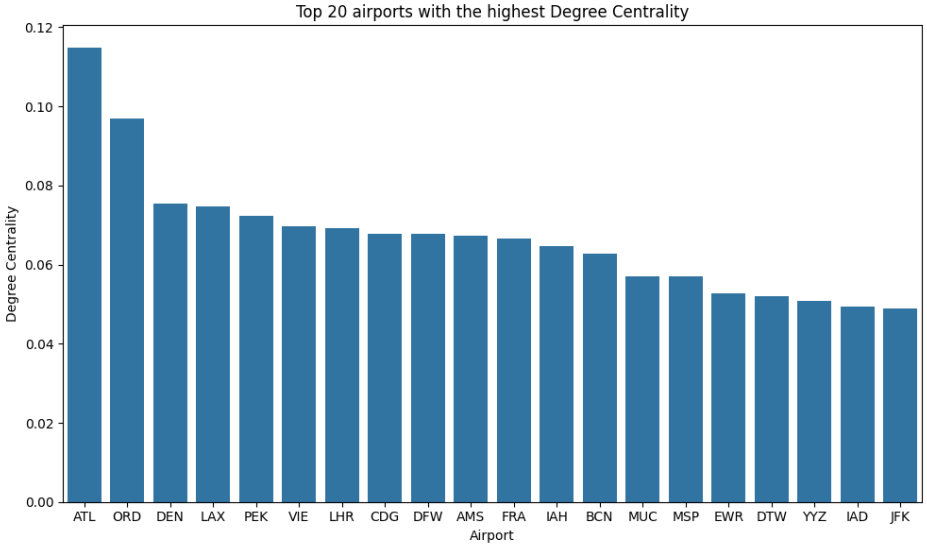
- **Degree Centrality**: Evaluate the relative importance of a node based on the number of direct connections it has. Airports with high degree centrality are those that serve as primary points of connectivity in the network by offering a larger number of direct flights to different destinations. This high connectivity makes them essential to the efficiency and accessibility of the overall air transportation network.
- **Betweenness Centrality**: Quantifies the extent to which a particular node serves as a bridge or intermediary in the network. In the context of air travel, it measures how frequently an airport lies on the shortest paths between other pairs of airports
- **Closeness Centrality**: It measures the average shortest path distance from a given airport to all other airports. Airports with high closeness centrality are those that can quickly reach, or be reached by, other airports in the network, making them strategically helpful for minimizing travel time and improving accessibility.
- **Eigenvector Centrality**: Measures how well-connected an airport is to others. An airport with high eigenvector centrality is one that not only has many connections but is also linked to other highly connected and influential hubs, amplifying its importance within the global air transportation network.
- **PageRank**: It evaluates the significance of an airport by considering not only the number of its connections but also the importance of the airports it is connected to. This metric highlights influential airports, even when many of their connections are to less prominent nodes.
- **Clustering Coefficient**: Quantifies how likely it is that an airport's neighboring airports (i.e., those directly connected to it) are also connected to one another. It provides insights into the local interconnectedness of an airport's surrounding network.
- **K-core**: Identifies a subnetwork of airports that are highly interconnected, where each airport in the subnetwork has at least k connections to other airports in that subnetwork. The k-core helps to highlight core airports that are central to a more tightly-knit cluster of airports within the broader network.

5.2 Results

5.2.1 Degree Centrality

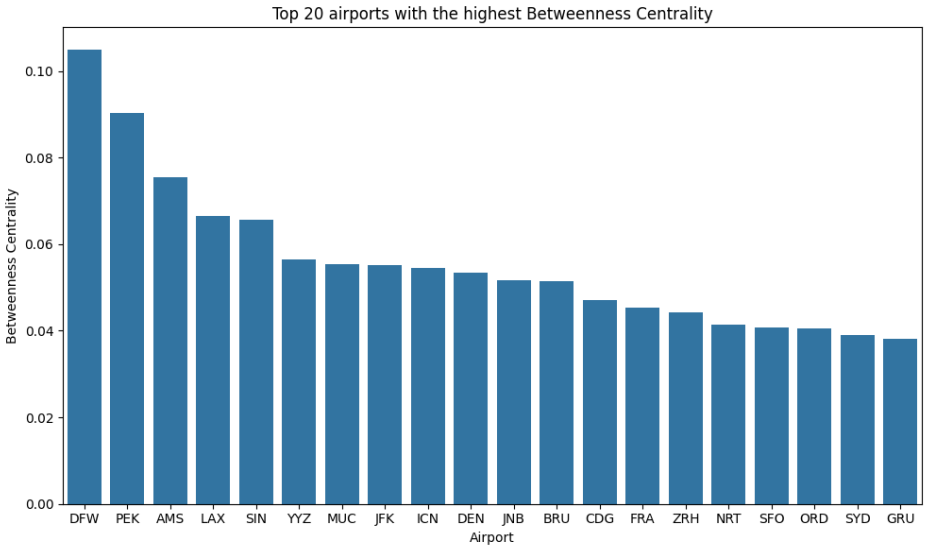
This chart shows the top 10 airports ranked by their degree centrality, which measures the number of direct connections an airport has with others. We can see as cities such

as **Atlanta** (ATL) and **Chicago** (ORD) has a degree centrality value much higher than others (almost higher than *0.09*), starting from **Denver** (DEN) the value descent, it remains on *0.07* for all the top 13 airports and then it falls under *0.06*.
Their high-degree values indicate that these urban centers serve as major direct connectivity hubs.
We can make some observation linked to the macro-areas represented by the continents. It's curious to see that the top 3 (also top 4) is totally composed by US cities, so we can assume the centrality of this state in direct airplane connection across the World, there are other US cites in the top 10.
Looking at european cities we can see that in the top 10 there is **Vienna** (VIE), **London** (LHR), **Paris** (CDG) and **Amsterdam** (AMS).
It's also curious to see that there is just 1 asian city in the top 10 rank, **Beijing** (PEK). We can imagine that there could be some problems linked to a "bottleneck effect" with just 1 city with high degree centrality.



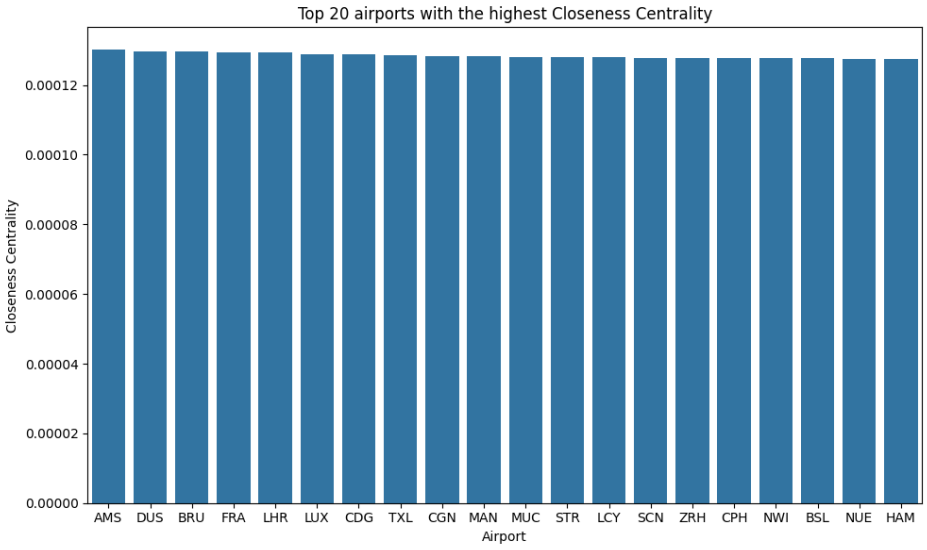
5.2.2 Betweenness Centrality

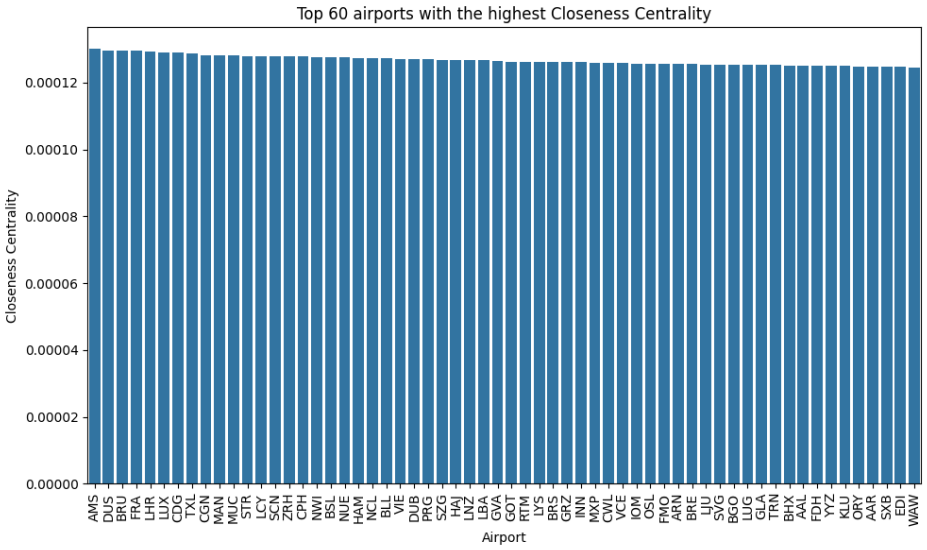
This bar chart shows the top 10 airports with the highest betweenness centrality.
Cities like **Dallas** (DFW) and **Beijing** (PEK) rank highest in betweenness centrality. This suggests that they play a key role as intermediaries, facilitating indirect connections across the network.
We can find a lot of cities that appears in between centrality top 10 and also in degree centrality, this suggests that these nodes are the main controller of the airplane flow. These cities are very strong bridges that allow the connection across the World.
The first 5 cities (Dallas, Beijing, Amsterdam, Los Angeles LAX and Singapore SIN) have little higher value than others, so we can assume that are more important. It's fascinating how there are 2 asian cities that have not so high degree centrality but high betweenness centrality.



5.2.3 Closeness Centrality

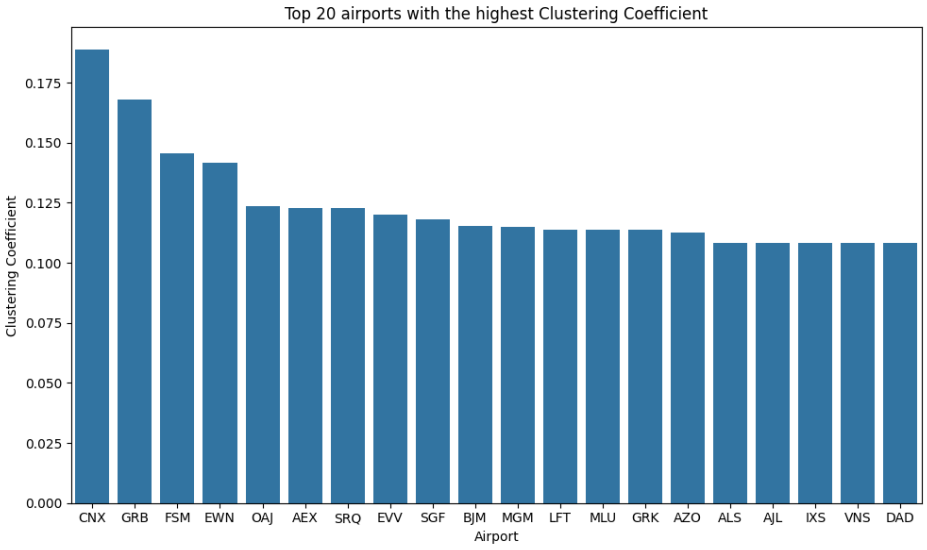
This chart ranks the top 10 airports according to closeness centrality, a measure of how well-connected an airport is within the network. High closeness means a node can quickly reach all other nodes, making it efficient for information spread.
It's not really self-explanatory because the value is stable for a lot of airport around `0.00012`, but we can assume that the network is well-connected.
We also plot the result with more than 20 airports.





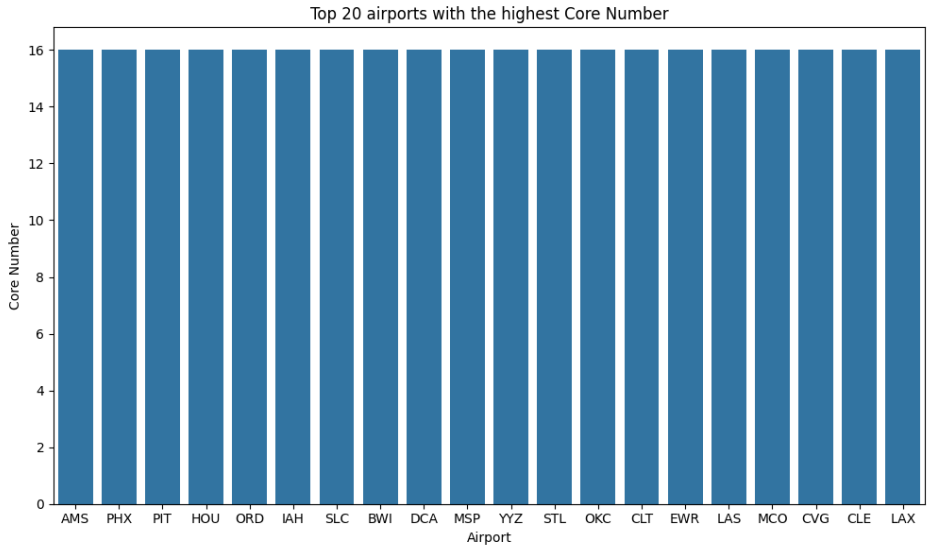
5.2.4 Clustering Coefficient

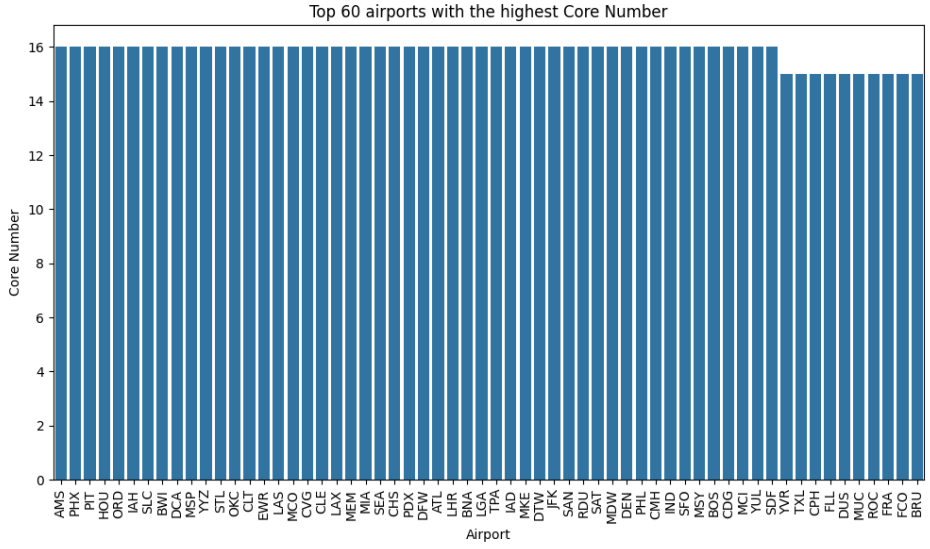
This bar chart shows the top 20 airports by clustering coefficient, suggesting they maintain tightly-knit local neighborhoods where many neighboring cities are also directly connected. Chiang Mai International Airport (CNX) leads with a coefficient of `0.18`, followed by Austin (GRB) with `0.17` and Fort Smith (FSM) and James City (EWN) with around `0.15`. The values decrease gradually to `0.125` for Jacksonville Airport (OAJ). These airports serve as local hubs where connected airports also tend to connect with each other.
It's interesting to see that there are littler american airports than the bigger ones, it means that the USA has a well-connected structure of internal flights in order to move across different states.



5.2.5 K-cores

K-Cores are used to identify the most tightly connected parts of a graph and can provide insights into the overall structure and behavior of a network. Here we can see that the 20 airports with the highest core number is equal to 16.
Because of the nature of the metrics we decided to see the others airport that has k-cores equal to 16 so we plotted other results. As we can see they are more than 50. So the network analyzed has a robust core group made by at least 50 airports across the world.

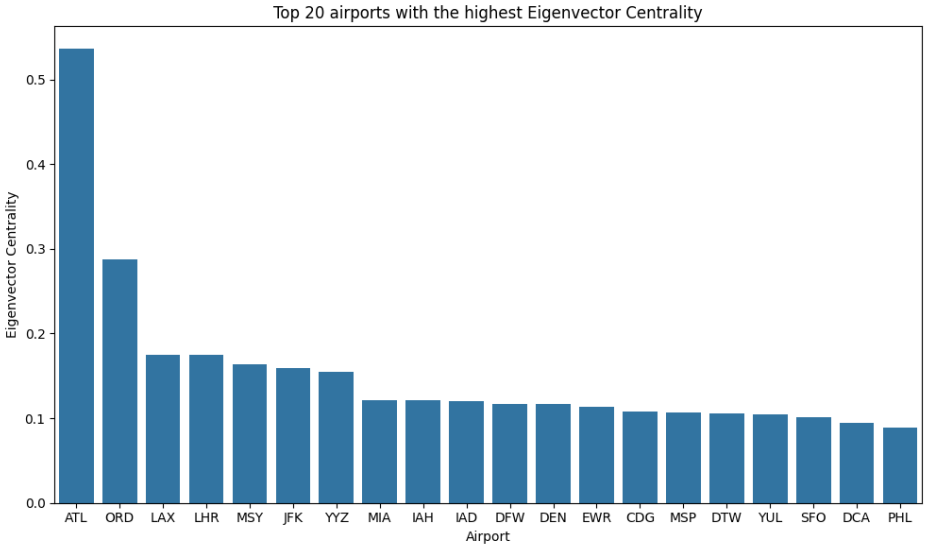




5.2.6 Eigenvector Centrality

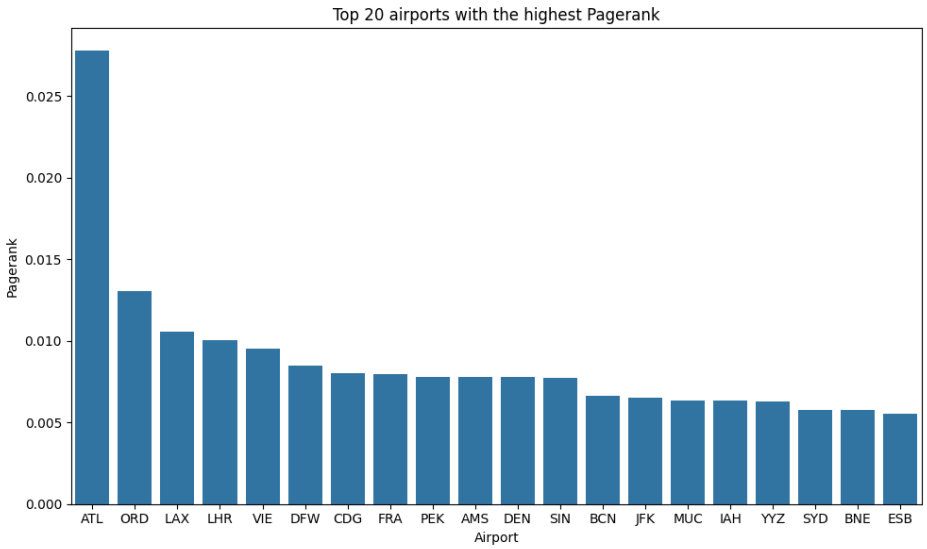
The rankings again favor influential hubs such as **Atlanta** (ATL) and **Chicago** (ORD). Their high scores reflect not only many connections but also links to other highly con-

nected cities. An airport with many connections to small, isolated airports would have a lower eigenvector centrality than one with fewer connections to major hubs.
We can see that the first 2 position are occupied by the same airport that has also high Degree Centrality, so we can assume that they are crucial for the network and they could represent a *bottleneck* in the network.



5.2.7 Pagerank

Results mirror those of eigenvector centrality, with cities like **Atlanta** (ATL), **Chicago** (ORD), and **Los Angeles** (LAX) emerging as pivotal nodes. This metric emphasizes the importance of quality as well as quantity in their connections.
This metric is also very similar to the Eignvector ones because of it's nature, but it evidence better the importance of ATL airport.



5.2.8 Metrics distribution

In the following figures are displayed node metric distributions in both linear and log scales. The linear scale clearly shows absolute differences and the overall spread of values across the network. However, many network metrics, such as degree or centrality measures, tend to be highly skewed or follow a heavy-tailed distribution. A log scale compresses the high-end values and expands the lower end, making it easier to observe patterns in the tail (e.g., power-law behavior) and identify significant yet less frequent nodes. Together, these views provide a more complete understanding of the network's structure.

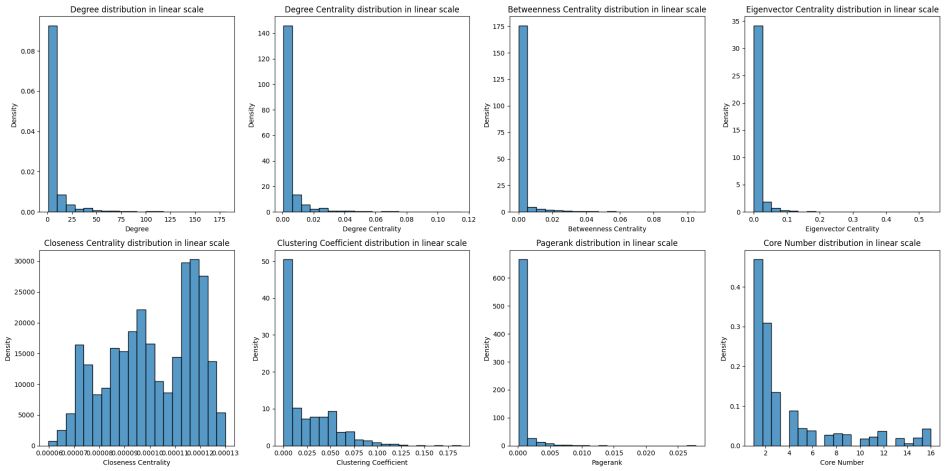


Figure 3: Metrics distribution in linear-scale

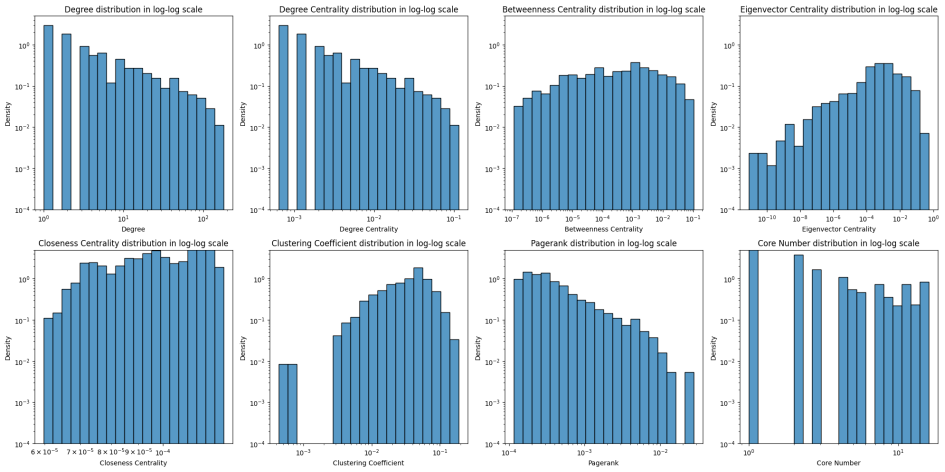


Figure 4: Metrics distribution in log-scale

5.2.9 Density

The computed density of `4.57e-3` indicates that only about `0.46%` of all possible connections between airports are present in the network. This very low density is expected

in an air transport network, where airports are not universally connected but rather linked through a few strategic routes and major hubs. The sparsity highlights that the network relies on selective connectivity to maintain efficiency and manage logistical constraints, rather than having a fully connected structure.

5.3 Additional Metrics Description

- **Average Path Length:** `10210.055112776308`

  This very high value likely reflects a weighted average, where the weights (distances in kilometers) are aggregated along paths. It indicates that, when considering these weights, the typical "distance" between airports is large. In contrast, the unweighted (hop-based) distances are much shorter.

- **Diameter of the Network:** `12`

  The diameter represents the maximum number of hops between any two airports. A diameter of 12 shows that, topologically, any airport can be reached from any other in no more than 12 steps, indicating a compact structure in terms of connectivity.

- **Number of Connected Components:** `1`

  This result confirms that the network is fully connected; every airport is reachable from any other, ensuring a unified air transport system.

- **Assortativity:** `-0.17902211456273712`

  The negative assortativity indicates a tendency for high-degree nodes (major hubs) to connect with low-degree nodes (smaller or regional airports). This disassortative mixing is typical in transportation networks, where central hubs connect to many less-connected nodes.

- **Number of Bridges:** `559`

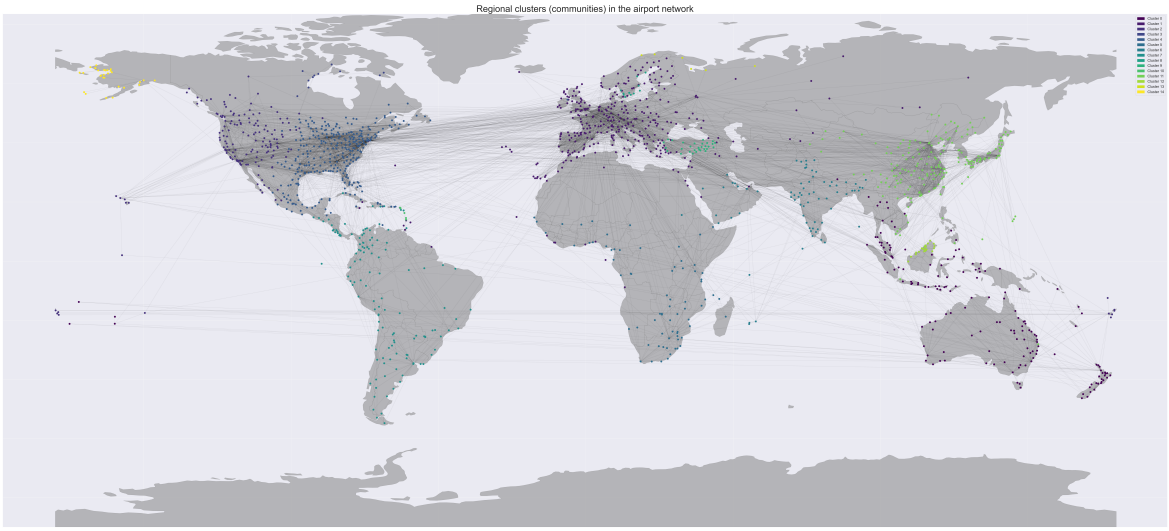
  Bridges are edges whose removal would disconnect parts of the network. The presence of 559 bridges suggests that there are several critical routes whose loss could fragment the network, highlighting areas of potential vulnerability.

5.3.1 Powerlaw

The power law fit yielded an exponent (α) of approximately 1.8023. In a distribution of the form $P(x) \propto x^{-\alpha}$, this value indicates a heavy-tailed behavior. Specifically, an exponent around 1.8 suggests that while many nodes (e.g., airports) have relatively few connections, a small number have extremely high connectivity. This pronounced heterogeneity implies that a few key hubs dominate the network structure, making them crucial for overall connectivity.

5.3.2 Clustering

Here is a representation of the clustering of the airports.



5.4 Robustness

The ability of the network to maintain a connection between all (or most) pairs of nodes when challenged. First, our team computed the first 10 hubs with the highest degree value. After that, we removed them from the network to check connectivity. The result was that we lost connectivity after the elimination of the hubs. Follow the fragmentation computation that resulted very low, so, the network is not fragmented. The third step is the computation of compactness. Initially, it was around `0.25`, but after removing the hubs, the network was not compact anymore.

5.4.1 Fragmentation Computation

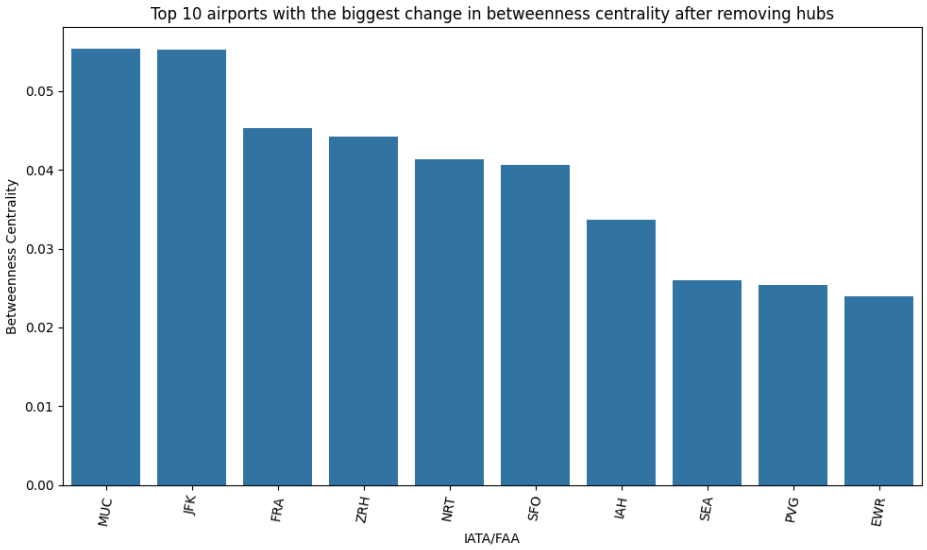
Graph fragmentation involves breaking down a graph into smaller, more manageable subgraphs. We first computed the fragmentation before removing the hubs, with a value equal to 0. Then we tried to remove the hubs, and we got a graph with fragmentation value of `0.05`.

5.4.2 Compactness Computation

Graph compactness tells how "tightly" a graph is connected. The value for the original graph is `0.25`. After removing the top 10 hubs, the graph is no longer compact.

5.4.3 Centrality Analysis

At the end, we proceeded with the computation of the Betweenness Centrality. Then we visualized the most influenced airports about the new computation of the metric.



6 Conclusions

The conclusions of the study on the global air transport network highlight several critical insights derived from the analysis of airport connectivity and centrality measures. The project aimed to identify the most relevant and strategic airports that play a pivotal role in facilitating global connectivity. Through the application of various centrality metrics, including betweenness centrality, closeness centrality, eigenvector centrality, and PageRank, the research successfully pinpointed key airports that serve as vital nodes within the air transportation system.

One of the primary findings of the study was the confirmation of the hypothesis that airports with high centrality between is essential to maintain global connectivity. These airports act as crucial bridges along the shortest routes between other airports, indicating their importance in linking different regions and facilitating efficient air travel. Robustness simulations conducted during the analysis demonstrated that the removal of these high-betweenness airports would lead to significant fragmentation and reduced connectivity within the network. This underscores the necessity of these airports in ensuring seamless air travel and highlights their strategic importance in the global transportation landscape.

Additionally, the findings offer insights into the overall structure and dynamics of the air transport network. This information can improve route planning and air traffic management, helping airlines optimize operations, reduce costs, and enhance service. Moreover, understanding the role of major hubs aids in assessing network resilience to disruptions, ensuring the stability and reliability of air travel vital for economic growth and global mobility.