ENSAI
Master for Smart Data Science
IT Tools 2: NoSQL

A Graph-Based Analysis of the Global Air Transport Network

Egle Cadullo, Jesse van der Ende

École nationale de la statistique et de l'analyse de l'information

Professor:	N. Parlavantzas
Date:	November 2025

1 Introduction

Air transportation is one of the most interconnected and complex systems in the modern world, linking thousands of cities through a global network of routes. Understanding its structure is essential for analyzing international connectivity, identifying transportation hubs, and assessing the resilience of the global air network. Traditional relational databases, while effective for structured data storage, are limited in their ability to represent and analyze highly connected data efficiently. Queries involving multi-hop relationships, such as finding the shortest path between airports or identifying the most central hubs, often require computationally expensive joins and complex table structures (Robinson et al., 2015).

Graph databases, in contrast, provide a natural way to represent and query networks by modeling data as nodes and relationships. In such databases, connections such as routes between airports becomes both intuitive and computationally efficient. Neo4j, one of the leading graph database systems, allows users to model entities such as airports, airlines, and routes as nodes connected by directed relationships, and to analyze connectivity using its built-in Graph Data Science (GDS) algorithms. The combination of flexible schema, efficient traversal, and analytical capabilities makes Neo4j well-suited for studying transport networks (Robinson et al., 2015; Sadalage and Fowler, 2012).

This study uses the OpenFlights dataset, an open-source collection of global flight routes, airports, and airlines. By importing and modeling this dataset in Neo4j, we aim to explore the structure of the international air transportation network through graph-based analysis. The objective is not only to examine the dataset's connectivity properties but also to illustrate how NoSQL technologies can provide valuable analytical insights in domains traditionally handled by relational systems.

The following research questions guide this analysis:

1. *Which airports and airlines act as the main global hubs in the air transport network?*
2. *How evenly is air connectivity distributed across countries?*
3. *Which airports/countries function as key bridges linking otherwise distant regions?*

By answering these questions, this project demonstrates how a graph-based representation can reveal patterns of connectivity, hierarchy, and accessibility within the global air transport system.

2 Data

The analysis relies on the OpenFlights database, an open-source collection of information about airports, airlines, and routes worldwide. The dataset was originally compiled from multiple public sources and has been widely used in studies of global air-transport networks. Although it has not been regularly updated since 2014, it remains one of the most comprehensive freely available descriptions of the international flight network.

2.1 Data description

The OpenFlights[1] dataset contains three relevant files:

- `airports.dat`: a list of approximately 14 000 airports worldwide, including identifiers, geographic coordinates, city, and country. The number is much larger than the count of major commercial airports, as the file also covers small regional airfields, heliports, and private landing strips recorded in public aviation registries.
- `airlines.dat`: list of approximately 6 000 airlines with names, codes, country, and operational status. This figure exceeds the number of active commercial carriers because the dataset includes defunct, regional, charter, and cargo operators as well as historical records.

[1] https://openflights.org/data

- `routes.dat`: approximately 67 000 directed flight routes linking source and destination airports and specifying the airline operating each connection.

Each file follows a simple table format. The `airports.dat` file provides identifiers, names, geographic coordinates, and contextual information such as city, country, altitude, and time zone. The `airlines.dat` file includes the airline name, its alias, and the IATA and ICAO codes, which are respectively the two- and three-letter identifiers assigned by the International Air Transport Association and the International Civil Aviation Organization. The `routes.dat` file specifies the airline operating each route, the source and destination airports (both by code and by numeric ID), the number of stops, and the aircraft types used on the route.

The variables available in each file are listed in Tables 1, 2, and 3.

Variable	Description
`Airport ID`	Unique numeric identifier assigned by OpenFlights.
`Name`	Official airport name.
`City`	City served by the airport.
`Country`	Country where the airport is located.
`IATA`	Three-letter IATA code (may be missing).
`ICAO`	Four-letter ICAO code (may be missing).
`Latitude`, `Longitude`	Geographic coordinates in decimal degrees.
`Altitude`	Altitude in feet above sea level.
`Timezone`	Offset from Coordinated Universal Time (UTC).
`DST`	Daylight-saving time indicator.
`TZ database time zone`	Olson time-zone string (e.g. `Europe/Paris`).
`Type`	Classification such as airport, station, or port (only type airport is included here).
`Source`	Origin of the data record.

Table 1: Variables in `airports.dat`.

Variable	Description
`Airline ID`	Unique numeric identifier.
`Name`	Official airline name.
`Alias`	Alternative or marketing name.
`IATA`	Two-character IATA code.
`ICAO`	Three-letter ICAO code.
`Callsign`	Radio callsign used in communication.
`Country`	Country of registration.
`Active`	Operational status (`Y` = active, `N` = inactive).

Table 2: Variables in `airlines.dat`.

Variable	Description
`Airline code`	IATA or ICAO code of the operating airline.
`Airline ID`	Numeric identifier linking to `airlines.dat`.
`Source airport code`	IATA or ICAO code of the origin airport.
`Source airport ID`	Numeric identifier of the origin airport.
`Destination airport code`	IATA or ICAO code of the destination airport.
`Destination airport ID`	Numeric identifier of the destination airport.
`Codeshare`	Indicates if the route is part of a codeshare agreement.
`Stops`	Number of intermediate stops on the route.
`Equipment`	Aircraft type(s) used, represented by three-letter codes.

Table 3: Variables in `routes.dat`.

2.2 Graph modeling approach

The data were imported into a Neo4j graph database to represent the global air transport network and to enable connectivity analysis. The objective of the modeling process was to design a structure that captures the hierarchical organisation of the data (countries, cities, and airports) together with the operational relationships between airlines and the flight connections they provide.

2.2.1 Node types

Four main node labels were defined as:

- `Airport`: represents an individual airport or airfield, with properties as described in Table 1.
- `Airline`: represents a company operating commercial, charter, or cargo flights, with properties as described in Table 2.
- `City`: represents a metropolitan area served by one or several airports, defined by the property `name`.
- `Country`: represents the country in which airports and airlines are registered.

2.2.2 Relationships and directionality

Directed relationships were created to capture both geographical and operational dependencies. Airports are linked to the cities and countries in which they are located, while airlines are associated with their country of registration. Operationally, airlines are connected to the airports from which they depart and fly to as destination. Finally, the flight connections between airports are represented by `ROUTE` relationships, which also contain route-specific information such as the operating airline, number of stops, and aircraft type. Although the airline identifier is stored within each route relationship, there is no direct edge between airline nodes and these `ROUTE` relationships, since Neo4j does not permit links between nodes and relationships. A summary of all the relationships is given in Table 4.

Relationship	**Meaning**
`(:Airport)-[:IN_CITY]->(:City)`	Associates each airport with its city.
`(:City)-[:IN_COUNTRY]->(:Country)`	Assigns each city to a country.
`(:Airline)-[:BASED_IN]->(:Country)`	Links airlines to their country of registration.
`(:Airline)-[:OPERATES_FROM]->(:Airport)`	Indicates airports from which an airline departs.
`(:Airline)-[:OPERATES_TO]->(:Airport)`	Indicates airports to which an airline flies.
`(:Airport)-[:ROUTE {airline_id, codeshare, stops, equipment}]->(:Airport)`	Represents a directed flight route between airports.

Table 4: Summary of relationship types used in the Neo4j graph model.

2.2.3 Indexes and constraints

To ensure data integrity and efficient query performance, a series of constraints and indexes were defined before importing the data. Uniqueness constraints were established on key identifiers to prevent duplicate node creation. Additionally, indexes were created on frequently queried attributes to improve the speed and lookup of the patterns. For the full indexes and constrains we refer to the code provided with our report.

3 Methodology

This section presents the methods that we will use to perform tha analysis of the dataset. The analyses were carried out in *Neo4j Desktop*, which provides local access to the Graph Data

Science (GDS) and Awesome Procedures On Cypher (APOC) plug-ins. These plug-ins are not available in the browser version but are essential for performing advanced network algorithms directly on the graph. The GDS library provides optimized, in-memory implementations of classical graph-theoretic algorithms such as centrality measures, community detection, and connectivity analysis. The APOC library extends Cypher with additional procedures for data aggregation, subgraph projection, and transformation, allowing complex queries to be executed more efficiently. The combination of Cypher, GDS, and APOC thus enables both descriptive and algorithmic analyses within Neo4j, without the need for external processing.

3.1 Identifying global hubs among airports and airlines

This part of the analysis focuses on identifying the airports and airlines that act as global hubs within the air transport network. Airports are evaluated according to their position in the airport-airport network, while airlines are assessed according to on how many airports they operate and how dominant they are at these airports. The following will present some measures that we will use to draw conclusions of our analysis.

3.1.1 Degree

For airports, hub status is first quantified by the **degree** of a node. Given a directed graph $G = (V, E)$ with airport nodes V and route relationships E, the out-degree, in-degree, and total degree of each airport v_i are defined as

$$\deg^+(v_i) = \big|\{\, e_{ij} \in E \,\}\big|, \qquad \deg^-(v_i) = \big|\{\, e_{ji} \in E \,\}\big|, \qquad \deg(v_i) = \deg^+(v_i) + \deg^-(v_i). \quad (1)$$

These measures represent, respectively, the number of outgoing, incoming, and total connections of an airport to other airports. The degree captures the local connectivity of each node and provides a direct measure of its importance within the network.

3.1.2 PageRank

The degree of a node only reflects how many direct connections it has, which means it does not capture influence beyond its immediate neighbours. To include a broader view of importance in the network, we also compute the **PageRank** metric (Page et al., 1999). PageRank is based on the idea that a node is important if it is connected to other nodes that are themselves important. PageRank works as follows. For a damping factor $\alpha \in (0, 1)$, PageRank solves iteratively

$$\pi_i = \alpha \sum_{j \in N_i^-} \frac{\pi_j}{d_j} + \frac{1-\alpha}{n} \quad (2)$$

where π_i is the PageRank score of airport i, N_i^- is the set of airports that have flights into airport i, d_j is the number of outgoing routes from airport j and n is the total number of airports. Airports with high PageRank values are those connected to other highly connected airports, indicating high global importance. For our analysis, we use the default value of $\alpha = 0.85$ of the GDS library.

3.1.3 Airline Connectivity and International Reach

Next, we analyze the airlines instead of the airports. The airlines are connected to the airports via the `OPERATES_FROM` and `OPERATES_TO` relations, meaning we have a bipartite network where we can split the network into a disjoint set of nodes. This allows us to study how big and international each airline is, and how dominant certain airlines are at certain airports.

The previous analysis considered the route network $G = (V, E)$ consisting solely of airports and their interconnecting routes. To evaluate the operational behavior of airlines, we extend the representation to a bipartite graph

$$G' = (V_A, V_L, E'), \tag{3}$$

where V_A is the set of airport nodes, V_L the set of airline nodes, and E' the set of relationships linking airlines to airports. Edges in E' correspond to the `OPERATES_FROM` and `OPERATES_TO` relationships. By construction, $E' \subseteq (V_L \times V_A) \cup (V_A \times V_L)$ and no relationships exist between nodes of the same type. This two-mode structure makes it possible to study how widely each airline operates and how concentrated the traffic is at each airport.

For each airline $l \in V_L$, we compute three frequency-based measures. First, the number of distinct origin airports from which it operates,

$$O(l) = \big|\{\, a \in V_A : (l, a) \in R_{\text{from}} \,\}\big|, \tag{4}$$

and the number of distinct destination airports it serves,

$$D(l) = \big|\{\, a \in V_A : (l, a) \in R_{\text{to}} \,\}\big|. \tag{5}$$

Here, R_{from} and R_{to} denote the set of edges corresponding to the `OPERATES_FROM` and `OPERATES_TO` relationships respectively. The quantities $O(l)$ and $D(l)$ now reflect the size of the airline's operational network in terms of airport coverage.

A third measure, the number of countries reached through any operated airport, captures the international span of the airline. Let $C(a)$ denote the country associated with airport a. Then the set of countries served by airline l is

$$S(l) = \{\, C(a) : (l, a) \in E' \,\}, \tag{6}$$

and the corresponding country reach is

$$C(l) = |S(l)|. \tag{7}$$

Airlines with large $O(l)$, $D(l)$, and $C(l)$ values are globally active carriers with broad operational and geographical coverage, while low values indicate smaller or regionally focused airlines.

3.1.4 Airline Diversity and Dominance at Airports

To characterise the diversity and dominance of airlines at each airport, we invert the perspective and examine the set of airlines linked to each airport $a \in V_A$. The number of distinct airlines operating from or to airport a is given by

$$A(a) = \big|\{\, l \in V_L : (l, a) \in E' \,\}\big|. \tag{8}$$

A high value of $A(a)$ indicates a competitive hub served by many carriers, whereas a low value suggests limited competition or a single-carrier base.

To quantify concentration more precisely, we compute a dominance index (Borenstein, 1989) based on the share of the most active airline at each airport. Let $r_{l,a}$ denote the number of routes associated with airline l at airport a (both from and to the airport). Then the dominance index is defined as

$$H(a) = \max_{l} \frac{r_{l,a}}{\sum_{l'} r_{l',a}}. \tag{9}$$

Values of $H(a)$ close to one indicate that a single airline operates most of the airport's traffic, whereas values closer to zero correspond to airports shared among many carriers.

3.2 Geographical Distribution of Air Connectivity

Beyond the behaviour of individual airports and airlines, we also examine how air traffic is distributed across countries. This country-level perspective allows us to study whether global connectivity is concentrated in a small set of countries or more evenly shared, and how internationally connected each country is. To do so, we compute three indicators: the **top-k concentration ratio**, the **Gini coefficient**, and a measure of **international reach**.

3.2.1 Top-k concentration ratio

Let C denote the set of all countries in the dataset, and let M_c represent the number of outgoing routes from airports located in country $c \in C$. The total number of routes in the network is

$$M_{\text{total}} = \sum_{c \in C} M_c. \tag{10}$$

The share of routes controlled by country c is then

$$s_c = \frac{M_c}{M_{\text{total}}}. \tag{11}$$

To measure concentration, we compute the **top-k concentration ratio** (Bain, 1956), defined as the cumulative share of the k countries with the highest route counts:

$$\text{CR}_k = \sum_{i=1}^{k} s_{c_i}, \tag{12}$$

where $c_1, c_2, \ldots, c_k$ are the countries ranked in descending order by M_c. A high CR_k value indicates that a small number of countries dominate the network.

3.2.2 Gini coefficient

To complement the top-k concentration ratio, we compute the **Gini coefficient** (Gini, 1955). This coefficient is a standard measure of inequality ranging from 0 (perfect equality) to 1 (maximum inequality). Given a sorted list of route counts $M_{c_1} \leq M_{c_2} \leq \cdots \leq M_{c_n}$, the Gini coefficient is defined as

$$G = \frac{2 \sum_{i=1}^{n} i \cdot M_{c_i}}{n \sum_{i=1}^{n} M_{c_i}} - \frac{n+1}{n}. \tag{13}$$

3.2.3 International reach

Additionally, similar as for the airlines, we examine international reach. We do that by computing, for each country, the number of distinct foreign countries reachable through direct flights. Let D_c denote this set:

$$D_c = \{c' \in C : \exists (a_1, a_2) \in E,\ a_1 \in c,\ a_2 \in c',\ c \neq c'\}, \tag{14}$$

where $a \in c$ indicates that airport a is located in country c. The size $|D_c|$ quantifies the international connectivity/reach of country c.

3.3 Identifying Key Bridge Airports Linking Distant Regions

The final part of the analysis examines which airports act as bridges between different parts of the global air transport network. Instead of looking only at local connectivity or airline activity, this section considers how airports help link distant regions of the network. To study this, we rely on two measures: (i) **betweenness centrality**, which captures how often an airport lies on shortest paths between others, and (ii) **community detection**, which identifies groups of airports that are more closely connected to each other than to the rest of the network. Together, these measures help reveal airports that play a connecting role between otherwise separate regions.

3.3.1 Betweenness centrality

Betweenness centrality (Freeman, 1977) measures the extent to which a node facilitates shortest paths between other nodes in the network. For each airport v_i, the betweenness score is defined as

$$C_B(v_i) = \sum_{s \neq v_i \neq t} \frac{\sigma_{st}(v_i)}{\sigma_{st}}, \tag{15}$$

where σ_{st} denotes the total number of shortest paths between airports s and t, and $\sigma_{st}(v_i)$ the number of those paths that pass through v_i. The fraction $\sigma_{st}(v_i)/\sigma_{st}$ captures how important airport v_i is for efficiently connecting s and t.

Airports with high betweenness centrality values do not necessarily have many direct connections, but their position in the network structure makes them crucial for linking otherwise distant or weakly connected regions.

3.3.2 Community detection with the Louvain algorithm

To analyse the higher-level organisation of the network, we identify groups of airports that are more strongly linked internally than externally. These groups, or communities, often reflect geographical regions or market structures. We use the **Louvain algorithm** (Blondel et al., 2008), which assigns airports to communities by maximising *modularity*, defined as

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j), \tag{16}$$

where A_{ij} is the adjacency matrix, k_i and k_j are the degrees of airports i and j, m is the total number of edges, and $\delta(c_i, c_j)$ equals 1 if airports i and j belong to the same community and 0 otherwise. The modularity Q evaluates how much more interconnected the communities are compared to a random graph with the same degree distribution.

Maximizing this gives us the matrix A, which provides us with the groups/communities. By analyzing which airports connect the most communities together we get insight into which airports are the key to linking distant regions.

4 Results

This section presents the empirical findings derived from the global air transport network constructed in Neo4j. The analysis proceeds by examining the centrality structure of airports and airlines, the geographical distribution of air connectivity across countries, and the role of key airports in bridging distant network regions.

4.1 Airport and Airline Connectivity Patterns

Table 5 reports the top airports ranked by total degree, reflecting the number of distinct outbound and inbound routes. The results show a clear dominance of large intercontinental airports. Hartsfield-Jackson Atlanta International Airport exhibits the highest connectivity with 1,826 total connections, substantially exceeding all other airports. Other major U.S. airports such as Chicago O'Hare, Los Angeles and Dallas-Fort Worth appear prominently, indicating the scale of the domestic and international network in the United States. Major European airports such as London Heathrow, Charles de Gaulle, Frankfurt, and Amsterdam Schiphol also rank within the top ten, reflecting their long-standing role as gateways between Europe and other world regions. The near symmetry between inbound and outbound connections underscores that these hubs serve balanced flows rather directional traffic.

Table 5: Top 10 airports by total degree

Rank	Name	City	Country	Out	In	Total
1	Hartsfield-Jackson Atlanta Int. Airport	Atlanta	United States	915	911	1826
2	Chicago O'Hare International Airport	Chicago	United States	558	550	1108
3	Beijing Capital International Airport	Beijing	China	531	530	1061
4	London Heathrow Airport	London	United Kingdom	525	522	1047
5	Charles de Gaulle Int. Airport	Paris	France	524	517	1041
6	Frankfurt am Main Airport	Frankfurt	Germany	497	493	990
7	Los Angeles International Airport	Los Angeles	United States	489	497	986
8	Dallas-Fort Worth International Airport	Dallas–Fort Worth	United States	469	467	936
9	John F. Kennedy International Airport	New York	United States	456	455	911
10	Amsterdam Airport Schiphol	Amsterdam	Netherlands	453	450	903

The PageRank scores in Table 6 provide a complementary measure by accounting not only for the number of connections but also for the importance of the connected airports. The ranking largely agrees with the previous results, with Atlanta and Chicago O'Hare again at the top. Interestingly, airports such as Singapore Changi and Denver appear higher in the PageRank ranking than in the degree ranking. This indicates that their connections lead to especially important parts of the network, giving them more influence than their route counts alone would suggest. Because PageRank places more weight on links to other well-connected airports, it highlights the role of these transfer hubs whose strategic connections span several world regions.

Table 6: Top 10 airports by PageRank score

Rank	Name	City	Country	PageRank
1	Hartsfield-Jackson Atlanta Int. Airport	Atlanta	United States	30.96
2	Chicago O'Hare International Airport	Chicago	United States	19.46
3	Los Angeles International Airport	Los Angeles	United States	18.63
4	Dallas-Fort Worth International Airport	Dallas–Fort Worth	United States	17.89
5	Charles de Gaulle Int. Airport	Paris	France	16.40
6	London Heathrow Airport	London	United Kingdom	16.29
7	Singapore Changi Airport	Singapore	Singapore	15.92
8	Beijing Capital International Airport	Beijing	China	15.83
9	Denver International Airport	Denver	United States	15.58
10	Frankfurt am Main Airport	Frankfurt	Germany	14.99

Airline-level indicators in Tables 7 and 8 reveal two distinct forms of connectivity. When measured by the number of origin airports served, the ranking is led by American Airlines, United Airlines, and Delta Air Lines. This pattern reflects the size of the U.S. aviation market, where airlines operate from a very large number of airports across the country. In contrast, when looking at how many different countries airlines connect to, European and Middle Eastern carriers move to the top: Air France reaches 126 countries, Turkish Airlines 110, and KLM 97. These airlines depend heavily on long-distance transfer traffic and make effective use of their geographical position to link regions that would otherwise be difficult to reach directly. The contrast between Tables 7 and 8 therefore highlights the difference between serving many airports and achieving broad international coverage.

Table 7: Top 10 airlines by number of origin airports

Rank	Name	Country	Origins	Destinations
1	American Airlines	United States	428	406
2	United Airlines	United States	425	394
3	Air France	France	377	337
4	KLM Royal Dutch Airlines	Netherlands	360	351
5	US Airways	United States	348	337
6	Delta Air Lines	United States	347	351
7	Alitalia	Italy	269	271
8	Turkish Airlines	Turkey	254	254
9	Lufthansa	Germany	243	244
10	China Eastern Airlines	China	218	218

Table 8: Top 10 airlines by number of countries served

Rank	Airline	Country	Countries Served
1	Air France	France	126
2	Turkish Airlines	Turkey	110
3	KLM Royal Dutch Airlines	Netherlands	97
4	United Airlines	United States	97
5	Lufthansa	Germany	96
6	Qatar Airways	Qatar	90
7	British Airways	United Kingdom	83
8	Emirates	United Arab Emirates	83
9	Ethiopian Airlines	Ethiopia	82
10	Delta Air Lines	United States	72

Further insight into the structure of airport–airline relationships is provided by the dominance index in Table 9. Several airports display high levels of concentration, where a single airline operates a majority of all routes. London Stansted, for example, has a dominance score of 0.723 due to the extensive presence of Ryanair. Similar patterns are observed at Houston Intercontinental (United Airlines), Sheremetyevo (Aeroflot), and Beijing Capital (Air China). Even the world's busiest airport, Atlanta, exhibits notable concentration with Delta Air Lines responsible for 60% of all its routes. High dominance values show that some airports rely heavily on a single airline, meaning there is little competition and that the airport's connectivity can be strongly affected by that airline's decisions.

Table 9: Dominance index for the top 10 airports who belonged to the 100 airports with the highest PageRank

Rank	Airport	Country	Dominance	Total Routes	Dominant Airline (#Routes)
1	London Stansted Airport	United Kingdom	0.723	343	Ryanair (248)
2	George Bush Intercontinental Houston	United States	0.666	485	United Airlines (323)
3	Sheremetyevo International Airport	Russia	0.660	397	Aeroflot Russian Airlines (262)
4	Jorge Newbery Airpark	Argentina	0.648	128	Aerolineas Argentinas (83)
5	Salt Lake City International Airport	United States	0.633	278	Delta Air Lines (176)
6	Ninoy Aquino International Airport	Philippines	0.624	237	Philippine Airlines (148)
7	Phoenix Sky Harbor International Airport	United States	0.620	287	American Airlines (178)
8	Denver International Airport	United States	0.606	429	United Airlines (260)
9	Beijing Capital International Airport	China	0.602	510	Air China (307)
10	Hartsfield–Jackson Atlanta International	United States	0.600	588	Delta Air Lines (353)

The diversity of airline presence at major hubs is shown in Table 10, which lists airports by the number of carriers operating there. Unlike the dominance index, this indicator captures the extent of competition and international reach. Charles de Gaulle and Frankfurt host more than 100 airlines each, followed by large Asian and Middle Eastern hubs such as Bangkok Suvarnabhumi and Dubai International. These airports function as highly open, globally connected nodes that attract a broad mix of international carriers, enabling them to serve as major transfer points across regions.

Table 10: Top 10 airports by number of airlines operating

Rank	Airport	Country	Airlines
1	Charles de Gaulle International Airport	France	109
2	Frankfurt am Main Airport	Germany	100
3	Suvarnabhumi Airport	Thailand	98
4	Leonardo da Vinci–Fiumicino Airport	Italy	92
5	London Heathrow Airport	United Kingdom	86
6	Dubai International Airport	United Arab Emirates	83
7	Amsterdam Airport Schiphol	Netherlands	81
8	Munich Airport	Germany	76
9	Singapore Changi Airport	Singapore	75
10	Madrid–Barajas Airport	Spain	69

4.2 Distribution of Connectivity Across Countries

The spatial distribution of global air connectivity is highly uneven. Table 11 summarises the countries with the widest international reach. France leads with access to 115 foreign destinations, followed by the United Kingdom (99), Germany (97), and the United States (93). The United Arab Emirates, despite its small size, reaches 91 countries through a concentrated set of large hubs. These results show that a country's international connections depend on both its size and its location. Large countries tend to have big networks, while smaller countries with well-placed hubs can still achieve a strong global presence.

Table 11: Top 10 countries by international reach

Rank	Country	Countries Reached	Airports Used
1	France	115	41
2	United Kingdom	99	33
3	Germany	97	30
4	United States	93	65
5	United Arab Emirates	91	5
6	Netherlands	85	8
7	Spain	80	25
8	Italy	78	24
9	Turkey	72	15
10	Russia	69	38

Table 12 presents concentration ratios that quantify the distribution of air routes among countries using cumulative shares. The top 5 countries account for 43.3% of all international routes, while the top 10 account for 56.0%. This concentration rises to 70.1% for the top 20 countries and 86.3% for the top 50. The distribution is therefore highly skewed: a small number of countries dominate global connectivity, while the majority contribute only marginally.

Table 12: Concentration ratios for global air connectivity.

Metric	Value
Top 5 countries (CR_5)	43.3%
Top 10 countries (CR_{10})	56.0%
Top 20 countries (CR_{20})	70.1%
Top 50 countries (CR_{20})	86.3%

This inequality is further illustrated by the Lorenz curve in Figure 1. The curve lies far below the diagonal of perfect equality, and the resulting Gini coefficient of 0.8072 confirms the extreme concentration of air routes. This level of inequality is very high. In practice, the results show that global air connectivity is organised in a clear hierarchy: a small number of countries act as major gateways for international travel, while many others have far fewer connections and sit at the edge of the global network.

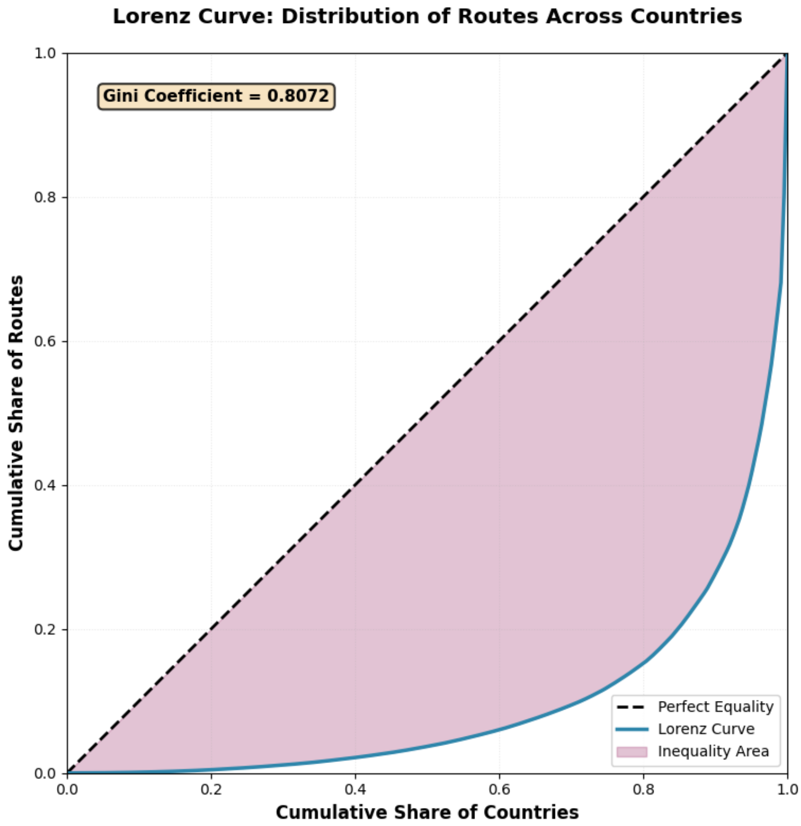

Figure 1: Lorenz curve showing the distribution of routes across countries. The shaded area between the curve and the diagonal represents inequality, quantified by the Gini coefficient.

4.3 Bridging Airports and Inter-Regional Connectivity

Betweenness centrality results in Table 13 identify airports that are most frequently located on shortest paths between other airports. These airports play a critical role as intermediaries in the global network. Los Angeles ranks first, reflecting its strategic position as a key connector between North America and the Asia-Pacific region. Major European hubs such as Paris Charles de Gaulle, London Heathrow, Frankfurt, and Amsterdam Schiphol also exhibit high betweenness scores, underscoring their function as intercontinental transfer points. Middle Eastern and Asian hubs (Dubai, Beijing Capital, Singapore Changi) likewise appear prominently, consistent with their emergence as global transit centres over the past two decades.

Table 13: Top 10 airports by betweenness score

Rank	Airport	Country	Betweenness
1	Los Angeles International Airport	United States	828,734
2	Charles de Gaulle International Airport	France	706,514
3	London Heathrow Airport	United Kingdom	567,244
4	Beijing Capital International Airport	China	551,905
5	Dubai International Airport	United Arab Emirates	539,624
6	Frankfurt am Main Airport	Germany	498,713
7	Amsterdam Airport Schiphol	Netherlands	472,118
8	Hartsfield–Jackson Atlanta Int. Airport	United States	463,055
9	Singapore Changi Airport	Singapore	451,338
10	Madrid–Barajas Airport	Spain	446,901

Table 14 examines how many network communities each airport connects, based on the Louvain community structure. Los Angeles again stands out by linking seven distinct communities, indicating its central role in bridging geographically distant clusters. Airports in the Gulf region (Dubai, Abu Dhabi, Doha) connect six communities, confirming their role as inter-regional hubs between Europe, Asia, and Africa. European hubs such as Heathrow, Charles de Gaulle, Frankfurt, and Schiphol also connect five communities each, consistent with their broad and diversified global reach. These findings complement the betweenness results by showing that bridging behaviour is not solely a function of traffic volume but also of an airport's location at the boundary of multiple structural clusters.

Table 14: Top bridge airports by number of connected communities

Rank	Airport	Country	IATA	Communities Connected
1	Los Angeles International Airport	United States	LAX	7
2	Dubai International Airport	United Arab Emirates	DXB	6
3	Abu Dhabi International Airport	United Arab Emirates	AUH	6
4	Hamad International Airport	Qatar	DOH	6
5	Oslo Lufthavn	Norway	OSL	5
6	Amsterdam Airport Schiphol	Netherlands	AMS	5
7	London Heathrow Airport	United Kingdom	LHR	5
8	Paris Charles de Gaulle Airport	France	CDG	5
9	Frankfurt am Main Airport	Germany	FRA	5
10	Istanbul Airport	Turkey	IST	5
11	Singapore Changi Airport	Singapore	SIN	4
12	Kuala Lumpur International Airport	Malaysia	KUL	4
13	Madrid Barajas Airport	Spain	MAD	4
14	Bangkok Suvarnabhumi Airport	Thailand	BKK	4
15	Toronto Pearson International Airport	Canada	YYZ	4

Overall, the results reveal a global air transport network that is highly centralized and geographically uneven. Only a few airports and airlines carry most of the global traffic, and most international routes are concentrated in a small group of countries. In addition, a limited number of key airports act as bridges between regions by connecting areas that would otherwise have very few direct links.

5 Conclusion

The analysis of the OpenFlights dataset reveals that the global air transport network is highly centralised and unevenly structured. A small group of major airports, primarily in North America, Europe, and Asia, concentrate much of the world's connectivity, appearing consistently at the top of both degree-based and PageRank rankings. These airports serve as major gateways for international travel and play a crucial role in shaping global mobility patterns.

Airline behavior shows a similar imbalance. A few large carriers operate from an exceptionally wide range of airports, while others stand out for their extensive international reach despite having smaller networks. Several airports also exhibit strong dominance by a single airline, indicating limited competition and a heavy dependence on anchor carriers. At the country level, connectivity is clearly skewed: only a handful of countries account for the majority of international routes, as shown by the high concentration ratios and the pronounced Lorenz curve.

The results also highlight the importance of certain airports as bridges between distant regions. Airports such as Los Angeles, Dubai, Singapore, and major European hubs connect multiple network communities and frequently lie on shortest paths between other airports. These locations play a key role in keeping the global network cohesive by linking areas that would otherwise be only loosely connected.

Taken together, the findings illustrate a global air transport system characterised by strong hierarchy, unequal distribution of connectivity, and a reliance on a relatively small set of strategic hubs. These structural features have implications for accessibility, international integration, and the resilience of the network in the face of disruptions.

References

Bain, J. S. (1956). *Barriers to New Competition*. Harvard University Press.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

Borenstein, S. (1989). Hubs and high fares: Dominance and market power in the us airline industry. *The RAND Journal of Economics*, 20(3):344–365.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, 40(1):35–41.

Gini, C. (1955). Variability and mutability. In *Memorie di metodologica statistica*. Libreria Eredi Virgilio Veschi, Rome.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. In *Technical Report*. Stanford InfoLab.

Robinson, I., Webber, J., and Eifrem, E. (2015). *Graph Databases: New Opportunities for Connected Data*. O'Reilly Media.

Sadalage, P. J. and Fowler, M. (2012). *NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence*. Addison-Wesley.