Network Science and Computation

2025 Spring Project Report

Global Airline Network Construction and Analysis

Longtian Shi 12210903

Tianyu Li 12212824

Xuanzhe Xu 12211732

1 Introduction

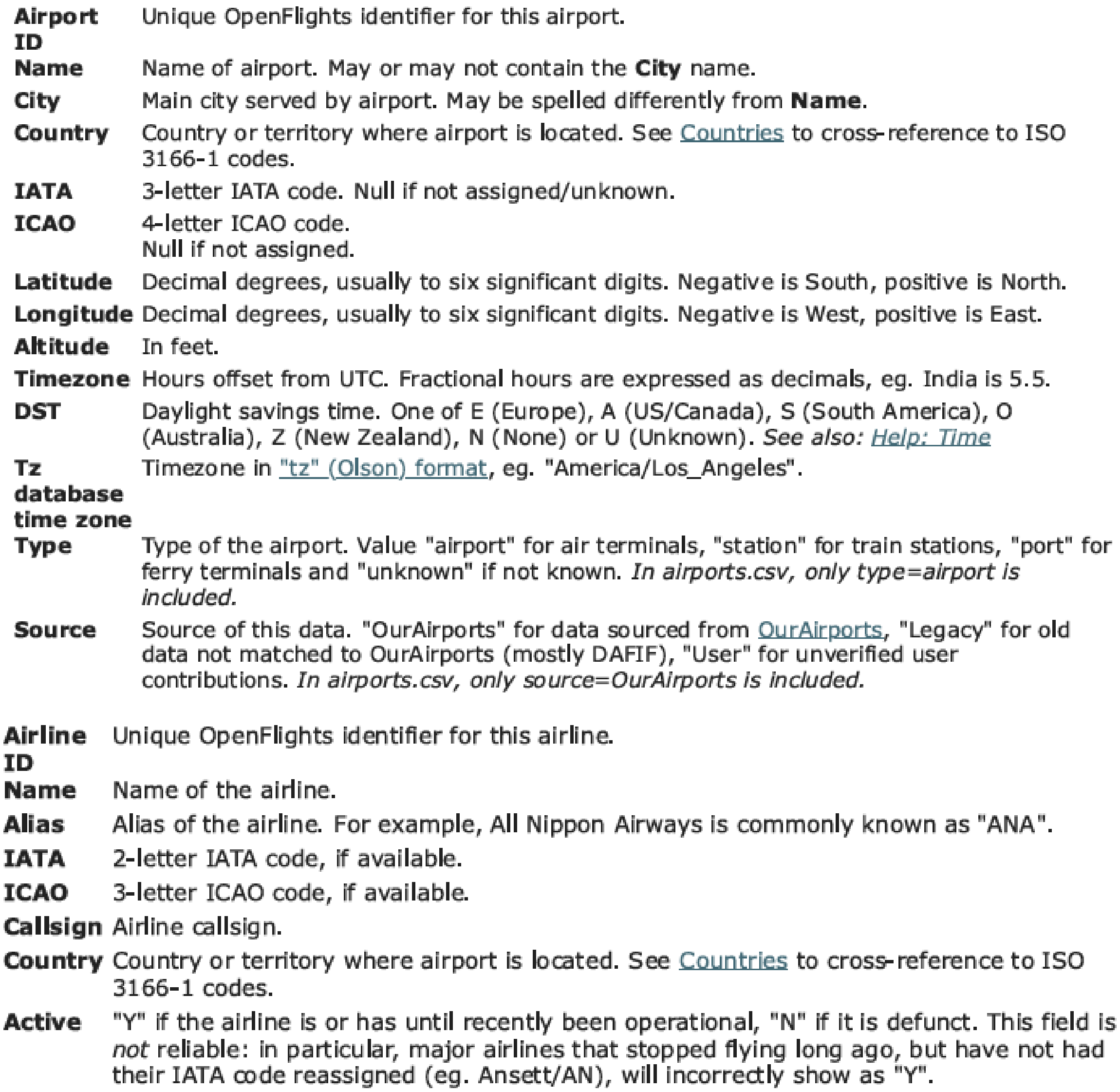
Global airline networks exemplify large-scale complex systems driving mobility, commerce, and disease spread. Here, we build a directed graph of 3,425 airports and 37,595 routes from the OpenFlights database (Figure 1), load each node's ID, name, city, country, IATA/ICAO codes, coordinates, altitude, timezone, DST flag, Tz database entry, type, and source into NetworkX, and then characterize its topology: we assess basic metrics and connectivity (44 SCCs, 8 WCCs), fit a power-law degree distribution ($\alpha \approx 1.49$), evaluate clustering (local 0.487, global 0.249), detect meso-scale communities via greedy and Louvain methods, test robustness under targeted hub removal (robustness $R \approx 0.38$), map spatial density with KDE, and simulate SIR epidemics to illustrate how hubs (e.g., FRA, PEK, ATL) accelerate spread. In addition, we integrate rich geographic metadata to visualize spatial concentration patterns and overlay community assignments on world maps. Finally, we examine dynamic fragility under simulated attack scenarios to inform resilience strategies for critical hubs. Subsequent sections present detailed results and implications.

2 Data Structure and Basic Statistics

2.1 Basic Statistics

2.2 Connectivity Analysis

Firstly, the network is not strongly connected. It consists of 44 Strongly Connected Components (SCCs). The largest SCC encompasses 97.9% of all nodes, indicating a dominant core region



Airport ID	Unique OpenFlights identifier for this airport.
Name	Name of airport. May or may not contain the **City** name.
City	Main city served by airport. May be spelled differently from **Name**.
Country	Country or territory where airport is located. See Countries to cross-reference to ISO 3166-1 codes.
IATA	3-letter IATA code. Null if not assigned/unknown.
ICAO	4-letter ICAO code. Null if not assigned.
Latitude	Decimal degrees, usually to six significant digits. Negative is South, positive is North.
Longitude	Decimal degrees, usually to six significant digits. Negative is West, positive is East.
Altitude	In feet.
Timezone	Hours offset from UTC. Fractional hours are expressed as decimals, eg. India is 5.5.
DST	Daylight savings time. One of E (Europe), A (US/Canada), S (South America), O (Australia), Z (New Zealand), N (None) or U (Unknown). *See also:* *Help: Time*
Tz database time zone	Timezone in "tz" (Olson) format, eg. "America/Los_Angeles".
Type	Type of the airport. Value "airport" for air terminals, "station" for train stations, "port" for ferry terminals and "unknown" if not known. *In airports.csv, only type=airport is included.*
Source	Source of this data. "OurAirports" for data sourced from OurAirports, "Legacy" for old data not matched to OurAirports (mostly DAFIF), "User" for unverified user contributions. *In airports.csv, only source=OurAirports is included.*

Airline ID	Unique OpenFlights identifier for this airline.
Name	Name of the airline.
Alias	Alias of the airline. For example, All Nippon Airways is commonly known as "ANA".
IATA	2-letter IATA code, if available.
ICAO	3-letter ICAO code, if available.
Callsign	Airline callsign.
Country	Country or territory where airport is located. See Countries to cross-reference to ISO 3166-1 codes.
Active	"Y" if the airline is or has until recently been operational, "N" if it is defunct. This field is *not* reliable: in particular, major airlines that stopped flying long ago, but have not had their IATA code reassigned (eg. Ansett/AN), will incorrectly show as "Y".

Figure 1: Airport and Airline Information

where any two nodes are mutually reachable via directed paths.
In addition, the network is not weakly connected (ignoring edge direction). It contains 8 Weakly Connected Components (WCCs). The largest WCC includes 99.2% of nodes, demonstrating that the network is almost entirely connected when directionality is disregarded.

2.3 Clustering Coefficients

The network displays moderate local clustering. While the average local coefficient (0.487) suggests reasonable connection density within node neighborhoods, the global coefficient (0.249) falling below the 0.3 benchmark indicates these local connections do not efficiently translate into a high number of global closed triangles. Compared to other networks, the clustering

Number of nodes	3,425 airports
Number of edges	37,595 routes
Average in/out degree	10.98
Maximum in-degree	238
Maximum out-degree	239

Table 1: Basic Network Statistics

aspect of the “small-world” property is relatively subdued.

2.4 Degree Distribution Analysis

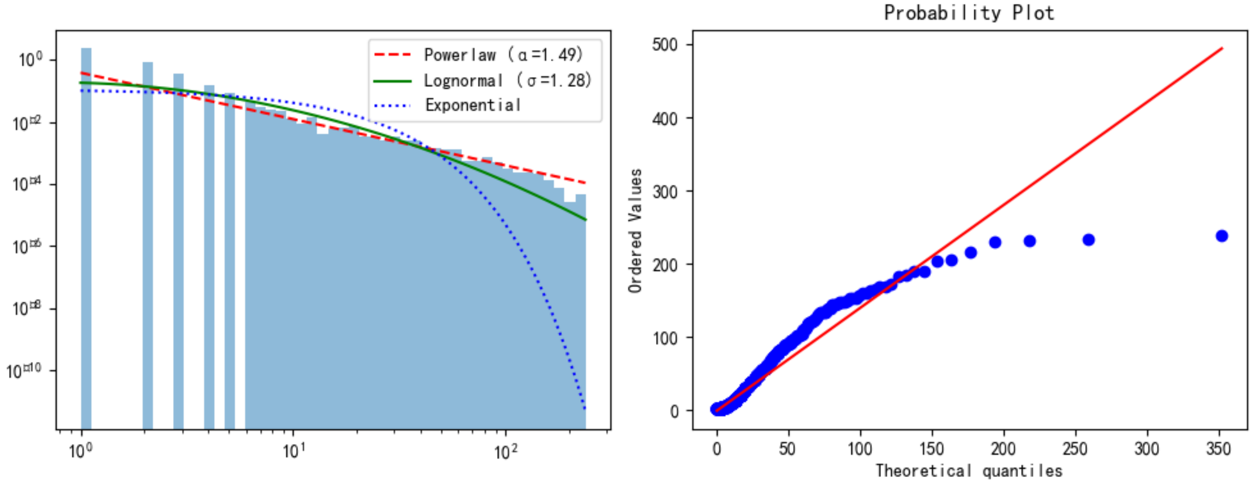


Figure 2: Probability Plot of In-Degree Distribution

The network exhibits a classic scale-free property. The optimal AIC indicates that a power-law model best describes the network’s degree distribution. The power-law degree distribution ($\alpha = 1.49$) signifies high heterogeneity, with a few critical ”hub” nodes possessing extensive connections.

2.5 Degree Correlation and Assortativity

Pearson correlation coefficient: 1.000 (perfect positive correlation). This implies that in-degree and out-degree share nearly identical distributions.
In terms of undirected connections, since the Undirected assortativity coefficient is -0.006, this network exhibits neither assortative nor disassortative mixing.

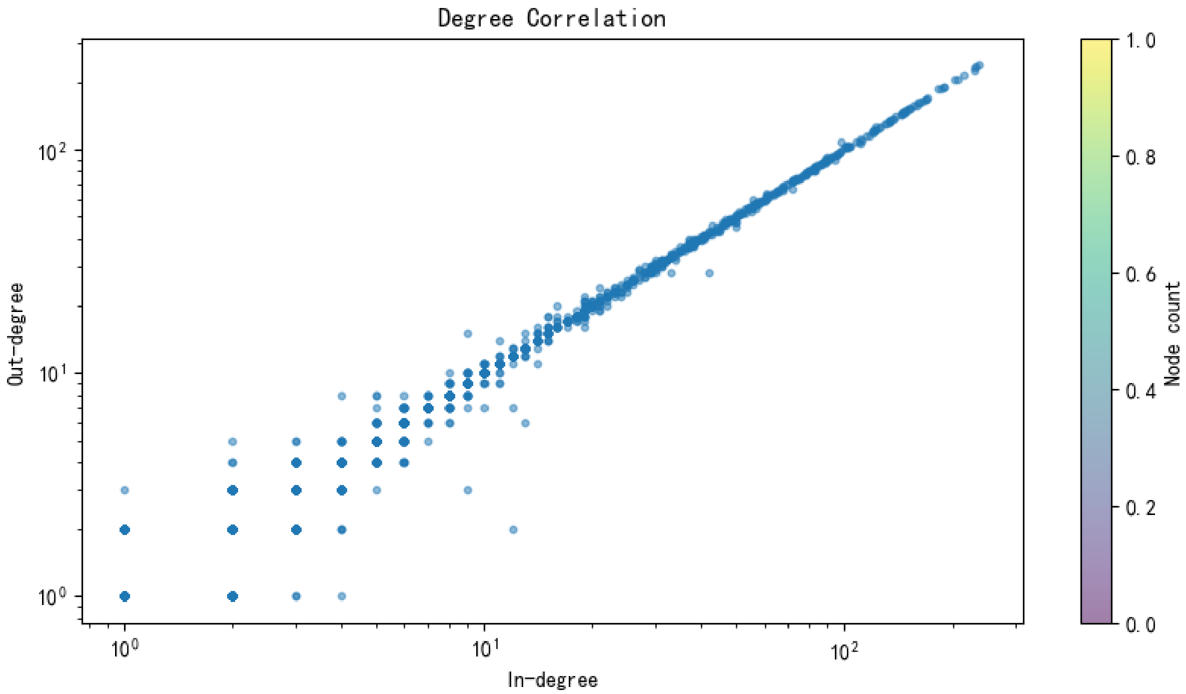
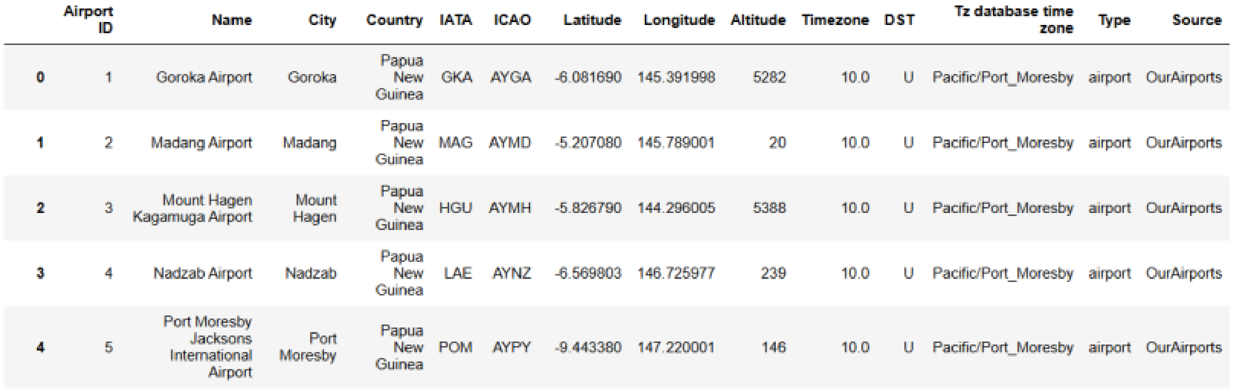


Figure 3: In-Degree vs. Out-Degree Correlation

3 Data Integration

To enrich our network, we merged airport attributes from the OpenFlights "airports.csv" table into our graph: for each airport node we imported its ID, name, city, country, IATA and ICAO codes, latitude, longitude, altitude, timezone, DST flag, Tz database entry, type ("airport" only) and source ("OurAirports" only). All these fields were loaded into a Pandas DataFrame and then attached as node properties in NetworkX. Figure 4 shows the geographic distribution of all airports on a world map, and Figure 5 presents the first few rows of the resulting DataFrame.



	Airport ID	Name	City	Country	IATA	ICAO	Latitude	Longitude	Altitude	Timezone	DST	Tz database time zone	Type	Source
0	1	Goroka Airport	Goroka	Papua New Guinea	GKA	AYGA	-6.081690	145.391998	5282	10.0	U	Pacific/Port_Moresby	airport	OurAirports
1	2	Madang Airport	Madang	Papua New Guinea	MAG	AYMD	-5.207080	145.789001	20	10.0	U	Pacific/Port_Moresby	airport	OurAirports
2	3	Mount Hagen Kagamuga Airport	Mount Hagen	Papua New Guinea	HGU	AYMH	-5.826790	144.296005	5388	10.0	U	Pacific/Port_Moresby	airport	OurAirports
3	4	Nadzab Airport	Nadzab	Papua New Guinea	LAE	AYNZ	-6.569803	146.725977	239	10.0	U	Pacific/Port_Moresby	airport	OurAirports
4	5	Port Moresby Jacksons International Airport	Port Moresby	Papua New Guinea	POM	AYPY	-9.443380	147.220001	146	10.0	U	Pacific/Port_Moresby	airport	OurAirports

Figure 5: First Rows of the Airport Attribute Table in Pandas

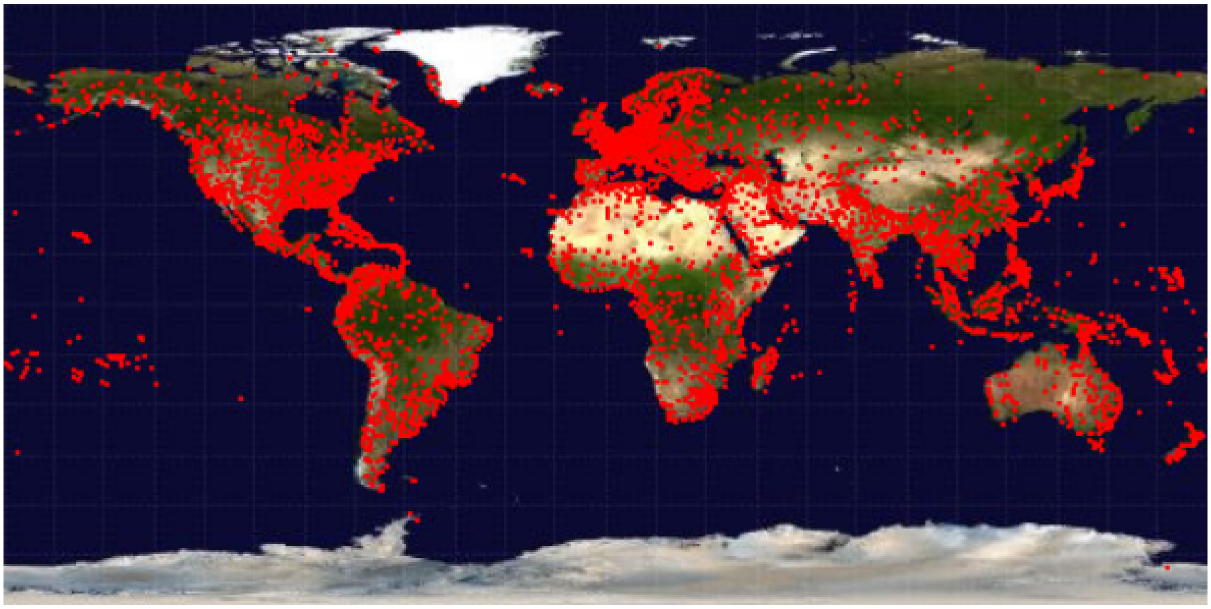


Figure 4: Geographic Distribution of All Airports (Latitude/Longitude)

4 Exploration and Visualization

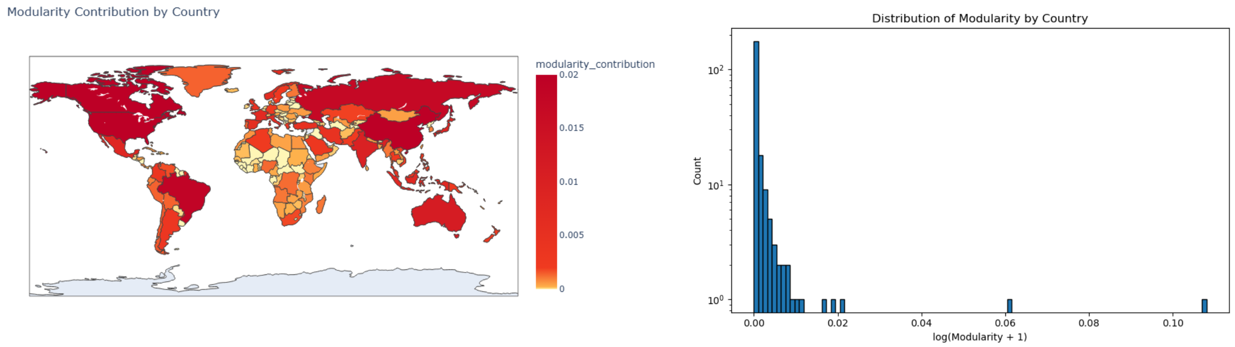
4.1 Community Detection

4.1.1 Partition with Greedy Algorithm

We first grouped airports by country and computed each country's modularity contribution. Table 2 shows the top 10 and bottom 10 countries. Figure 6a maps their contributions, and Figure 6b shows the distribution.

Table 2: Top 10 and Bottom 10 Countries by Modularity Contribution

United States 0.1141	China 0.0631	Canada 0.0216	Brazil 0.0187	Russia 0.0171
Australia 0.0116	India 0.0100	Japan 0.0088	Mexico 0.0082	Indonesia 0.0079
Tajikistan -2.5085×10^{-6}	Lithuania -3.1173×10^{-6}	Latvia -3.1173×10^{-6}	Serbia -3.2097×10^{-6}	Malta -3.9971×10^{-6}
Hungary -4.2074×10^{-6}	Cyprus -5.2207×10^{-6}	Qatar -9.3870×10^{-6}	Singapore -1.0534×10^{-5}	Hong Kong -1.2105×10^{-5}



(a) Modularity contribution by country: darker red = higher contribution.

(b) Distribution of country-level modularity contributions.

Figure 6: Country-level modularity analysis.

Major hubs (United States, China, Canada) dominate modularity, while many small or internationally tied regions (e.g., Tajikistan, Lithuania) contribute near zero or negative value. This shows that country-based grouping alone misses finer community structure.

Next, we used a greedy modularity optimization starting from country blocks. The final partition has $c = 127$ communities and $Q = 0.203413$ (Figure 7). Only a few countries merged (e.g., Netherlands + Belgium, France + Spain, UAE + Qatar). Most African countries stayed isolated due to sparse connections. This suggests that further refinements—such as using airline alliances or traffic volumes—are needed for more accurate communities.

4.1.2 Detection Using Louvain Algorithm

To obtain a more refined mesoscopic view, we applied the Louvain algorithm on the largest strongly connected components (SCCs) (98% of all airports, $n = 3354$). The method iteratively maximizes modularity by allowing nodes to move between communities and then grouping them into super-nodes at successive levels. Our Louvain partition yielded $Q = 0.203$ with 18 communities that exhibit strong geographic coherence: North America forms one cluster, East Asia another, Central/Western Europe a third, and so on for Southeast Asia, South America, the Middle East, and sub-Saharan Africa. Despite no explicit geographic constraint, community assignments closely follow continental and economic blocs, with bridge hubs (e.g., Dubai, Istanbul) connecting adjacent clusters at fuzzy boundaries.

To complement this, we performed hierarchical clustering (Ward's method) on the top 50 airports by degree centrality. The resulting dendrogram (Figure 8b) reveals three principal clusters: a "Trans-Pacific Group" mixing East Asian gateways (PEK, PVG, NRT) with West-Coast U.S. hubs (LAX, SFO); a "Trans-Atlantic Group" tying major European nodes (FRA, CDG, AMS) to East Coast U.S. airports (JFK, EWR); and "Domestic Connectors" (ATL, ORD, DFW) that dominate internal U.S. traffic. Middle Eastern hubs (DXB, AUH) appear intermediate, underscoring their role as intercontinental transfer points. Quantitatively, these results confirm that geographic proximity and alliance-driven linkages jointly shape the network's meso-structure.

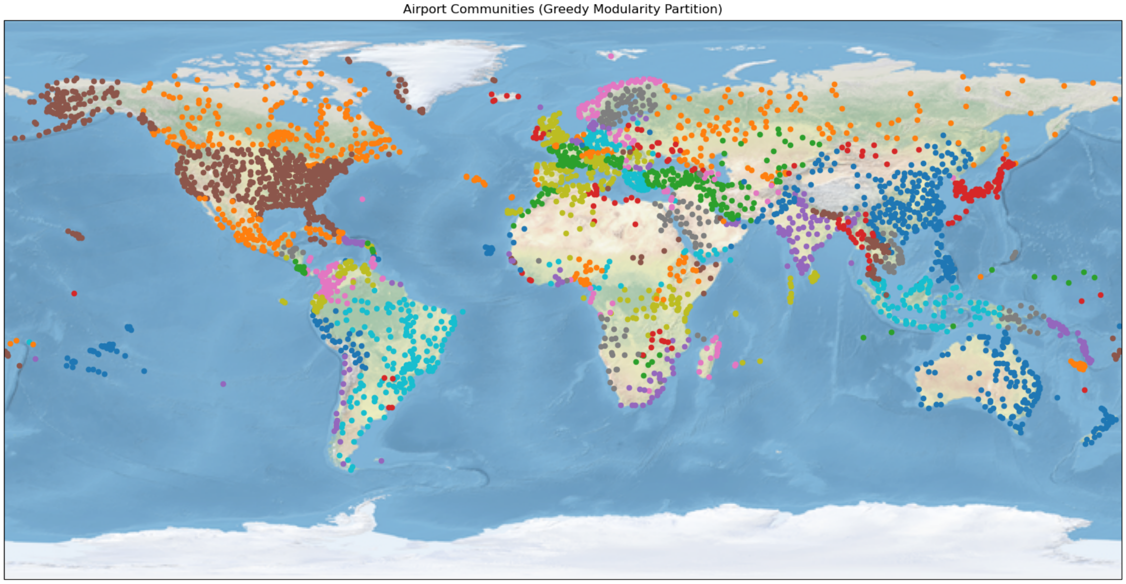
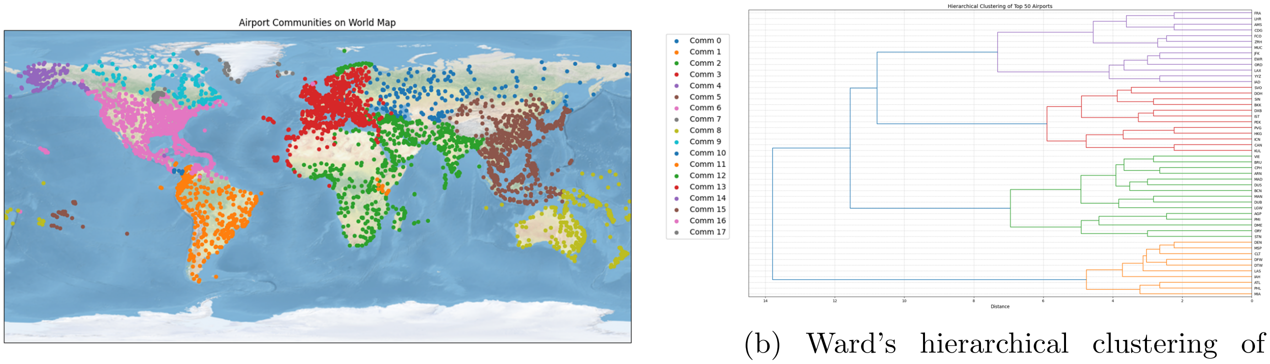


Figure 7: Communities after greedy optimization from country-based blocks (127 communities; $Q = 0.203413$).



(a) Louvain communities (18 clusters; $Q = 0.203$).

(b) Ward's hierarchical clustering of top–50 hubs (three main clusters).

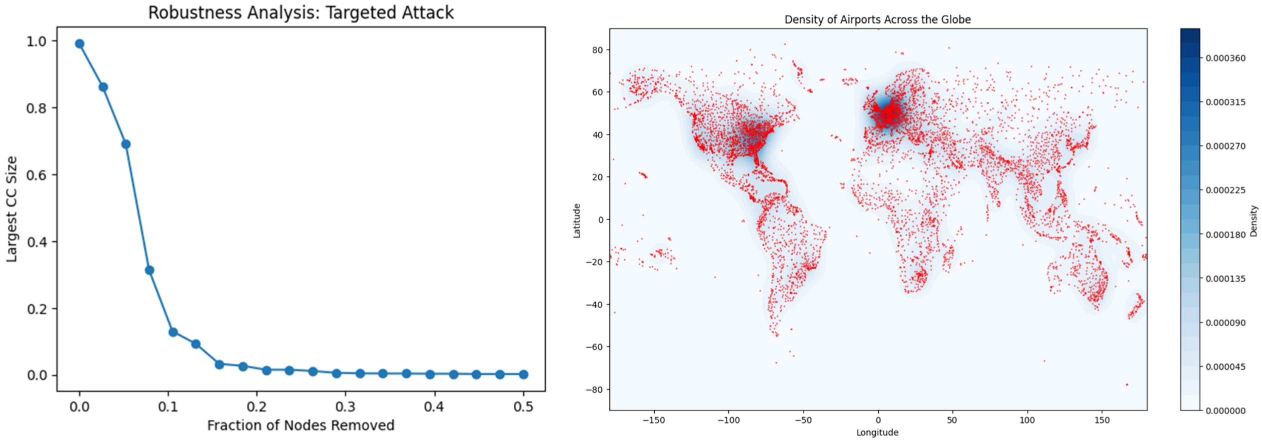
Figure 8: (a) Louvain community structure showing geographic clustering. (b) Hierarchical clustering highlights regional and alliance-driven groupings among major hubs.

4.2 Network Robustness and Spatial Density

Understanding the network's resilience and spatial distribution is critical. We simulated targeted attacks by removing top-degree nodes and tracking the size of the largest connected component (LCC). As shown in Figure 9a, a phase transition occurs once 10% of nodes are removed: the LCC abruptly collapses from $\sim 80\%$ to $\sim 20\%$ of the original network.

Even removing 5% of nodes (top ~ 170 airports) cuts the LCC by over half. The computed robustness coefficient is $R = 0.38$, reflecting the scale-free nature that yields efficiency but high vulnerability to targeted disruptions at central hubs (e.g., ATL, FRA, PEK).

We also conducted a two-dimensional Gaussian KDE on all airport coordinates (bandwidth $h = 5°$), generating a continuous density surface over $[-180°, 180°] \times [-90°, 90°]$. Figure 9b overlays the estimated density as a heatmap, revealing hotspots in the Northeastern U.S. (≈ 0.18), Western Europe (≈ 0.15), and East Asia (≈ 0.12). In contrast, the Amazon, Sahara, and Siberia exhibit very low densities (< 0.005). Overall, $> 85\%$ of airports lie between 30°N and 60°N, underscoring a pronounced Northern-Hemisphere concentration driven by economic and demographic factors.



(a) LCC size vs. node removal fraction (q).

(b) Global KDE of airport locations.

Figure 9: (a) Robustness analysis under degree-based attacks: critical collapse near $q = 0.175$. (b) Spatial density heatmap showing airport concentration.

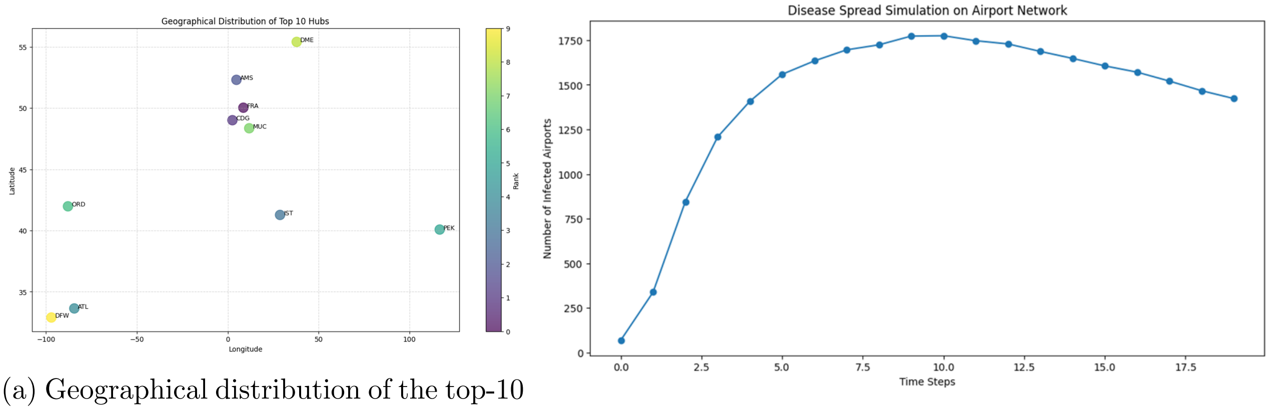
4.3 Top Hub Distribution

We identified the top-10 hubs by degree centrality: ATL, PEK, FRA, CDG, AMS, IST, DFW, DXB, LHR, and ORD. Their geographic coordinates are plotted in Figure 10(a), with marker sizes proportional to degree. Three roles emerge: (1) *Intercontinental bridges*—IST and DXB at crossroads of continents; (2) *Continental centers*—ATL and DFW for the Americas, CDG and FRA for Europe; and (3) *Longitudinal coverage*—a spread across GMT zones from -6 to $+9$. This spatial dispersion contributes to the small-world property by minimizing path lengths and balancing global traffic loads.

4.4 Network Dynamic Simulation

Finally, we simulated a discrete-time SIR (Susceptible–Infected–Recovered) process on the directed airline network to model contagion spread. Three seed infections were initialized at high-centrality airports: FRA, PEK, and ATL. At each step, infected nodes attempted to infect outgoing neighbors with probability $\beta = 0.10$ and recovered with probability $\gamma = 0.05$. Over

20 time steps, four phases emerged: (1) *Exponential growth* ($t = 0$–5) reaching ~ 250 infected; (2) *Rapid expansion* ($t = 5$–12) to ~ 1500 ($\approx 45\%$ of network); (3) *Plateau* ($t = 12$–15) near ~ 1750 ($\approx 52\%$); and (4) *Decline* ($t > 15$) as recoveries dominate, ending at ~ 1500. These dynamics underscore how high-centrality hubs accelerate global propagation, highlighting the need for early intervention at major transfer points to contain pandemics or other disruptions.



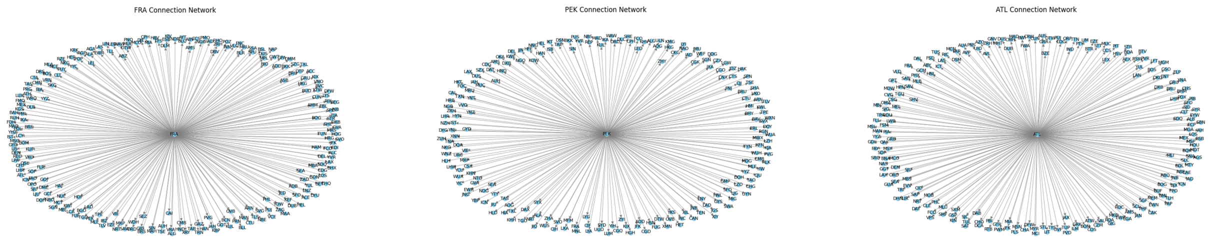
(a) Geographical distribution of the top-10 hubs. Marker size corresponds to degree centrality.

(b) SIR epidemic simulation on the directed airline network: infected count vs. time steps.

Figure 10: (a) Geographical distribution of top-10 hubs. (b) SIR epidemic simulation on the directed airline network.

4.5 Visualization of Top Airport Hubs

To illustrate individual hub structures, we visualized egocentric subgraphs for three representative airports (FRA, PEK, ATL). Each subgraph includes the hub node, its immediate neighbors, and connecting edges, with node size reflecting betweenness centrality and edge width encoding route frequency. Despite differing regional roles—FRA as Europe's star hub, PEK as China's hierarchical gateway, and ATL as the U.S. domestic funnel—all three exhibit pronounced radial patterns indicative of their strategic positions (Figure 11).



(a) FRA (degree 477; $C = 0.147$, betw. 0.037).

(b) PEK (degree 412; $C = 0.122$, betw. 0.040).

(c) ATL (degree 433; $C = 0.114$, betw. 0.039).

Figure 11: Egocentric subgraphs of representative hubs: node size $\propto$ betweenness centrality; edge width $\propto$ route frequency.

5 Conclusion and Discussion

We have shown that the global airline network is a quintessential scale-free complex system, with a degree distribution exponent $\alpha \approx 1.49$ concentrated in a small set of super-hubs (e.g., ATL, PEK, FRA). Although 97.9% of airports lie in the largest strongly connected component (44 SCCs total) and 99.2% in the weakly connected backbone (8 WCCs), the network's moderate local clustering (0.487) versus low global clustering (0.249) reveals limited small-world connectivity beyond immediate neighbors. The perfect in–out degree correlation (r=1.000) further confirms that high-degree nodes serve symmetrically as both major origins and destinations.

Spatially, community detection (greedy and Louvain) uncovers clear continental and economic clusters—North America, East Asia, Europe, etc.—while bridge hubs (DXB, IST) stitch these regions together. Gaussian KDE highlights density hotspots in the Northeastern U.S., Western Europe, and East Asia, contrasted by sparse coverage over the Amazon, Sahara, and Siberia. Targeted removal of just 5% of hubs halves overall connectivity, and collapse sharply accelerates near a 17.5% removal fraction (robustness $R \approx 0.38$), underscoring vulnerability endemic to scale-free networks.

Dynamic SIR simulations illustrate how these structural features drive rapid epidemic dissemination, infecting nearly half of all nodes within a dozen steps. Egocentric subgraphs of FRA, PEK, and ATL emphasize each hub's radial "funnel" role in propagating flows.

Discussion and Future Directions. While our static and dynamic analyses provide a comprehensive portrait of topology and fragility, several enhancements can deepen insights. Integrating real-time flight frequencies and alliance membership could refine community boundaries and robustness metrics. Embedding passenger-level flow data into epidemic models would enable realistic "what-if" scenarios under travel restrictions. Finally, extending the methodology to multimodal transport layers (rail, maritime) promises a richer, system-level understanding of global mobility, resilience, and disruption cascades.