Indian Institute of Technology, Gandhinagar

ES 404: Networks and Complex Systems

Instructor - Proff. Udit Bhatia

Analysis of the US Aviation Network

Submitted by:

Ekansh Raghav
Roll No: 22110080

Task 1: Network Construction & Preliminaries

Dataset Description

- **What does each node represent?**
  (Describe the representation of nodes in the network.)
- **What do the edges signify?**
  (Explain the meaning of edges and their properties.)
- **Where/how did you acquire this data?**
  (Provide the source and method of data collection.)

Construct Your Network

Describe any data cleaning/preprocessing steps before building the adjacency list/-matrix:

- Remove duplicate and inconsistent entries.
- Handle missing or incomplete data.
- Normalize node identifiers.
- Construct an adjacency list/matrix for analysis.

Initial Observations

- **Number of nodes (N):** (Insert value)
- **Number of edges (E):** (Insert value)

Solution

This project aims to deepen our understanding of key network measures by applying them to a real-world transportation network. We use the US airport network from 2010, originally compiled from Bureau of Transportation Statistics (BTS) data and later refined by Opsahl.

1. Dataset Description

Nodes

Each node represents an airport. The provided dataset uses airport codes that have been converted to unique IDs. In a more detailed study, you may also join these IDs with metadata from `USairport_2010_codes.txt` to recover airport names, locations, or other attributes.

Edges

Each weighted, undirected edge corresponds to a connection (route) between two airports. In the BTS version, edges' weights represent the number of passengers flown. In the refined dataset, when multiple flights exist between the same pair of airports, the weights are summed. Ties of weight zero (cargo-only flights) and self-loops are removed.

Data Acquisition

The data was downloaded from the BTS Transtats site (filtered for 2010 with all months and for the full US geography) and later processed into a network format (tnet/UCINET). Further details about data collection can be found in the blog post *"Why Anchorage is not (that) important: Binary ties and Sample selection."*

2. Constructing the Network

Data Cleaning and Pre processing Steps

Importing Data: Read in the edge list (and separately, the airport codes) into the Python programming language using pandas).

Before constructing the network, data cleaning and preprocessing are essential to ensure accuracy:

Data Cleaning: This involves removing duplicate records, correcting errors, and handling missing values

Cleaning: We follow these steps

- Remove edges with zero weight (indicating cargo-only routes).
- Remove any self-loops where the same airport appears as both origin and destination.
- Ensure that duplicated connections (multiple flights along the same route) are summed.

Adjacency Structure

Build an appropriate graph representation (e.g., using NetworkX's Graph/DiGraph object undirected networks; for weighted edges, use weighted graphs).

US Airport Network

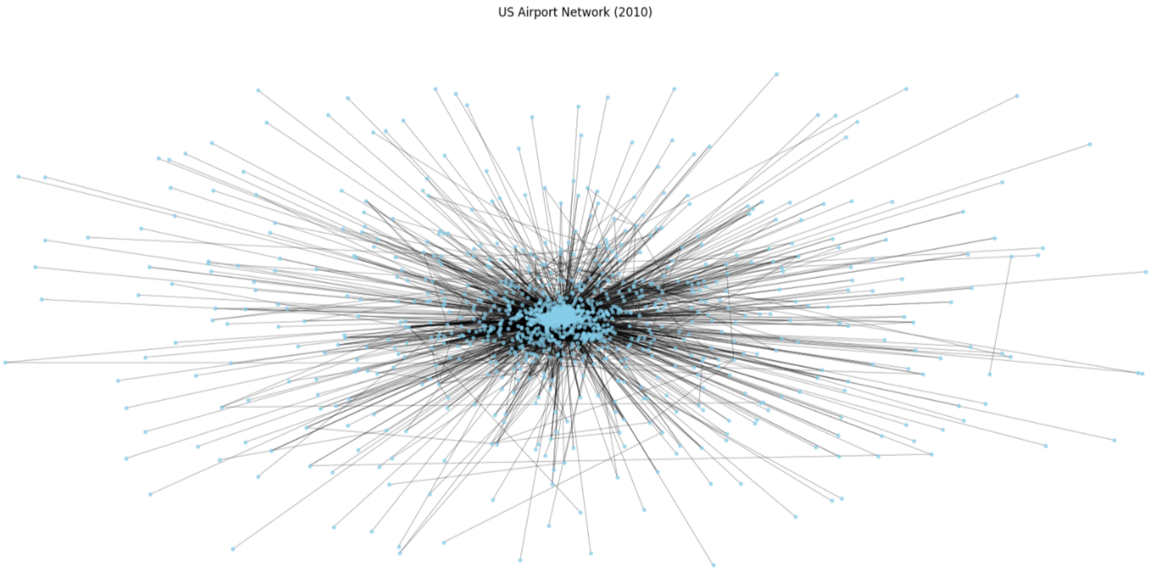

Figure 1: US Airport Network

3. Initial Observations

Number of Nodes (N) and Edges (E): Determining these provides a basic understanding of the network's size and complexity. A higher number of edges relative to nodes may indicate a densely connected network, while a lower ratio suggests sparsity.

Results

After running the analysis, we obtain the following results:

- Number of Nodes (N): After cleaning, our network contains 1,574 nodes (airports).
- Number of Edges (E): The network features 17,215 edges, representing the connections between these airports.

Degree Distribution: Plotting the degree distribution reveals how connections are spread across nodes. In many real-world networks, this distribution exhibits a heavy-tailed pattern, where most nodes have few connections, and a few nodes have a large number of connections. This indicates the presence of hub nodes that play a central role in the network's connectivity.

Degree distribution (log-log)

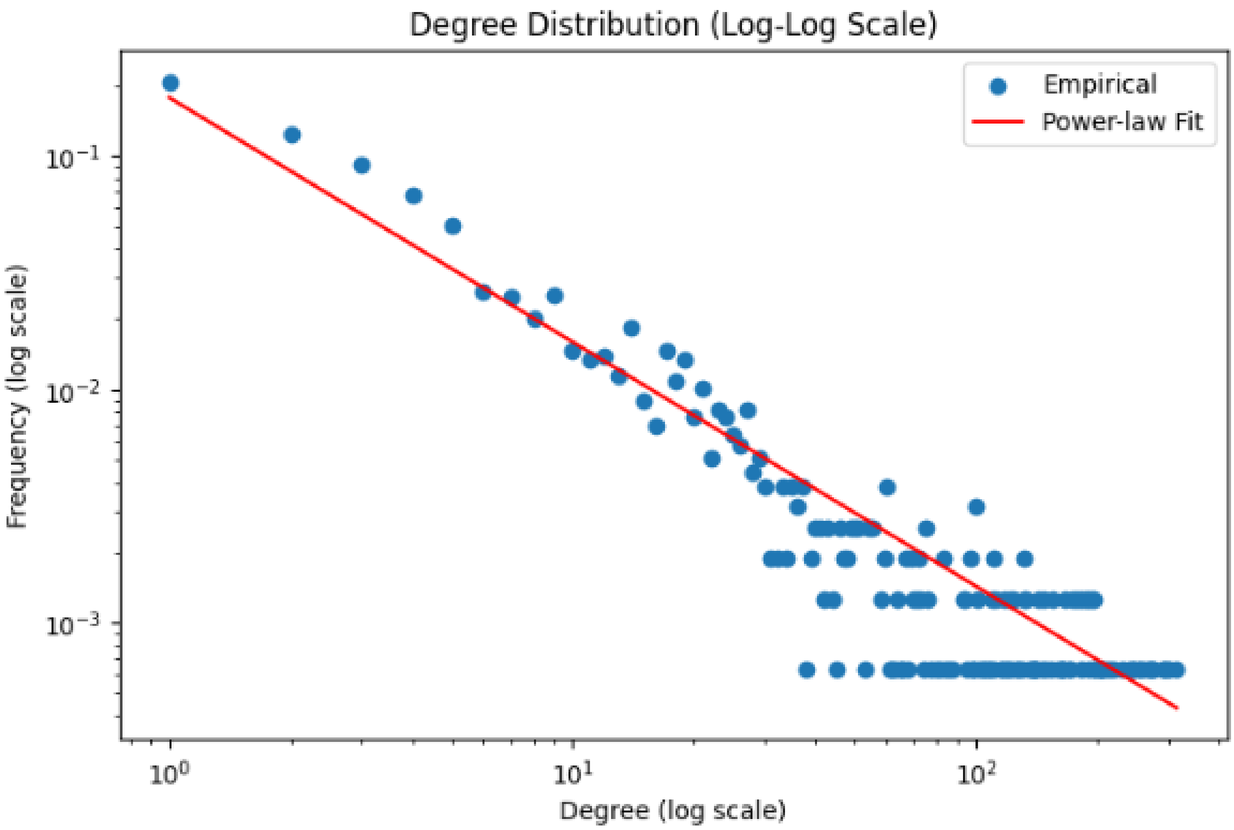

Figure 2: Degree distribution

Skewness and Heavy Tails: A right-skewed (positive skewness) degree distribution suggests that while most nodes have a low degree, there are significant outliers with exceptionally high degrees. This heavy-tailed nature is characteristic of scale-free networks, where certain nodes (hubs) are highly connected, influencing the network's robustness and vulnerability.

The fitted power-law coefficient (slope of the log-log plot) provides an indication of the level of inequality in connectivity; a steeper slope implies a sharper drop-off in the number of high-degree nodes.

Task 2: Centrality Analysis

Centrality metrics are used to identify the most influential or "central" nodes within the network. Different measures capture various aspects of centrality.

2.1 Eigenvector Centrality

Theory: Eigenvector centrality assigns scores to nodes based not only on the number of connections a node has (its degree) but also the quality (centrality) of those connections. It reflects the idea that connections to high-scoring nodes contribute more than connections to low-scoring nodes.

Implementation: After implementing the code, we obtain the Top 10 airports by Eigenvector Centrality:

Airport	Eigenvector Centrality
ATL	0.3188
ORD	0.2769
LAX	0.2732
DEN	0.2405
DFW	0.2390
SFO	0.2150
LAS	0.1998
PHX	0.1908
MCO	0.1730
IAH	0.1612

Table 1: Top 10 Airports by Eigenvector Centrality

Interpretation: Nodes (airports) with high eigenvector centrality are influential because they are connected to many other airports that are themselves well connected. In the US airport context, these are typically major hubs such as ATL (Atlanta) or ORD (Chicago), which are part of the backbone of the air travel system.

2.2 Katz Centrality

Theory: Katz centrality extends the idea of eigenvector centrality by adding a constant to account for the influence of distant nodes. It measures the total number of walks between nodes, with walks of longer length exponentially damped to ensure convergence.

Implementation: After implementing the code, we obtain the Top 10 airports by Katz Centrality:

Interpretation: Katz centrality emphasizes nodes that are connected—even indirectly—to many others. Airports with high Katz values can reach a large number of other nodes either directly or through intermediate nodes, reinforcing their influential role in the network.

Airport	Katz Centrality
JHW	0.4807
BFD	0.4720
EVV	0.1404
DLH	0.1213
ATW	0.1193
MGM	0.1168
ILG	0.1086
MSL	0.1061
SWF	0.1014
MOB	0.0993

Table 2: Top 10 Airports by Katz Centrality

2.3 Betweenness Centrality

Theory: Betweenness centrality measures the frequency at which a node appears on the shortest paths between other nodes. Nodes that score high in betweenness serve as critical bridges or connectors in the network.

Implementation: After implementing the code, we obtain the Top 10 airports by Betweenness Centrality:

Airport	Betweenness Centrality
ANC	0.2467
TEB	0.1611
FAR	0.1371
BUR	0.1301
HPN	0.1177
DLH	0.0903
FBK	0.0832
IAD	0.0788
FAI	0.0780
NUL	0.0707

Table 3: Top 10 Airports by Betweenness Centrality

Interpretation: Airports with high betweenness centrality act as critical intermediaries in air travel, facilitating movement between different regions of the network. These nodes often include less frequently mentioned hubs that nonetheless play a pivotal role in ensuring network connectivity.

2.4 Closeness Centrality

Theory: Closeness centrality measures how quickly a node can interact with all other nodes by calculating the inverse of the sum of the shortest path distances from the node to all others. High closeness centrality indicates that a node is, on average, ”close” to all other nodes and can spread information (or in this case, passengers) efficiently.

Implementation: After implementing the code, we obtain the Top 10 airports by Closeness Centrality:

Airport	Closeness Centrality
LAX	18.0365
ATL	18.0365
ORD	18.0365
DEN	18.0365
SEA	18.0364
SFO	18.0364
DFW	18.0364
LAS	18.0364
JFK	18.0364
PHX	18.0364

Table 4: Top 10 Airports by Closeness Centrality

Interpretation: Airports with high closeness centrality can reach all other airports quickly. In the context of air travel, this might correspond to geographically central hubs or airports that are well integrated into the network, thus ensuring efficient routing and transfer times.

2.5 HITS (Hyperlink-Induced Topic Search)

Theory: The HITS algorithm produces two separate scores for nodes:

- **Authority Score:** Indicates a node that is linked by many good hubs, suggesting that it is a reliable source of connections.
- **Hub Score:** Indicates a node that links to many good authorities, suggesting it serves as an efficient connector.

Implementation: After implementing the code, we obtain the Top 10 airports by HITS Authority Score and Hub Scores:

Airport	HITS (Authority Score)
ATL	0.0367
ORD	0.0318
LAX	0.0314
DEN	0.0277
DFW	0.0275
SFO	0.0247
LAS	0.0230
PHX	0.0219
MCO	0.0199
IAH	0.0185

Table 5: Top 10 Airports by HITS (Authority Scores)

Airport	HITS (Hub Score)
ATL	0.0367
ORD	0.0318
LAX	0.0314
DEN	0.0277
DFW	0.0275
SFO	0.0247
LAS	0.0230
PHX	0.0219
MCO	0.0199
IAH	0.0185

Table 6: Top 10 Airports by HITS (Hub Scores)

These measures are especially useful in directed networks, but they also provide insight into the mutual reinforcement between nodes in undirected networks.

Interpretation:

- **Hub Nodes:** Airports that serve as excellent linking points to many influential airports.
- **Authority Nodes:** Airports that are significant targets for connections, often indicating major travel destinations.

By comparing HITS results with other centrality measures, we can see if the same nodes consistently appear as pivotal points in the network, reinforcing their central role.

Interpretation:

The top nodes identified by eigenvector and Katz centrality may correspond to major transit hubs (typically large international airports). Betweenness centrality highlights airports that serve as bridges between different regions. Closeness centrality favors nodes that can quickly interact with many others, which might identify geographically central hubs. HITS can show airports that are excellent "hubs" (connecting to many high-quality nodes) or "authorities" (receiving many connections from these hubs).

Analysis of Centrality Rankings

Upon analyzing the top-ranked airports from each centrality measure, we observe both overlaps and differences:

Consistent Airports Across Centralities

Airports that frequently appear in the top 10 of multiple centrality measures include:

- **ATL** (Atlanta)

- **ORD** (Chicago O'Hare)
- **LAX** (Los Angeles)
- **DEN** (Denver)
- **DFW** (Dallas/Fort Worth)

These airports rank highly in:

- Eigenvector Centrality
- Closeness Centrality
- HITS (Hub & Authority Scores)

Interpretation: These are large international hubs with high passenger traffic and strong connections to other major airports. Their high Eigenvector and HITS scores indicate connectivity to other central airports, while their high Closeness Centrality suggests efficient access to the rest of the network.

Katz Centrality Highlights Regional Airports

Top airports by Katz Centrality include:

- JHW, BFD, EVV, DLH, etc. (mostly small or regional airports)

Interpretation: Katz centrality favors nodes connected to many low-scoring neighbors. Since it considers paths of all lengths with attenuation, dense local neighborhoods can lead to higher scores even for less globally significant nodes. This highlights how Katz Centrality is sensitive to local structure, particularly in networks with loosely connected subgraphs.

Betweenness Centrality Highlights Bridging Nodes

Top airports by Betweenness Centrality include:

- **ANC** (Anchorage)
- **TEB, FAR, BUR, HPN** (regional connectors)

Interpretation: These airports act as key connectors rather than major hubs. For example:

- **Anchorage (ANC)** bridges Alaska to the continental US, making it a critical node in shortest-path routing.
- **TEB (Teterboro) and BUR (Burbank)** serve as secondary airports in congested metro areas (NYC, LA), playing key roles in routing.

These results show how Betweenness Centrality captures control over flow rather than direct connectivity.

Closeness Centrality Favors Large Hubs

Top airports include:

- LAX, ATL, ORD, DEN, SEA, SFO, DFW, LAS, JFK, PHX

Interpretation: These airports are positioned to reach other airports in fewer hops. Their geographic centrality and dense connectivity ensure a low average path length to the rest of the network.

HITS Scores Align with Eigenvector Centrality

Hubs and Authorities both rank:

- ATL, ORD, LAX, DEN, DFW, etc.

Interpretation: These airports act both as **originators (hubs)** and **popular destinations (authorities)**, showing mutual importance. The alignment of hub and authority rankings suggests an undirected or symmetrically weighted dataset.

Task 3: Modularity and Community Detection

Goals

- Use a community detection algorithm (e.g., Louvain or Girvan–Newman).
- Compute the modularity score.

Community Detection

Community detection aims to partition a network into groups (communities) where nodes are more densely connected to one another than to nodes in other groups. In transportation networks, these communities might correspond to regions or clusters of airports that share strong interconnections.

Modularity

Modularity is a measure ranging from –1 to 1 that quantifies the strength of division of a network into communities. A high modularity (typically above 0.3) indicates a strong community structure where more edges fall within communities than would be expected by chance.

Implementation

We use the Louvain algorithm, a popular method that greedily optimizes modularity. It is well-suited for large real-world networks like air traffic. We use the greedy modularity maximization method (available in NetworkX) to detect communities, compute the modularity score, and optionally visualize the network with different colors for each community.

Results

Number of Communities and Partition

Number of communities detected: 20

The Louvain algorithm assigned each airport to one of 20 distinct communities.

This relatively large number of communities indicates fine-grained partitioning across the network. Each community likely represents clusters of airports that are highly interconnected, possibly reflecting regional operations, airline hubs, or passenger flow patterns.

Modularity Score

Modularity score: 0.2541

Interpretation

A modularity of 0.2541 is moderate, suggesting that the network has a somewhat noticeable but not strongly pronounced community structure.

While the network does exhibit some clustering behavior, it is not deeply modular. This could be due to the highly interconnected nature of air transportation, where hub airports serve diverse regions.

The presence of many inter-community connections (especially through large hubs like ATL, ORD, LAX) reduces the overall modularity.

Visualization

In the community-colored visualization:

- Dense clusters of smaller regional airports can be seen forming communities.
- Major hub airports are often located on the boundaries between communities, acting as bridges.
- Some visual separation is present, but many cross-community links are visible, which supports the low-to-moderate modularity score.

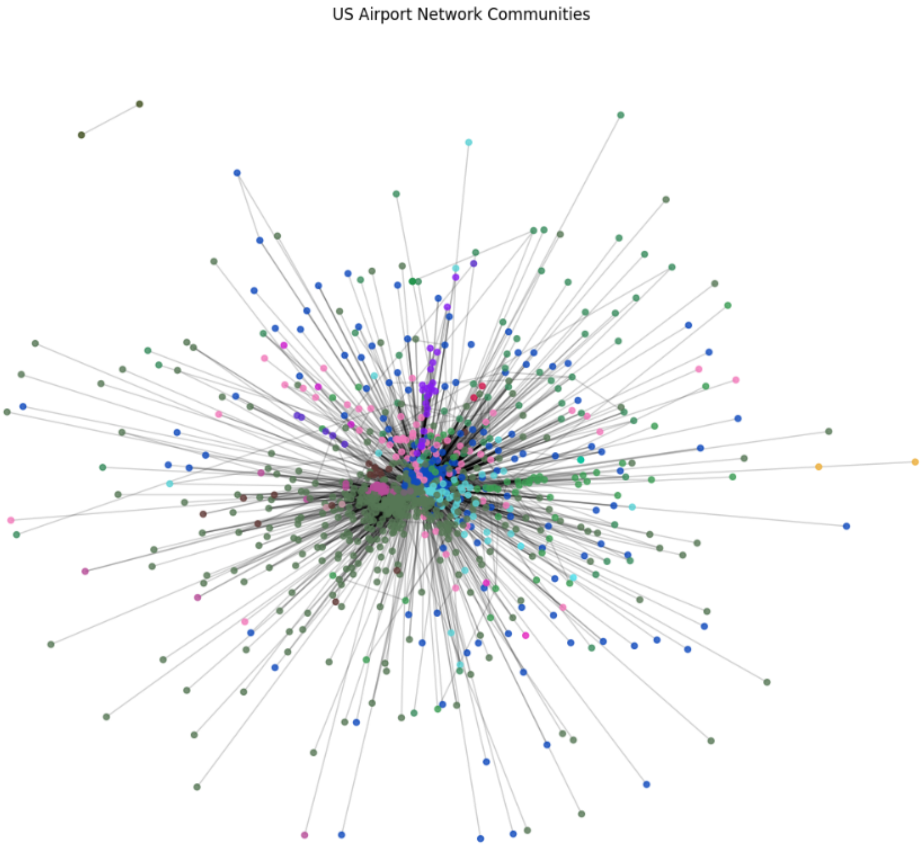

Figure 3: US Airport Network Communities

Are the Communities Meaningful?

Yes, to a certain extent.

The communities reflect real-world airline operations and geographic grouping:

- Regional communities such as clusters of airports in the Midwest, Southeast, or Pacific Northwest can be identified.
- Airports from remote regions like Alaska often form their own community due to their localized connectivity.

However, the presence of large multi-hub airlines operating nationwide routes reduces modularity:

- Airports like ATL (Delta), ORD (United), or DFW (American) connect to many communities, which decreases community separation.

This reflects the dual nature of the air transport network: it is locally dense but globally interconnected.

Task 4: Assortativity & Degree-Degree Correlations

Objectives

- **Degree Assortativity:** Do high-degree nodes connect preferentially to other high-degree nodes?
- **Degree-Degree Correlation Plot:** Plot average neighbor degree vs. node degree.

Degree Assortativity

Degree assortativity quantifies whether nodes in a network tend to connect to other nodes with similar degrees (number of connections). It is calculated as the Pearson correlation coefficient between the degrees of connected node pairs:

- **Positive value** → Assortative network
(e.g., social networks, where well-connected individuals tend to associate with other well-connected individuals).
- **Negative value** → Disassortative network
(e.g., technological or biological networks, where hubs connect to many low-degree nodes).
- **Zero or near-zero** → Neutral (no clear pattern).

The degree–degree correlation plot complements this by showing the average neighbor degree (k_{nn}) as a function of a node's degree (k). This helps visualize whether high-degree nodes connect to other high-degree nodes or low-degree nodes.

Results

Assortativity Coefficient

- **Pearson degree assortativity coefficient:** -0.1133

Interpretation:

- The negative coefficient indicates that the network is disassortative.
- In the context of the US airport network, this means major hub airports (high-degree nodes) tend to connect with smaller regional airports (low-degree nodes).
- This is expected behavior for transportation systems that aim to connect distant or less-connected locations to central hubs.

Degree–Degree Correlation Plot

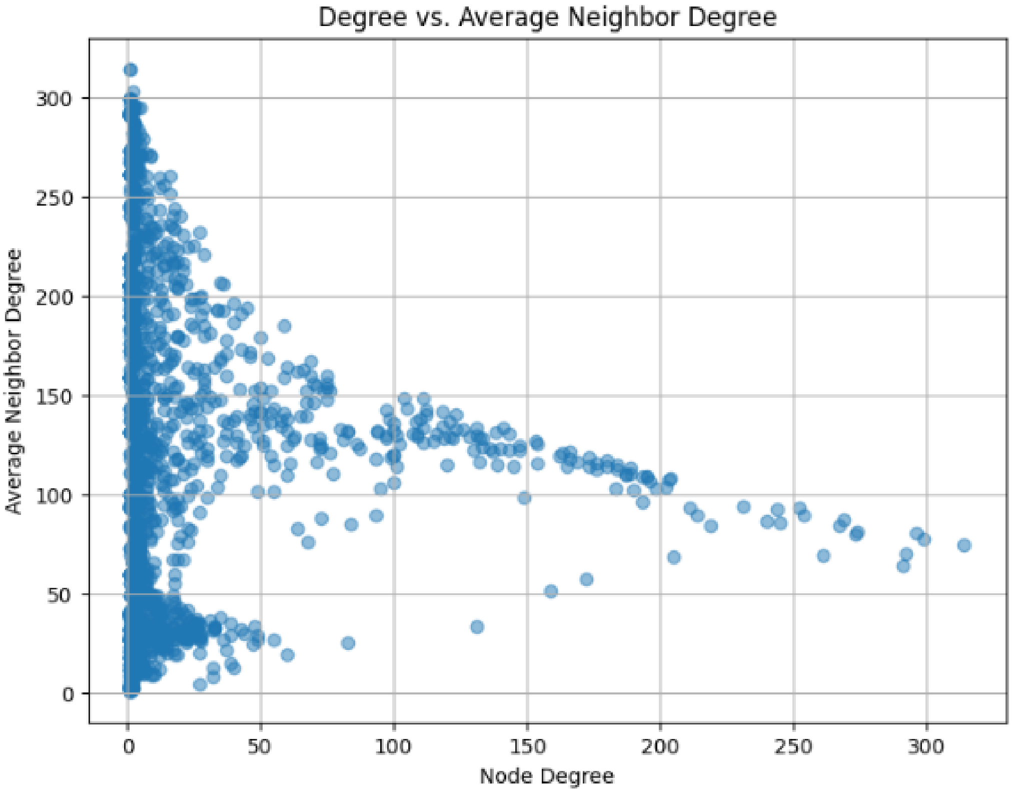

Figure 4: Degree–Degree Correlation Plot

Interpretation:

- The plot clearly shows a downward trend: as node degree increases, the average neighbor degree decreases.
- High-degree airports (those with many direct connections) often link to airports that are less connected — consistent with a hub servicing multiple spokes.
- There's noticeable scatter at low degrees, indicating that smaller airports have a more varied range of connections, but the trend flattens and declines at higher degrees.

This pattern is logical and expected in aviation networks, where central hubs like ATL, ORD, or LAX serve as connection centers linking numerous smaller, regional airports to the broader network.

Task 5: Clustering Coefficients

Analysis

- **Compute the global clustering coefficient (transitivity).**
- **Compute local clustering coefficients for each node (or for a subset).**

Clustering Coefficients

The clustering coefficient measures how likely it is that two neighbors of a node are also connected to each other — essentially capturing the ”friend of a friend” idea.

Global Clustering Coefficient (Transitivity)

The **Global Clustering Coefficient** is the proportion of closed triplets (triangles) to all triplets (both open and closed) in the network.

Global Clustering Coefficient: 0.3841

Random Graph Clustering Coefficient (same size & density): 0.0138

Interpretation:

- The network's clustering is significantly higher than that of a random graph.
- This suggests that the real-world airport network exhibits strong triadic closure, meaning airports often form tightly-knit groups (e.g., multiple regional airports linked via a central hub).

Local Clustering Coefficients

For a specific node, the **Local Clustering Coefficient** measures how close its neighbors are to being a complete clique.

Top 10 nodes by degree and their clustering values:

Node ID	Degree	Local Clustering Coefficient
114	314	0.1839
709	299	0.1965
1200	296	0.2163
877	292	0.1636
766	291	0.1361
389	274	0.2150
500	273	0.2129
1068	269	0.2447
711	267	0.2351
1016	261	0.1593

Table 7: Top 10 nodes by degree and their clustering coefficients

Interpretation:

- These high-degree nodes (likely hub airports) have moderate clustering values (0.15 to 0.25), showing that while they're highly connected, their neighbors are not fully interconnected — consistent with the hub-and-spoke structure.
- Node 1 has a clustering coefficient of 1.0, indicating that its neighbors form a complete clique — likely a small regional or local network.

Distribution of Local Clustering Coefficients

The distribution of local clustering coefficients provides insights into the overall network structure. Airports with high local clustering values tend to form dense regional subnetworks, while hub airports typically have lower clustering due to their connections spanning multiple communities.

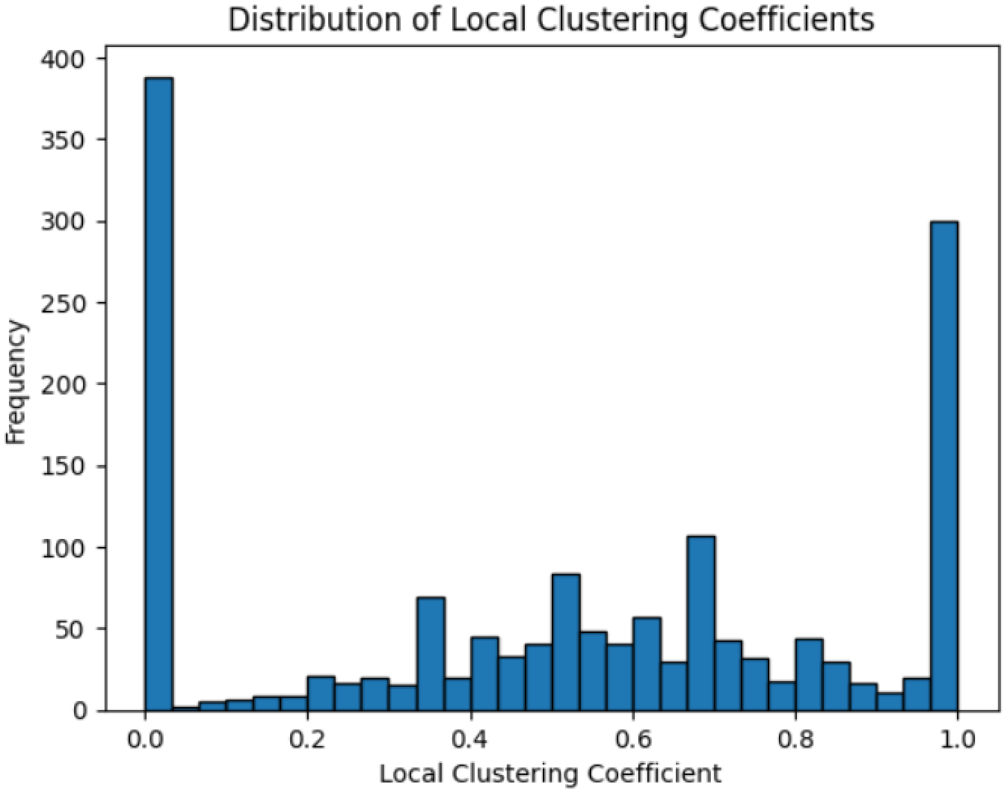

Figure 5: Degree–Degree Correlation Plot