SNA U1 - project - Networks Theory

Luis David Martínez Gutiérrez
Data Engineering
Universidad Politécnica de Yucatán
Km. 4.5. Carretera Mérida — Tetiz
Tablaje Catastral 4448. CP 97357
Ucú, Yucatán. México
Email: 2109099@upy.edu.mx

Didider Omar Gamboa Ángulo
Universidad Politécnica de Yucatán
Km. 4.5. Carretera Mérida — Tetiz
Tablaje Catastral 4448. CP 97357
Ucú, Yucatán. México
Email: didier.gamboa@upy.edu.mx

Abstract

This report presents an extensive analysis of the US Airlines network, focusing on its structural properties and centrality measures. The network, comprising 332 nodes and 4252 links, is classified as a directed network. Key metrics such as average path length, clustering coefficient, and diameter were calculated. Various centrality measures were analyzed, identifying major hubs and influential nodes. The degree distribution and community detection provided insights into the network's structure and the presence of distinct communities. The findings highlight the importance of major hubs in maintaining connectivity and the potential for optimizing flight routes and services based on network analysis.

I. Introduction

The chosen network is the US Airlines network, which represents the connections between various airports in the United States. This network is crucial as it helps in understanding the connectivity and efficiency of the air transportation system in the US. The analysis of this network can be applied to optimizing flight routes, improving airport services, and enhancing the overall passenger experience.

This dataset has been used in various research papers to analyze the structure and dynamics of transportation networks. Generally, the results from such analyses highlight the highly connected nature of major airports and the presence of community structures within the network.

The US Airlines network can be classified as a directed network because the connections (flights) between airports have a direction, indicating the departure and arrival points. It is not a bipartite network since it doesn't involve two distinct sets of nodes. It's also not planar because it involves numerous intersecting connections that can't be represented on a plane without overlaps.

SNA U1 - project - Networks Theory

II. Network Characteristics.

A. Size of the Network

The size of the network refers to the total number of nodes, which in this case are the airports. This metric gives us an idea of the scale of the network.

Size of the Network: The US Airlines network consists of 332 nodes.

B. Number of Links

The number of links (edges) represents the total number of direct flights between airports. This helps us understand the connectivity within the network.

Number of Links: The network has 4252 links.

C. Average Path Length

The average path length is the average number of steps along the shortest paths for all possible pairs of network nodes. It provides insight into the efficiency of the network in terms of how easily passengers can travel between airports.

Average Path Length: The average path length of the network is approximately 2.74.

D. Clustering Coefficient

The clustering coefficient measures the degree to which nodes in a network tend to cluster together. It indicates the likelihood that two airports connected to the same airport are also connected to each other. A high clustering coefficient implies a network with tightly knit groups.

Clustering Coefficient: The average clustering coefficient of the network is approximately 0.63.

E. Additional Distance Metrics

Average Distance: Similar to the average path length, it represents the average number of steps along the shortest paths in the network. **Diameter**: The longest shortest path between any two nodes in the network. It provides insight into the network's size in terms of reachability. **Eccentricity**: The greatest distance between a node and any other node. It helps identify the nodes that are the farthest apart in the network. **Radius**: The minimum eccentricity of any node. It represents the shortest maximum distance from any node to all other nodes. **Periphery**: The set of nodes with the highest eccentricity. These nodes are the farthest from all other nodes in the network. **Center**: The set of nodes with the lowest eccentricity. These nodes are the most central in the network.

Average Distance: The average distance is the same as the average path length, which is 2.74. Diameter: The diameter of the network is 6. Eccentricity: The maximum eccentricity in the network is 6. Radius: The radius of the network is 3.

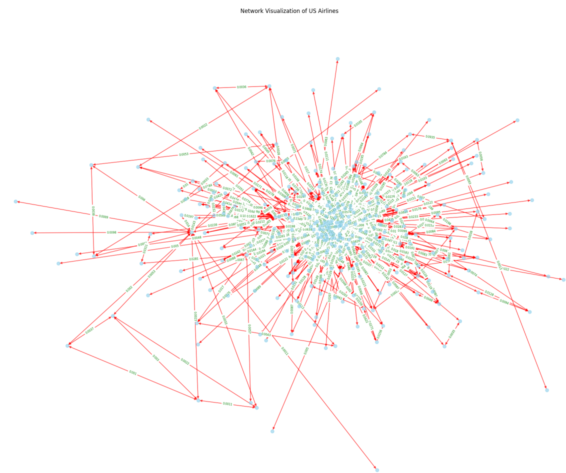

Fig. 1. Spring Layout with Default Parameters

Periphery: The peripheral nodes are those with the maximum eccentricity, including nodes like 9, 10, 11, etc. Center: The central nodes are those with the minimum eccentricity, including node 313.

F. CODE 2: Circular Layout

Graph Layout: The nodes are positioned using a circular layout. This positions all nodes on a circle, spreading them out evenly around the circumference.
Colors and Labels: Nodes are blue, edges are light gray, and labels are red.
Visualization: The circular layout helps to evenly distribute nodes around a central point, making it easier to see individual nodes and their connections, but it can make it harder to see the overall structure and clusters within the network.

III. How the Different Graphs Help in Visualization

A. CODE 1 (Default Spring Layout)

Strengths: Good for visualizing the overall structure and clustering within the network.
Use Case: Ideal when you want to see the general layout and identify clusters and hubs.

B. CODE 2 (Circular Layout)

Strengths: Excellent for identifying individual nodes and their direct connections due to even spacing.
Use Case: Useful when the focus is on the connections of individual nodes rather than the overall network structure.

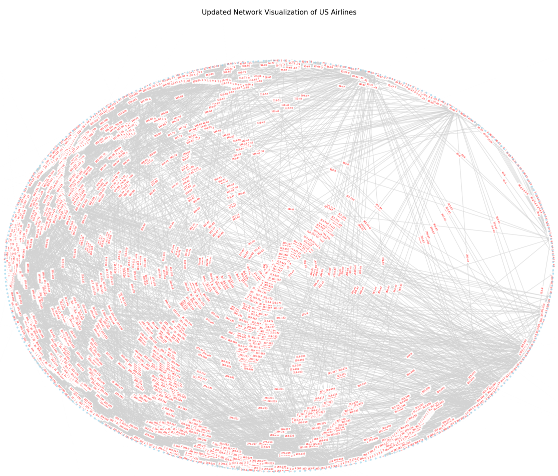

Fig. 2. Circular Layout

```
Measure: Size of the Network: 332
Number of Links: 4252
Average Path Length: 2.7381247042550867
Clustering Coefficient: 0.6252172491625031
Average Distance: 2.7381247042550867
Diameter: 6
Eccentricity: {'1': 5, '2': 5, '3': 5, '4': 5, '5': 5, '6': 5, '7': 5, '8': 4, '9
Radius: 3
Periphery: ['9', '10', '11', '12', '14', '15', '17', '18', '19', '20', '21', '22'
Center: ['313']
```

Fig. 3. Circular Layout

C. How the Different Graphs Help in Visualization

D. CODE 1 (Default Spring Layout)

Strengths: Good for visualizing the overall structure and clustering within the network.
Use Case: Ideal when you want to see the general layout and identify clusters and hubs.

E. CODE 2 (Circular Layout)

Strengths: Excellent for identifying individual nodes and their direct connections due to even spacing.
Use Case: Useful when the focus is on the connections of individual nodes rather than the overall network structure.

Each layout offers unique advantages depending on the specific aspects of the network you wish to analyze. The spring layout (default and adjusted) is better for overall structural insights, while the circular layout excels in clarity of individual node connections.

IV. Centrality Measures.

Centrality measures help identify the most important nodes in a network based on different criteria. For the US Airlines network, the following centrality measures are considered:

A. *Degree Centrality*

Degree centrality measures the number of direct connections a node has. Airports with high degree centrality are major hubs with many direct flights.

B. *Betweenness Centrality*

Betweenness centrality measures the extent to which a node lies on the shortest paths between other nodes. Airports with high betweenness centrality play a critical role in connecting different parts of the network.

C. *Closeness Centrality*

Closeness centrality reflects how close a node is to all other nodes. Airports with high closeness centrality can reach other airports more quickly and efficiently.

D. *Eigenvector Centrality*

Eigenvector centrality measures the influence of a node in the network. Airports with high eigenvector centrality are connected to other highly influential airports.

E. *Observations*

Degree Centrality: Indicates which airports are the main hubs. These hubs are critical for network efficiency and passenger traffic.
Betweenness Centrality: Highlights airports that are crucial for maintaining connectivity between different parts of the network. These airports are key for route optimization and emergency response.
Closeness Centrality: Identifies airports that can reach others quickly, which is important for efficient travel times.
Eigenvector Centrality: Reflects the overall influence of airports in the network. Airports with high eigenvector centrality are integral for maintaining network stability.

	Degree Centrality	Betweenness Centrality	Closeness Centrality	Eigenvector Centrality	Name
118	0.419940	0.048395	0.144455	3.368532e-58	Chicago O'hare Intl" 0.7914 0.3710 0.5000\,
261	0.356495	0.043513	0.373403	7.949398e-19	Dallas/Fort Worth Intl" 0.7449 0.4578 0.5000\,
255	0.305136	0.022892	0.337740	1.402415e-21	The William B Hartsfield Atlan" 0.8091 0.4506 ...
152	0.283988	0.022363	0.170822	4.940028e-49	Pittsburgh Intll" 0.8305 0.3852 0.5000\,
182	0.283988	0.026587	0.229017	1.035827e-34	Lambert-St Louis Intl" 0.7789 0.4018 0.5000\,
230	0.262840	0.018125	0.275244	1.301940e-26	Charlotte/Douglas Intl" 0.8268 0.4356 0.5000\,
166	0.256798	0.018262	0.204353	3.999268e-43	Stapleton Intl" 0.7049 0.3920 0.5000\,
67	0.235650	0.021579	0.068324	8.220781e-70	Minneapolis-St Paul Intl/Wold-" 0.7644 0.3432 ...
112	0.211480	0.006950	0.099924	1.000477e-59	Detroit Metropolitan Wayne Cou" 0.8146 0.3687 ...
201	0.205438	0.023056	0.255905	2.073334e-33	San Francisco Intl" 0.6159 0.4126 0.5000\,

Fig. 4. Spring Layout with Default Parameters

V. Degree Distribution.

The degree distribution of a network shows how the connections (or edges) are distributed among the nodes (or vertices). It provides insights into the overall structure and robustness of the network.

Observations
The degree distribution can reveal whether the network follows a scale-free distribution, where a few nodes (airports) have a very high degree, while most nodes have a low degree. A heavy-tailed distribution indicates the presence of hubs, which are critical for the network's connectivity and resilience.

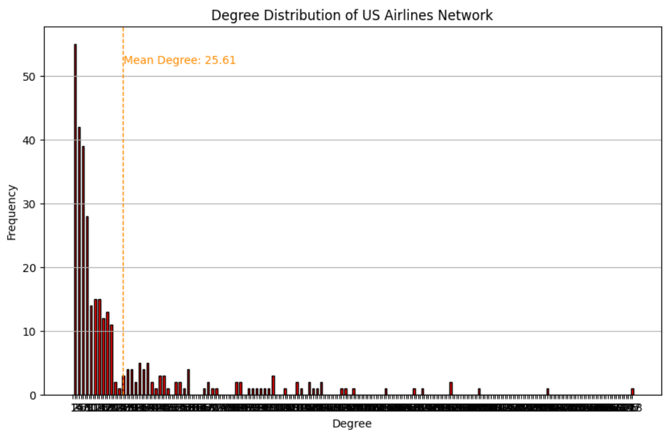

Fig. 5. Spring Layout with Default Parameters

VI. Community Detection.

Community detection is a process used in network analysis to identify groups of nodes that are more densely connected to each other than to the rest of the network. These groups, known as communities, can reveal the underlying structure of the network and can help in understanding how the network functions.

In this analysis, we apply the Girvan-Newman method for community detection. The Girvan-Newman method is a hierarchical clustering algorithm that detects communities by progressively removing edges with the highest betweenness centrality. Betweenness centrality measures the number of shortest paths that pass through an edge. By removing these edges, the network is split into smaller, more cohesive sub-networks, which represent the communities.

A. Explanation of the Graph

1) Graph Layout: The nodes are positioned using the spring layout algorithm, which simulates a physical system where nodes repel each other while edges act like springs pulling nodes together. This results in a visually appealing representation where clusters and hubs are more visible.

2) Colors and Labels: Each community is assigned a unique color using the 'cool' colormap. Nodes within the same community are colored similarly, making it easy to identify the different communities. The node labels are included to identify the airports.

3) Visualization: This visualization shows the network's community structure, highlighting how the network is divided into sub-networks or communities. Each community likely represents a group of airports that have more frequent or direct connections with each other than with the rest of the network.

B. Observation

An important observation from the community detection analysis is the presence of tightly-knit groups of airports that form distinct communities. These communities can be based on geographical proximity, airline alliances, or major flight routes. Identifying these communities is crucial for optimizing flight operations, managing air traffic, and improving passenger connectivity. For instance, a disruption in a major hub within a community can significantly impact the connected airports within the same community, but might have a lesser impact on airports outside the community.

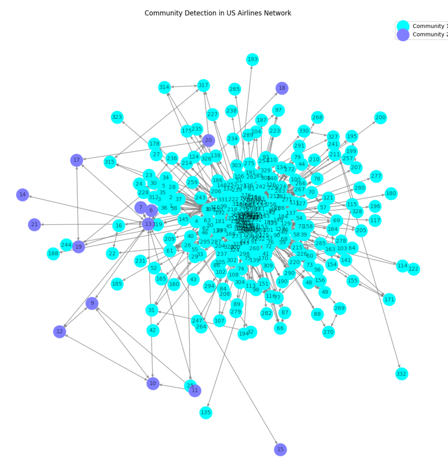

Fig. 6. Spring Layout with Default Parameters

VII. Conclusion

Based on the analysis, the US Airlines network exhibits a highly centralized structure with a few key hubs playing crucial roles in connectivity. The degree distribution suggests a hub-and-spoke model, which is common in airline networks. Centrality measures highlight the importance of specific airports in maintaining network efficiency and resilience. Community detection reveals regional clusters that can inform strategic planning and operational improvements.

The analysis of the US Airlines network reveals that it is a highly connected and centralized network with major hubs playing a crucial role in maintaining connectivity. The degree distribution shows that most airports have a low degree, while a few have a very high degree, indicating a scale-free network. Community detection highlights the presence of distinct communities within the network, which could be related to geographical regions or airline alliances.

References

[1] Hagberg, A., Schult, D., Swart, P. (2008). NetworkX: Network Analysis with Python. Retrieved from https://networkx.github.io/documentation/stable/

[2] Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science Engineering, 9(3), 90-95. Retrieved from https://matplotlib.org/stable/contents.html

[3] Girvan, M., Newman, M. E. J. (2002). Community structure in social and biological networks. Proceedings of the National Academy of Sciences, 99(12), 7821-7826. Retrieved from https://www.pnas.org/content/99/12/7821

[4] Stanford large network dataset collection. (n.d.). Retrieved from https://snap.stanford.edu/data/

[5] Pajek datasets. (n.d.). Retrieved from http://vlado.fmf.uni-lj.si/pub/networks/data/