June,24 2023

NAVIGATING THE SKIES: EXPLORING CENTRALITY AND REGIONAL DYNAMICS IN THE U.S. AIRLINE NETWORKS.

Astrid Cristina Tello Willis
MASTER THESIS
Student nu. 2626530

MSc. in Spatial, Transport and Environmental Economics
Vrije Universiteit

Thesis Supervisor: Dr. Eric Pels

To my gran, who dreamt of being a pilot.
To my dad who found happiness in flying.
To my mom who never cut my wings.

ACKNOWLEDGMENTS

Writing this thesis was a journey, both academically and personally. I’m filled with gratitude for the experience, the learnings, struggles and everyone I crossed flights with.

Certainly, this was made possible thanks to the guidance and invaluable inputs of my supervisor. Eric, thank you for your constant support, for your availability, disposition, feedback and flexibility throughout this process. This would not have been possible without you.

I'm grateful for the moral support of my friends who showed unconditional love. Special thanks to Beth who was there from the start providing inputs to tackle the problem better and pro-tips from a PhD candidate, I owe you greatly, you made the days easier and better. To my Latam friends whose recurrent questions about the progress and words of wisdom kept me on track and motivated even when times got rough. Ceci, Margarita, Marilú, Lore, Ana, Cati, Estefanía and the rest of my *Lunecitas*, this is also to you.

Finally, to my family; my mom Patricia, my godmother Miriam, and my brothers Joao & Fernando, who cheered me incessantly, regardless of the wheres and whens, and who have been my greatest admirers without reason. This is mostly for you, you’re the absolute reason why this was possible. *No cabe en una vida mí gratitud*.

Más vale tarde que nunca. ¡Se logró!

ABSTRACT

The document discusses the network dynamics of the U.S. airline transportation system during a 25 year period, taking cities as nodes and direct non-stop flight as edges. The study explores the global and node network topography, and its evolution to understand if there are regional dynamics that affect its behavior.

In line with previous studies, the network exhibits a small-world property, with a 2.2 short average path length and a 66% clustering coefficient. Chicago is identified as the most central city in the network based on three centrality measures: degree centrality, eigenvalue centrality, and betweenness centrality.

However, the network presents centrality anomalies, where the most connected cities are not always the most central, using both an unweighted network and one weighted by distance, highlighting the presence of community structures. Using the Louvain algorithm to identify 6 different communities and Guimera et al (2005) methodology to classify nodes within an intra and inter community scores, each city is classified to verify its role in the network to find that there exist regional behavioral differences in the proportions of the types of nodes, and the type of travels that happen in each region.

Geography is found to play a role in the network, but it does not necessarily determine centrality, nor the influence of one city to the rest of the country. The most connected cities are not always the most geographically central ones. Factors such as existing infrastructure, demand patterns, and connectivity to other airlines contribute to a city's centrality in the network.

Overall, the study provides insights into the structure, dynamics, and evolution of the U.S. airline network, highlighting the importance of individual cities, hub airports, and the interplay between geography and network connectivity.

TABLE OF CONTENT

I. INTRODUCTION

Áureo de Paula (2017) starts his chapter of the book *Advances in Economics and Econometrics* saying that "Networks are vulgar… they are commonplace, ordinary". The fact that they go unnoticed, however, does not limit its importance. Their structures "define how information, prices and quantities reverberate in a particular economic system". This is true for almost every network that exists, albeit its invisibility.

In particular, networks are the underlying structures of almost everything around us. From the transportation roads that we use, to urban project planning, its structures define patterns that are, sometimes invisible to the eye, but contain information and insights. Understanding these structures could be beneficial in many fronts: from the effective management in cases of crisis by identifying critical locations in networks (Demsar et al), to understanding saturation of information or even to model how a virus might spread.

Generally, however, these structures are endogenous to the subject of study and generally evolve to more complex ones as time passes since the interactions on the objects and its link feed each other and affect one another. Therefore the importance of their study is to try to find the order and balance in a complex structure, its connections and interdependencies. This is the case for most airline networks, and the U.S. in particular, where aircraft excess supply after the Second World War pushed for a reutilization of war aircrafts for commercial uses with very low prices that increased demand in the region, then followed periods of regulation and deregulations which shaped its supply side (Cook, 1996). And with the growth of tourism and globalization, the opening of frontiers and new destinations, airlines reorganize their routes trying to maximize their profits, depending on their cost structure, competition, airport capacity, among other factors, making each network unique, complex and constantly evolving.

Thus, the objective of this research is to study the evolution of the U.S. airline Network from 1997-2022 to find what are the underlying factors that have shaped its structure. The hope is to analyze the network and its most central cities both from a geographical and topological perspective, measured by their betweenness centrality, to evaluate if their position has influenced the system, or what other factors play an important role in shaping the network structure. Understanding the network and its spatial distribution has a direct impact in various areas, from cost efficiency on the supply side, and time efficiency from the demand side, to capture localities that are underserved. Even disturbances can be mitigated by analyzing which nodes are the most relevant for the network or could impose a higher risk of attacks or weather crises. The use of complex networks theory allows for a characterization of the network and while similar research has been performed o a worldwide scope, understanding the within border characteristics allows for better policy decisions and finding possible opportunities for airlines.

The document is organized as follows: Section II. reviews basic concepts in graph theory and highlights some of the existing literature in airline network topology and its case studies.

Section III. explains the methodology and its relevance for the research. Section IV. clarifies the data selection process, its manipulation and some basic statistics, followed by Section V. which describes the results and analysis. Finally, Section VI. concludes with a summary, some policy suggestions and a brief discussion and limitations of the study.

II. LITERATURE REVIEW

1. Graph concepts.

The first studies of networks began in 1736 with the Konigsberg bridges problem written by Euler. Although it did not represent any graph theory *per se*, historians and economists consider this the cornerstone of the field interested in the relationship between nodes and edges and the mathematical theory that stemmed out of it.

Aldous and Robin (2000) define a graph as a visual representation that contains points, called *vertices or nodes* and lines that join each point together called *edges or links*. The term networks or graphs is often used interchangeably and their studies carry numerical information on the agents that analyze the relationship between the objects pertaining to the graph.

A graph has *multiple edges* if two or more edges join the same pair of vertices, and a vertice can have a *loop* if an edge links to itself. The *degree* of a node (v) measures the number of edges incident with v, and if it is a loop it gets counted twice. A *walk* is a sequence of vertices and edges of a graph, while a *trail* is a walk in which all the edges, but not necessarily all vertices, are different, whereas a *path* is a walk in which all the edges and vertices are different.

A graph then can be defined by its direction; a *directed graph* is where the edges connotate a walk from point a to b, or an *undirected graph* which only represents the connection between the nodes. And in terms of its density a *sparse graph* describes a one that contains many vertices but few edges, or a *simple graph,* one with no loops or multiple edges.

A *clique* is a subset of a network in which the nodes are adjacent to each other, It represents a complete subgraph that requires every pair of distinct nodes to be connected with a unique edge. Their existence highlights the presence of communities. A *motif* introduced by Milo et al (2002) is defined as a small connected subgraph with a well defined structure, which occurs significantly more frequently than it would appear in a random graph[1]. They are sometimes considered the building blocks of the network.

A network is said to be *scale free* if the fraction of the nodes with degree k follows a power law distribution $k^{-\alpha}$ with $\alpha > 1$. As Broido and Clauset (2019) explain "a power law is the only

[1] A random graph refers to the way to construct a graph where an edge is included with some probability (p), and excluded with the (1-p) probability.

normalizable density function f(k) for node degrees in a network that is invariant under scaling. For instance f(c, k) = g(c) f(k) for any constant c, and thus free of natural scale."

For computing purposes, matrices are an alternative to graph representation which allows the analysis of the information better. If G is a graph of n vertices from 1,2.3,...n, the *adjacency matrix* is one in which the entry in row i and column j is the number of edges joining the vertices i and j. Every node is represented as an entry in both row i and column j and a 1 is given depending on the number of edges that the node 1 has with node 2. The main diagonal is represented by 0. The *Incidence matrix* contains 1 in the presence of an edge connecting the nodes, or 0 if there is none.

To characterize the graph it is defined by its size and structure: The *diameter* measures how far the most distanced nodes in the network are measured as the shortest path between them. The *shortest path* is the path that minimizes the weights (or distance) of the edges between two nodes.

A robust density measurement is the *clustering coefficien*t which computes how the neighbors of the nodes are connected to one another. It is calculated by the sum of edges connecting a vertex's neighbors over the total number of possible edges between the vertex's neighbors (Hansen et al. 2020).

To measure the node's importance in a network, several centrality measures exist: degree centrality measures the number of nodes connected, betweenness centrality measures the importance of the node in the flow of information and eigenvector centrality measures how influential the node is.

2. Graph theory in Networks

Nowadays one of the most studied networks is the transportation system, responsible for moving people and things on a daily basis across the globe. The way people commute and travel determine its structure, and the networks themselves, impact someone's life, directly and indirectly, from spillover effects, to negative externalities. In the U.S. alone, almost 280 million vehicles were registered in 2021, and more than 850 million passengers were carried by U.S. airlines in 2022 (United States Department of Transportation). This massive transportation volume makes the study of its qualities extremely important since a "failure in its operations could seriously impact regional economic development" (Zhang et al., 2021) and understanding the drivers to its evolution could allow for better optimization and functionality.

In aviation, as the network grows and develop, its complexity calls for a better understanding of its structure especially as saturation comes at play, if at any point, airlines and airports reach maximum capacity, and infrastructure on a network level and a carriers' can not cope with an increasing size, the importance of studying the airline transportation network becomes paramount. As Lin and Ban (2014) highlight, improving the efficiency of travel and optimizing its underlying topographical network structures will be the ultimate goal of aviation. This has

been done internationally taking the world transportation network with a focus on both passengers and freight, or country-specific around the world but with a higher volume of studies from the U.S. and China given data availability. Others try to evaluate the system depending on the type of network structure they choose to work with, whether it is a Hub and Spokes or Point to Point structures.
This following section will review some of the literature in transportation networks and summarize some of its findings.

3. Topological Characteristics and Network Evolution

Network topology reviews how nodes and edges are organized in the system to understand its structure and how information flows in the system. Researchers model data to characterize the networks, understand their properties, and replicate their behavior in more complex situations.

3.1 Small-World

One of the most recurrent findings in airline theory is represented by the Small-World graph first characterized by Watts and Strogatz (1998) which displays higher density and smaller shortest path than what a random graph would, allowing the nodes in the network to be "agents of speed and synchronizability", where information travels faster, or spreads easier.

There is vast evidence that the airline networks behave this way. Guimera et al (2005) find the global transportation system is a scale-free small world one. However, contrary to their beliefs, the central cities are not the most connected ones. They concluded these anomalies are a response to the community structures governed not only by geographical location but geopolitics, commerce and international ties. This is important, because global connectivity allows the exchange of ideas, people and goods, an uneven distribution would seclude some regions from development and trade. Although trade patterns are rarely homogeneously distributed across the same countries (Markusen et al., 1991), finding those areas with less accessibility allows for better development prospects or policy agendas.

For the U.S. Li-Ping et al. (2003) characterized the network as a directed, weighted one to find an average path-length of 2.4 and a clustering coefficient of 0.61 showing the presence of a small-world property. Xu and Harris (2008) studied intercity passenger travel from 1996-2008 weighted by passenger traffic, non-stop distance and average one-way fare to find a small-world network accompanied by "dissortative mixing patterns and rich-club phenomenon"[2]. Tao et al (2014) explored the evolution of the US airport network from 1990 to 2010 to find that, in spite of its massive growth during the years, it is coherent with previous studies. Other authors who conclude the same include Lin and Ban (2014) and Cheung and Gunes (2012).

[2] ***Disassortative Mixing*** vertices with high degree tend to connect to low degree vertices.

For Asia, Wang et al (2011) explored the Chinese transportation network computing all centrality measures for individual cities, and confirmed that the network displays small-world properties, with low average path length and high clustering coefficient while finding that centrality indices are correlated with population and regional GDP capturing spillover effects. Li and Cai (2004) studied both directed and undirected paths to find a small world order property and a double pareto law. Dai et al (2018) did so to the Southeast Asian Network for the period of 1979 -2012. The region also showed a scale-free network with an increasing number of hub cities and shows, since 1996 small-world property with a disassortative mixing pattern, suggesting an increasing number of hubs as the region gains economic strength. They even performed a decomposition analysis to tier the cities between capital, secondary and tourist destinations to find spatial characteristics, showing interconnection characteristics depending on the locations of the cities, where the peripheral, remote cities have low significance in the network and low connectivity, while the core is densely connected and geared in the north (and capital cities). The inbetween layer, shows high volatility over time, but just as Guimera et al, track these changes to socio-economical and political dynamics, more than on geographical distribution.

Other international descriptions of small-world phenomena include Italy, Michele and Funaro (2007), confirming the presence of a scale-free network, given that degree and betweenness centrality exhibit a power-law, and find the presence of hubs and the lack of communities. India, Bagler (2008) with a weighted network to find that it shows the behavior of a complex, small-world behavior characterized by a truncated power-law degree and shows some degree of hierarchy, and with main transport flow concentrated between large airports. Australia Hossain and Alaim (2017); Brazil Rocha (2009) who indicate similar findings.

3.2 Sub-groups and its impact on network topology evolution

To determine the structures of a graph, understanding the subsets of a network and how they relate to one another can help identify the core of the system and its evolution. In particular cliques determine a set of nodes where each node is connected to each other with a unique set of edgest, and motifs, characterized first by Milo et al (2002) describe a subgraph that repeats itself or is shown in a pattern more often than a random graph. It is considered the building blocks of complex networks given its presence and influence in the rest of the nodes.

Yang et al. (2021) studied the Chinese airline industry, implementing a motif detection technique to research how components build up to connect to larger networks and finding patterns by removing cliques and crucial vertices in the data. They found that the network has a multicentric and hierarchical structure with the majority of its operation only important on a regional level. As for the reach of its network, the U.S. and Canada are the most popular regions for cliques. They concluded that the combination of airports in the cliques may be affected by airline capacity, and industry dynamics such as traffic rights and airline cooperation.

4. Centrality Measures and nodes dynamics

An important area of graph theory study is the analysis of individual nodes in the system and their influence on the entire structure. For this, centrality measures have played a role in the studies of airline networks as they highlight the most important centers or nodes. Specifically, they have been studied to analyze the robustness of the network and its resistance and resilience to outer shocks to provide policy guidance, or to measure the network efficiency as certain nodes convey more information than others and allow for better propagation.

For robustness, Zhang et al. (2021) worried about the impact of shocks in the network and studied the robustness of the Belt and Road region that comprises 66 countries across the Eurasia region. They define robustness as the ability to maintain its designed functionality without having network disturbances. They perform a ranking according to centrality measures to then compute a sequential removal of nodes to verify the impact of their absence. They conclude that networks are more vulnerable to sequential targeted attacks than random attacks; for unweighted networks, betweenness and degree centrality had stronger impact in maintaining order to a shock than eigenvector or closeness centrality, for a weighted one, recursive power had a more important role than recursive centrality. However, they were not able to find a single particular centrality measure that identified the robustness of the network on its own.

Verma et al (2014) think of resilience of the Worldwide Network as the most important feature but one that is often taken for granted. They find that for long distance travel the network is redundant (repeated routes), and therefore resilient, however, it is weak for short distance and would fall apart if seemingly insignificant nodes were removed. They highlight the existence of a core composed of 2.3% international airports that are strongly connected and resilient, yet, even if those were removed, around 90% of the world's airports would remain interconnected. Therefore they classify the world into clusters based on the physical proximity of airports and conclude that the weak periphery plays a crucial role as those serve airports without any alternative routes.

For efficiency, Lin and Ban (2014) studies 20 years of network system changes, illustrating the dynamics of the American airline system and the evolution of its real physical networks focusing on centrality measures and efficiency, computed as a global index based on the shortest path length. They find out that the backbone of the system was constructed in the early stages of its inception and have remained the same in spite of the network growth and new cities being added mainly due to geopolitical causes. Efficiency was improved over the years with the addition of shorter and medium length routes observing a decrease in the betweenness centrality throughout the period, but they also found a negative correlation with the size of the network and efficiency.

5. Characterizations of the carrier's networks

Finally, another extensive research topic revolves around the physical organization of the network. Authors paid special attention to how the famous Hub & Spoke (H&S) networks or the Point to Point structures differ and interact from each other, trying to understand their differences since the type of system determines a carriers' cost structure and connectivity routes. Hub and Spokes systems were first introduced as early as 1950. They rely on a central hub, whose objective is to collect all passengers traveling, to then distribute them to their final destinations (the spokes), following a star-like pattern. At their inception they were considered the most efficient structure that reduced costs and maximized airloads and they are highly dependent on the maximization of frequency and economies of scale. With the increase in competition and the introduction of Low Cost Airlines (LCCs) who operate in a more direct, point to point network that serves profitable routes, much research has been performed to find the "best" operating structure.

For example, Lederer and Nambimadom (1997) studied the optimal network design for profit maximizing airlines, by creating a model that seeked to minimize an airlines and passengers' cost, measured by travel times. They concluded direct structures are beneficial when distances are small or demand of the route is very high. If these two do not hold, then a hub and spoke network would be optimal. They also find that the number of cities have an impact on the optimality of a network, if it serves a small number, direct networks are more optimal, whereas a hub and spoke is better for a large number of cities.

Marti et al. (2015), analyze the efficiency and financial situation of Spanish airlines by comparing which of the two systems manages resources better and measure productivity between carriers depending on the system in which they operate. They find that large H&S airlines rate lower on efficiency scores measured as their operating income (labor costs, assets and supplies). However they are not able to conclude which system is better, since airlines that operate in P2P networks show drastic differences, some carriers score as "perfect" in the efficiency score, while others show the worst performance among all carriers.

Generally, these studies are limited by the fact that most carriers nowadays have a mixed system and do not fully behave like a perfect H&S or P2P, and most binary systems of recognition fail to capture this duality. For this reason Roucolle et al. (2017) defined 3 sets of continuous indicators analyzing the evolution of the U.S. Network and studying the similarities of LCC and Legacy networks doing a Principal Component Analysis and then studying the evolution of the networks. The first component corresponds to six variables related to node centrality and their indices and they call it *NetHub* since it points to its global structure in the network and how close the behavior is to a Hub and Spoke. The second is made out of four variables: density, transitivity, eigenvector centrality and centralization and they call it *NetWeave* since it presents the availability of alternative routes, and finally *NetSize* relative to network segments and nonstop routes. With this information, they characterize the network of all major airlines depending on the behavior of the three components and define the carriers as Hub & Spoke with one Hub, Hub & spoke with multiple hubs, Point to Point, or Circular.

They conclude that LCC airlines score higher in NetHub on average, although they seem to be converging over time as Legacy airlines are increasing their Nethub scores over time. Conversely, in size Legacy carriers present a higher score, but LCCs have been growing over time. The main difference is shown with NetWeave, where LCCs provide greater alternative routes. Finally, there seems to be a convergence of the business model of both types of carriers, confirming the mix of systems of airlines that focus less on form, and more on profit maximization.

III. METHODOLOGY

The research is composed of a four step process: the first is to perform basic graph metric analysis to the U.S. passenger network to understand its topology and evolution throughout a 25 year period. The second relies in finding the node centrality measures to determine which cities are the most important and whose removal could alter the structure and soundness of the network. Thirdly, the Louvain Community Algorithm is used to determine underlying communities in the data to deep dive in the influence of regional centers and their interactions in the network to finally follow Guimera and Amaral (2004) methodology that measures intra-community and inter community scores to characterize the nodes in 7 distinct discrete categories to evaluate the network in this new light and draw conclusions.

In general, the research follows a similar structure to Guimera et al. (2005) who studied the composition of the global airline network using data of more than 800 airlines and 3,993 cities, to characterize the worldwide transportation system and analyze the influence of community structures and nodes. They found anomalies in the data where the most central cities are not the most connected ones and conclude that community structures show the geopolitical factors that shape the patterns in the network. Their paper proposes both a way to characterize the network, and classify the role of individual cities in light of their community interactions.

For the U.S. this characterization is important since regional interactions are not exclusive to worldwide networks, but trade and transportation have been found to have regional country patterns. Finding out if the network is evenly distributed and cohesive is the first step to identify if there are better ways to organize or structure the network altogether. And by finding influential nodes within regions allow for better planification and preparedness. The expectation is that intra-geopolitics, as opposed to the author’s findings, do not have such a strong effect within a country given that they all share the same culture, history and political system, and that networks developed inside the country are shaped by geographical locations and population. Finally, the main contribution of the research is to study the network in a comprehensive way by combining the traditional approach of centrality measures and adding a community structure approach to fully grasp the dimensions and interactions of the system.

1. Network Characterization:

To measure network characteristics and dynamics, the clustering coefficient is calculated as the sum of edges connecting a vertex's neighbors over the total number of possible edges between the vertex's neighbors (Hansen et al. 2020).

Mathematically, it accounts the number of edges connecting the neighboring nodes as $C_i = \frac{2\, E_i}{k_i(k_i - 1)}$ Where k_i is the number of edges connecting the node i with its neighbors, and E_i is the number of edges between neighbors of nodes i.

The Clustering coefficient then is computed as $C = \frac{1}{N}\sum_{i=1}^{N} C_i$

A higher number implies a more direct network system, whereas a low clustering coefficient would mean more of a hub and spoke structure. Put it differently, a 1 in clustering coefficient would mean that all the cities are connected, whereas a value close to 0, implies they are not.

To understand the role of the nodes in the system three centrality measures will be calculated: The Degree Centrality measures the influence of a node with its local environment as it sums the number of direct flights passing through each city. The higher the degree, the higher the connectivity.

$$C_D(v) = \frac{\sum_i vij}{N-1}$$

The Betweenness Centrality measures the influence of a node in the flow of information of the system. it detects the influence of a node on how the information passess through it. As a formula it measures the number of times the node lies on the shortest paths of all other nodes in the network or its "ability to make connections to other groups in the graph" (McKnight, 2013).

$$C_B(v) = \sum_{s,t \in V} \frac{\sigma(s,\, t|v)}{\sigma(s,\, t)}$$

Where V is the set of nodes, $\sigma(s,\ t)$ is the number of shortest paths, and $\sigma(s,\ t|v)$ is the number of those paths passing through the node v.

The Eigenvector Centrality measures the level of influence of a node within the network, relative to the influence of the closest connections of the node. If its closest connections have a strong influence then the node will possess a high score. It is then possible to have a low degree centrality, but to have a high eigenvector centrality, In social networks, a person who has a large number of unpopular friends, for example (Goldbeck, 2013), or in an airline system, a small city that is connected to only big cities.

It is calculate through the adjacency matrix $Ax = \lambda x$. Where A is the adjacency matrix of the graph with an eigenvalue λ.

IV. DATA DESCRIPTION AND MANIPULATION

1. DATA

The Data set was obtained from the Bureau of Transportation Statistics of the United States, from the Database Air Carrier Statistics (Form 41-Traffic) focused only on the national carriers. Specifically, the data was obtained from the “T-100 Domestic Segment (U.S. Carriers)”, which contains yearly non-stop segment data and was taken for a 25 year period, from 1997-2022.

The database contains aggregated information reported on a monthly basis by carrier on and route with the number of departures scheduled and performed, number of passengers and seats, flight distance, airtime and ramp to ramp time origin and destinations city and airports.[3]

During the period, several mergers and acquisitions happened in the industry therefore the unique_carrier code was used to avoid double counting. For example, Delta and Northwest in 2008 announced their merger, others changed their operations.. Below is a table of the mergers.

AIRLINES & UNIQUE CARRIER	Merger Date Announcement	Change in Reporting	UNIQUE CARRIER
MERGERS & ACQUISITIONS			
US Airways (“US“) & America West (“HP“)	2005	Oct 2007	“US”
Delta (DL) & NorthWest (“NW “)	2008	Jan 2010	“DL”
Continental Micronesia (“ CS“) & Continental Airlines (“CO“)	Dec 2010	Jan 2011	“CO“
Atlantic Southeast (“EV “) & ExpressJet (“XE”)		Jan 2012	“EV”
United (“UA“) & Continental (“CO“)	2010	Jan 2012	“UA”
Southwest (“WN”) & AirTran (“FL”)	2011	Jan 2015	“WN”
American Airlines (“AA”) & US Airways (“US”)	2013	July 2015	“AA”
Alaska (“AS”) & Virgin America (“VX”)	2016	April 2018	“AS”
NAME CHANGES WITH UNIQUE CODES REMAINING THE SAME			
Pinnacle to Endeavor (“9E”)	N/A	Aug 2013	“9E”
American Eagle to Envoy (“MQ”)	N/A	April 2014	“MQ”

Table 1: Mergers, acquisitions and transitions

1.1 Data aggregation

Data was obtained on a yearly basis and added together. After merging, the database for the 25 year period contained 9,582,235 unique rows of information. In the United States, during 2001, there was a strong restructuring and reformation of the Airline sector, therefore, variables like Carriers, Origin/Destination Cities and flights, show an important growth during that year.

The general data shows 272 unique carriers where 248,868,185 flights took flight offering 25,269,592,783 seats, transporting a total of 18,864,925,869 passengers who traveled

[3] See Data snapshot in Annex

7,555,150,186 miles to get to their destinations. It is estimated by Statista (2022) that between 12-20% of travel for business purposes in the past 3 years.

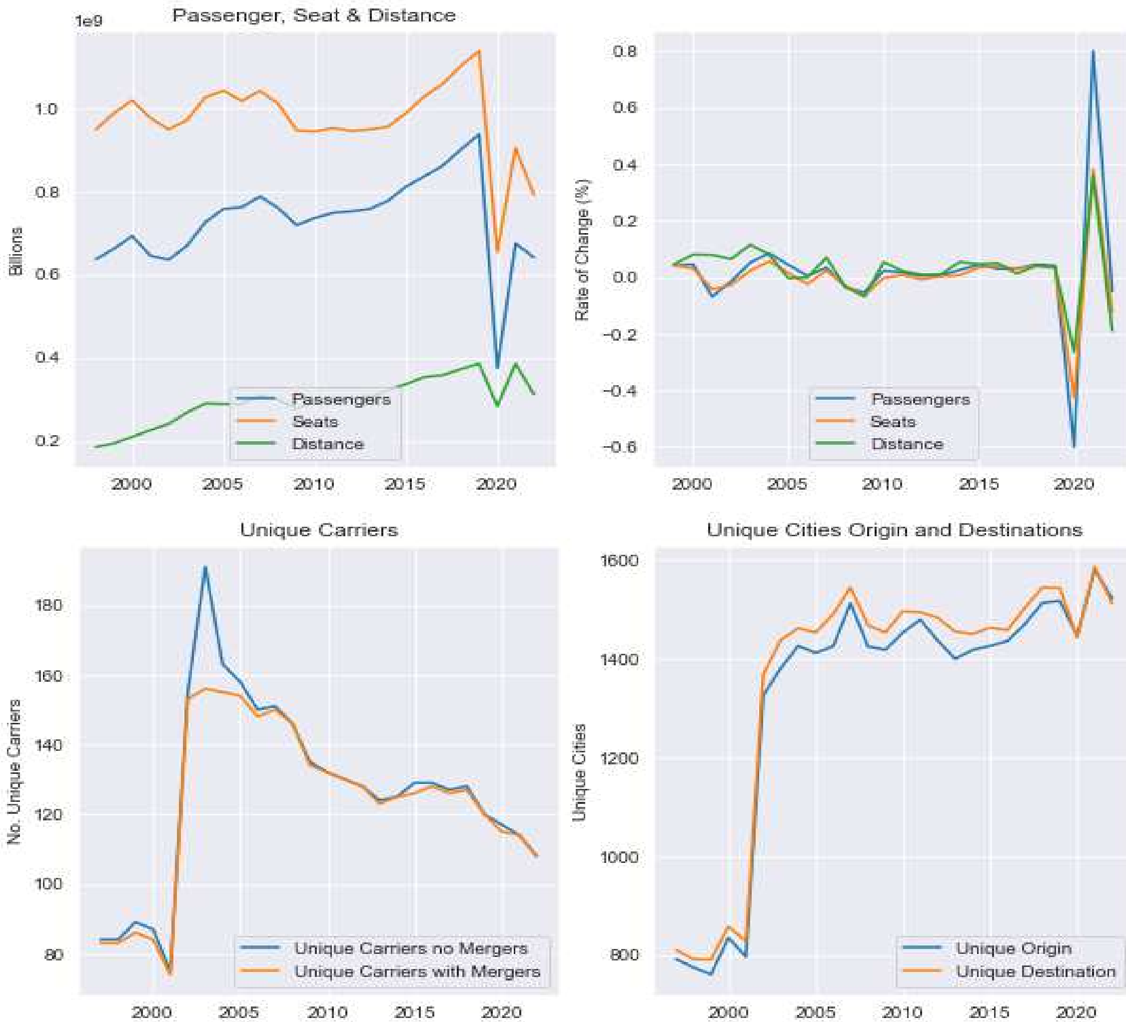


Figure 1: Evolution of Basic Metrics

Overlooking the yearly trends, the passengers traveling (and the seats offered) had a positive growth trend all until pre Covid levels, with the exception of a ditch in 2001 and 2002 given the 9-11 attacks that lowered demand and halted some flights. This is the same year when the U.S. underwent some regulatory transformations.that required every aircraft, in spite of its size, to be registered to the FAA (U.S. Department of Transportation, 2000) & U.S. Code house of Rep (2004). Thus the number carriers registered an increase of 106% from 2001 to 2002, and kept increasing 23% the following year. This also impacted the origins and destination points, given that regional airports and smaller locations started to play a role on the network, but as seen in the data, passenger, seats, and distance did not increase, given the smaller size aircrafts.

From 2002 onwards, that trend has been decreasing steadily leaving only 91 carriers registered up to the timing of the study. A major drop was seen in 2008, where the world recession hit every sector in the economy, causing carriers to shut operations, and throughout the period, the consolidation of operations, where fuel prices, shortage of pilots, and economies of scale have contributed to this consolidation.

In terms of reach, the network has also expanded throughout the years. Overall it has served 3,097 unique origin- and 3,154 destination cities. Both variables' behavior have similar trends, showing the biggest expansion in 2021. Given the consolidation of the carriers and the expansion in reach, it can be said that each carrier has expanded and matured to a more robust network, offering a wider destination list. The U.S. network is an ever evolving creature; cities' origins and destinations appear and disappear over time, depending on the demand and airline carriers' offerings and needs, expanding and contracting the network and certain cities gaining popularity changing the network structure as time goes by.

1.2 Data Cleaning and Manipulation

The airline industry in the U.S. is massive, from military aircrafts, to personal propeller planes, it is said that it is the most developed airline industry in the world (Xu & Harris, 2008). Therefore, there is an important classification that the U.S. Bureau of Transport uses to identify carrier groups and non-cargo national airlines dependent on the yearly operating revenues of the airline:

- Major Carriers.- Revenues > $1,000,000,000 usd
- National Carriers - Revenues between $100,000,000 - 1,000,000,000
- Large Regionals Carriers -Revenue between $20,000,000 and $99,000,000
- Medium Regional Carriers - Revenue < $20,000,000 or less than 60 seats.
- Commuter Carriers - Air taxi operators
- Small Certified without any specification - Maximum seating capacity of < 61 seats

Apart from this a definition for Domestic Only - All Cargo Carriers and International Carriers are also specified in the Data. Therefore and from the extent of the study, data was cleaned and selected as follows:

- Flights that were scheduled, but not departed and had 0 passengers (19% of the data entries were dropped).
- Entries where the country of origin or destination were different from the U.S. Given that the purpose of the study is to understand the networks within the US, any trip originating from the outside is assumed to be a returning flight (438,824 trips were coming from outside of the US).

After the first data clean, more than 6.8 million entries remained distributed as follows:

ID	DESCRIPTION	Number Entries	Total Passengers	Unique Carriers
1	Large Regional Carriers (Annual revenue of $20 to $100 million)	168,007	95,427,905	68
2	National Carriers (Annual revenue > $100 million to $1 billion)	1,163,200	2,073,662,090	59
3	Major Carriers (Annual revenue over $1 billion)	4,219,953	14,242,813,865	27

4	Medium Regional Carriers (Annual revenue under $20 million)	41,685	8,303,849	30
5	Small Certificated Carriers (carrier holding certificate issued under 49 U.S.C. section 41101 and operating aircraft designed to have a max seating capacity of ≤60 seats or a max. payload of 18,000 pounds or less.)	932,688	90,341,618	53
6	Commuter Carriers (air taxi operator which performs at least five round trips per week between two or more points and publishes flight schedules which specify the times, days and plans which such flights are performed.)	304,065	211,558,400	68
7	Commuter Carriers (Air Taxi providing Essential Air Service)	4,951	66,507	10

Table 2: Carrier's Classification

* Passengers and unique carriers are the sum or total count for the 25 year period.

Since the database is aggregated by month, new variables were created to identify trends:
To determine the average number of passenger and seats per flight:

- Passengers per flight (PPF): Number of Passengers/Departures Performed
- Seats Per Flight dividing (SPF): Number of Seats/Departures Performed

To determine the Occupancy rate of Flights

- Occupancy Rate (OR): Passengers / Seats.

The new variables' distribution add some information on the behavior of the data. The PPF and PFS show a bimodal-like histogram that reflects the presence of smaller carriers that carry less than 100 seats, and thus less people. Both metrics have the highest frequency at the 10 or less seats/passengers, showing a greater distribution of smaller aircrafts. For the OR, data is left skewed, where on average, flights have a vacancy of 32%, while the majority of flights are full to 90% of its capacity. This data makes sense, since airlines work to maximize their occupancy rate with pricing schemes, this is specially true to airlines that serve whose network is direct (as opposed to hub-and spoke) whose main form of increasing profit is through an increase demand, or a higher occupancy rate Lederer and Nambimadom, (1998)

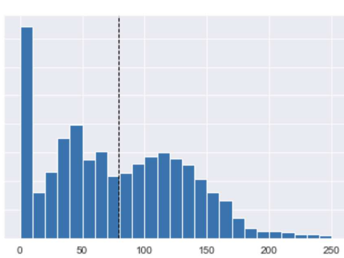


Figure 2: PPF Histogram

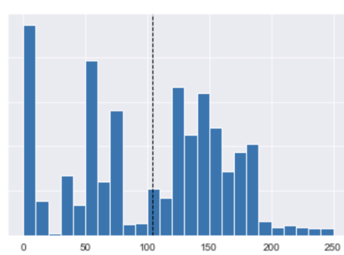


Figure 3: SPF Histogram

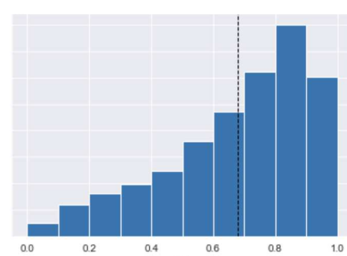


Figure 4: OR Histogram

1.3 Major Carriers

To narrow down the scope of the study, only 27 airlines who are categorized as major carriers were considered and only kept data of those who had more than 10 years of data during the 25 year period (see Annex for timeline). This accounts for 17 carriers, that translates to 3,877,498 rows of data pertaining to the following airline carriers with data ranging from the data stated in the data Period:

IATA CODE	Name of Airline	Classification of carrier	Years	No. Years
5Y	Atlas Air	Charter Services	2010-2022	13
AA	American Airlines	Legacy	1997-2022	26
AS	Alaska Airlines	Legacy	1997-2022	26
B6	JetBlue Airways	Low Cost Carrier	2005-2022	18
CO	Continental Airlines	Legacy	1997-2011	16
DL	Delta Airlines	Legacy	1997-2022	26
EV	ExpressJet Airlines	Regional and charter	2006-2010	11
F9	Frontier Airlines	Low Cost Carrier	2007-2022	16
HA	Hawaiian Airlines	Legacy	2009-2022	14
HP[4]	America West Airlines	Regional	1997-2007	11
MQ	Envoy Air	Regional	2000-2022	23
NK	Spirit Airlines	Low Cost Carrier	2013-2022	10
NW	Northwest Airlines	Legacy	1997-2009	13
OO	Skywest Airlines	Regional	2006-2022	17
UA	United Airlines	Legacy	1997-2022	26
US	U.S. Airways	Legacy	1997-2015	19
WN	SouthWest Airlines	Low Cost Carrier	1997-2022	26

Table 3: Airline Codes & Classification

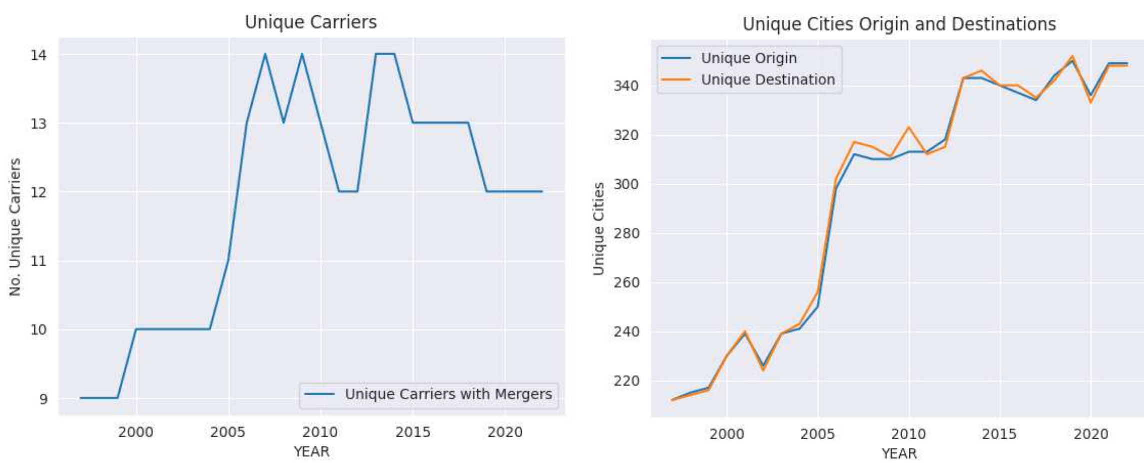


Figure 5: Major Carriers evolution, Count, Origin & Destination

During the 25 year period there were 4,219,532 unique non-stop passenger flights between 28,693 distinct city pairs, 516 distinct origin cities, and 530 destination cities.

2. DESCRIPTIVE STATISTICS

The new data reduces the variability, the focus on major carriers, shed a different light on its underlying structure. The distribution behaves more like a normal distribution with a mean of 101, meaning that on average, the Major Carriers Transport 100 passengers per route. Meanwhile capacity (measured as seats), have a mean of 136 seats per flight, with some big

[4] Merged with U.S. Airways and reported jointly in 2007.

aircrafts ranging up to more than 250, but with less presence in the market. In general, capacity ranges from 100-189 seats per flight. In terms of the occupancy rate, distribution is constrained to maximum capacity, and on average the historical flights have been 75% full to capacity.

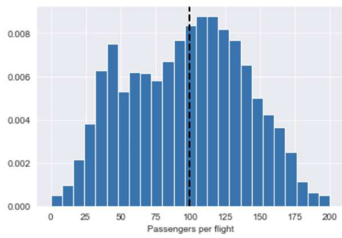


Figure 6: Major Carriers' PPF Histogram

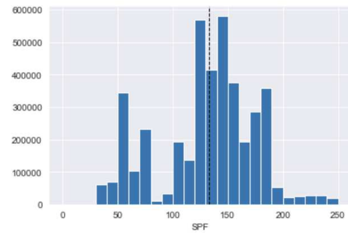


Figure 7: Major Carriers SPF's Histogram

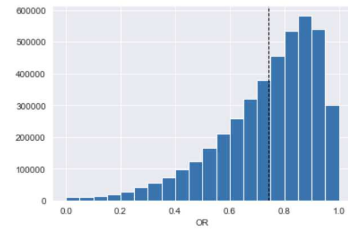


Figure 8: Major Carriers' OR Histogram

A common practice in the industry is to leave only information from the third quarter to avoid seasonality effects (Roucolle et al, 2022). By removing all but the data for 3Q we remove high peak travelings like Thanksgiving, Christmas,and Vacation Breaks when there's a peak in travel. This reduces the data sample to less than 1 million rows and easier computation.

index	DEPARTURES SCHEDULED	DEPARTURES PERFORMED	SEATS	PAX	DISTANCE	AIR_TIME
count	993901	993901	993901	993901	993901	655421
mean	36.087	35.752	4686.222	3615.224	851.795	3815.835
std	49.362	48.026	6664.522	5271.539	623.139	5759.612
min	0	1	26	1	0	0
25%	2	3	344	242	389	264
50%	24	24	2656	1866	689	1834
75%	53	52	6020	4668	1105	5101
max	1,654	1,018	116,785	95,777	5,812	157,772

Table 4: Basic Statistics

On an airline level the one with the most passengers, capacity, trips, and departures performed, is Southwest airlines, followed by all the major Legacy Airlines. Although the comparison is unfair given the time period they have been in operation. This gets normalized by analyzing the SPF or PPF variables. In terms of occupancy rate (Passenger/Seats) computed per route, the airlines who have an overall higher score are the Low Cost Airlines (JetBlue, Frontier and Spirit) which goes in line with the definition to maximize capacity to reduce average costs. The Average Seat Mile (ASM) was also computed to be able to compare all airlines, given their time in the market and their capacity. In this case, Hawaiian airlines is outstandingly leading the average seat miles, however, geography plays an important role given that the airline travels at least 2,200 miles (3,450 km) to get to any other city in North America, followed by American Airlines and then JetBlue.

CAR	PAX	SEATS	DISTANCE	AIR_TIME	DEPT	PPF	SPF	OR	ASM
5Y	23.1	78.7	273.3	33.7	0.29	21.84	71.70	40%	300.53
AA	553,406.9	697,209.9	92,901.4	381,936.8	4,610.64	11,360.47	14,500.72	78%	8,256.78
AS	130,715.2	169,248.0	42,184.7	114,203.5	1,156.98	4,307.82	5,732.43	74%	4,800.08
B6	113,372.0	138,835.8	25,384.1	136,293.3	1,018.88	2,505.20	3,184.46	79%	7,199.74
CO	130,620.1	169,527.4	44,956.8	66,991.8	1,215.81	4,752.87	6,269.68	75%	4,341.29
DL	637,542.7	809,741.2	126,308.1	368,860.7	5,087.76	17,839.90	23,226.66	77%	5,107.48
EV	48,427.3	61,131.9	14,617.9	70,433.7	1,134.65	1,347.16	1,712.82	79%	934.23
F9	59,515.0	70,661.0	24,749.2	51,646.4	428.60	3,563.19	4,484.76	80%	2,683.74
HA	31,585.0	40,431.1	5,496.6	21,292.3	263.11	523.28	649.44	79%	14,581.42
HP	56,352.8	76,355.2	14,819.5	12,609.6	548.66	1,483.42	2,035.41	73%	5,004.20
MQ	82,372.7	113,555.1	23,761.7	108,025.5	2,290.49	1,887.33	2,593.83	72%	959.27
NK	57,499.1	70,283.1	19,481.1	47,906.8	392.29	2,731.50	3,473.74	79%	3,582.84
NW	150,827.0	204,388.7	37,622.2	42,115.4	1,535.42	5,001.51	7,044.99	70%	3,200.14
OO	124,851.2	160,740.4	38,654.7	193,650.3	2,812.12	3,404.71	4,394.21	77%	1,135.99
UA	444,369.5	558,268.4	151,546.6	310,547.9	3,626.38	16,910.29	21,694.44	78%	4,751.13
US	214,197.3	293,923.8	47,025.4	103,244.4	2,182.68	6,388.54	8,973.09	71%	3,376.66
WN	757,498.3	1,023,261.5	136,816.9	471,186.4	7,229.04	19,144.66	25,411.51	75%	3,844.21

*In Thousands.

Table 5: Top Carriers basic information

On a yearly basis the trends show a steady increase in the passenger carried with the exception of Southwest, Delta, American Airlines ad United, that show greater than average incrementality, this is explained however, by the mergers and acquisitions, as stated earlier:

- Southwest in 2015 reported AirTran flights
- Delta in 2010 reported Northwestern flights
- Americans in 2015 reported the U.S. Airways flights
- United in 2012 reported Continental flights

The passenger growth trend for the smaller regional airlines followed a steady and slow growth path from 2008 - 2019, until Covid impacted the market. For 2022, the recovery is almost 100% to all of the carriers, with the exception of Delta (Fig. 9). Eventually, airport and route capacity might start playing a role, if airport capacity is maximized, the only way to cope with increased demand will be through bigger airplanes, the growth rate would start resembling the replacement aircraft rate.

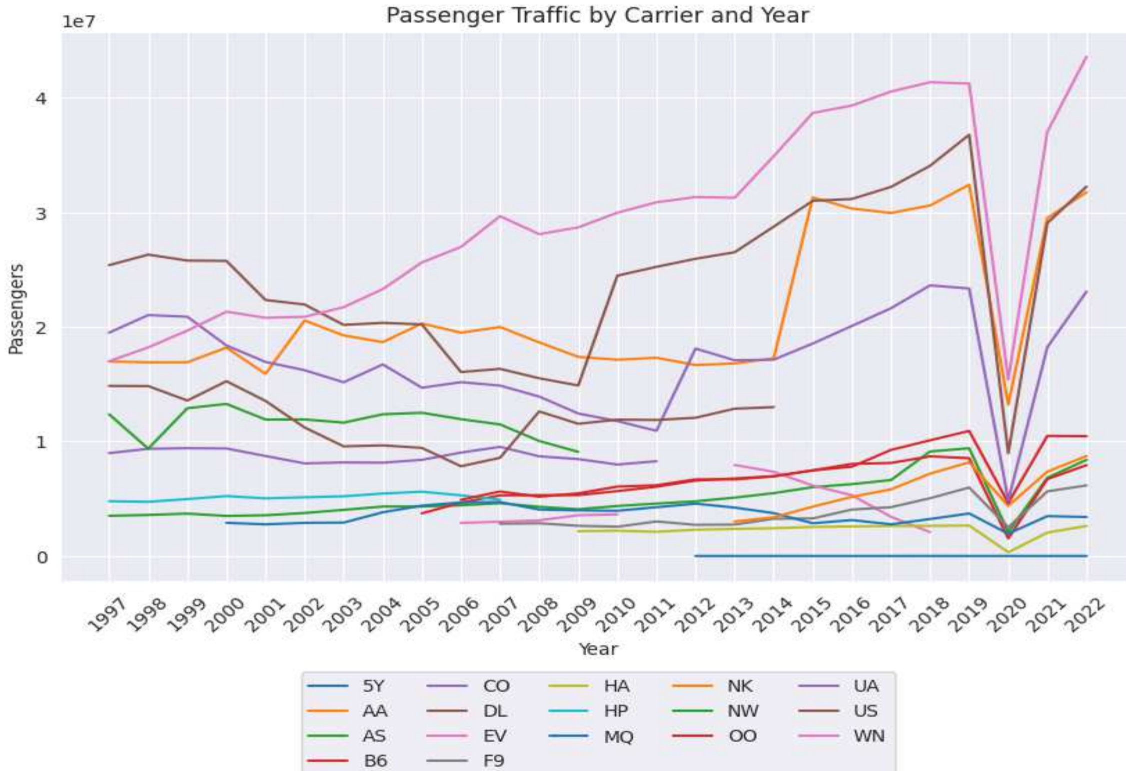


Figure 9: Passengers Volume per Major Carrier

V. RESULTS

Network theory is based on the studies of nodes that are connected through edges whose dynamics can be analyzed. In airline transportation studies the majority of the research considers cities or airports as nodes, and the flights as edges (an exception is Sun et al (2014) who proposed flight paths as edges). For this study, nodes are taken on a city level, that is, if a city has more than one airport, i,e, New York; Laguardia, and JFK, the values are aggregated to one city. The edges are direct, non-stop flights.

1. TOPOLOGY & GLOBAL NETWORK DYNAMICS

The U.S. network is massive, composed of 459 nodes and 11,290 edges. The diameter of the network, measured as the longest distance to travel from the two farthest nodes, is 5 flights. The average path length for all destinations, computed as the summation of all the shortest paths divided by all pairs is 2.1, meaning that on average, it would take 2 flights to arrive at any destination and showing the presence of cities as hubs. Overall, 74% of the cities can be reached in 2 or less flights, and only 12 cities are 5 flights away from each other. Among them are Rota and Saipan located in the US Pacific Islands. For them it would take a total of 5 flights to get to Anniston, AL, Fort Drum, NY or Beaufort, SC.

The network is dense, as shown by the average clustering coefficient, where 66% of two connecting nodes have a neighbor that is also connected to one another. This allows passengers to have more alternatives to travel from one city to the next, and is specially important in case of airport failures, bad weather or even terrorist attacks Verma et al., (2014).

The network shows the small-world property, given its high values of clusterings and low values for shortest path, and is also inline with previous literature. For comparison, a random graph with the same number of nodes and edges shows a diameter of 3, a shortest path length of 1.9, but a clustering coefficient of 0.10.

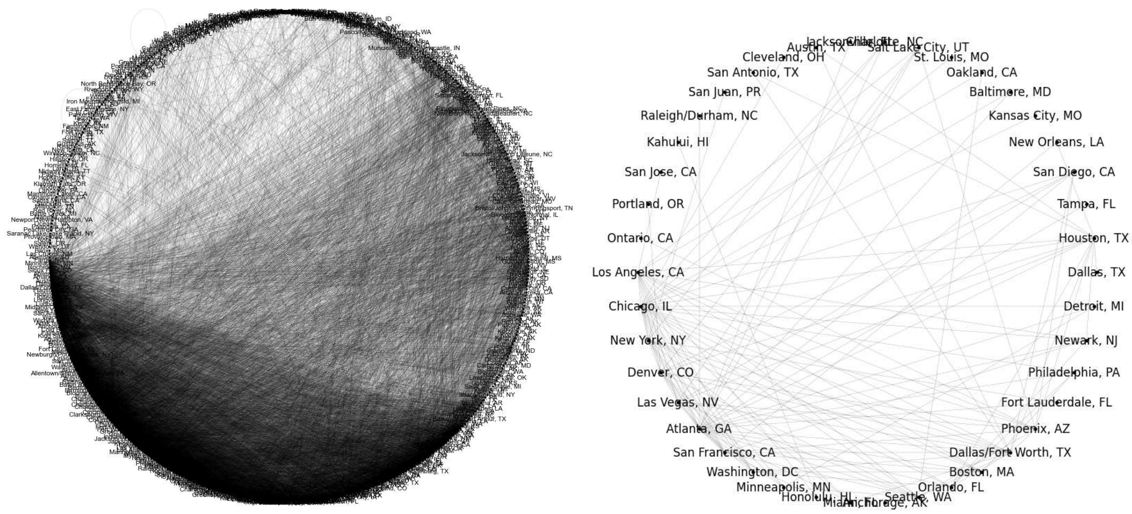


Figure 10: Small World Graph

Above, the left image shows a representation of the network connectivity, highlighting the small world phenomenon. The circumference of the circle contains every city in the network, and each line is an edge showing the nonstop flight from two cities. For visual clarity, a subsample of the top 250 connections are represented on the right hand side. The cities with the most connections are the expected cities, those known for their importance as hubs, economic centers or tourist destinations.

2. NODE NETWORK DYNAMICS

The previous section characterizes the network dynamics as a global and gives a preview of how large and cohesive the U.S. airline network is. However, node dynamics allows for a better understanding of how the network works, as it shows the importance of individual players. It helps answering questions such as: Who are the most connected cities? Where do most passengers fly to? What are the cities most crucial to the network? All these questions are only answered by analyzing individual network's nodes and centrality measures.

Three measures of degree centrality are obtained: degree centrality that shows the proportion of direct flights connecting through each node; betweenness centrality, that measures the number of times the node acts as a bridge for the shortest path from other nodes; and eigenvector centrality that evaluates how influential is a node in the system, in spite of its size, by measuring how influential are the nodes that it is connected to. It is expected that these three centrality measures are correlated, a city with higher degree centrality must lie in more shortest paths of other nodes, and would be better connected to other influential cities.

To measure the centralities, a list of city pairs (origin, and destination) was set adding the number of passengers aggregated throughout the years to understand the overall importance of each city in terms of the passengers it carries. Here direction was considered to measure one way flights. A total of 28,693 pairs of cities were found in the data, averaging 496,386 passengers transported. The pair of cities where more passengers were transported, was Los Angeles CA, to Chicago IL.

Origin City	Destination	Passengers
Los Angeles, CA	Chicago, IL	9,706,076
Chicago, IL	Los Angeles, CA	9,524,572
New York, NY	Chicago, IL	9,258,890
Chicago, IL	New York, NY	9,176,461
Chicago, IL	Denver, CO	8,629,687
Denver, CO	Chicago, IL	8,525,974
Los Angeles, CA	New York, NY	8,051,676
New York, NY	Los Angeles, CA	7,988,678
Las Vegas, NV	Los Angeles, CA	7,948,265
Los Angeles, CA	Las Vegas, NV	7,898,195
New York, NY	Atlanta, GA	7,865,974
Los Angeles, CA	San Francisco, CA	7,806,463
Atlanta, GA	New York, NY	7,702,647
San Francisco, CA	Los Angeles, CA	7,611,555
Washington, DC	Chicago, IL	7,414,230

Table 6: Passenger Volume per route

Then, an edge list was created to measure the importance of the city in the network. There are 25 cities who lie within the 95th percentile of number of connections and on average, cities have 49 connections. The most connected cities are Chicago (286), Denver (250), Dallas/ Fort Worth (238), Atlanta (231), Houston (219). There is not a strong correlation between population size and number of connections (R^2= 0.35), however, those cities who are present within the 95th percentile in population size, are also present in the 95th percentile of the ones with the most connections. Basically, these cities are centers of gravity that are known for attracting people, industries, education and tourism.

With the edge list it is possible to compute the centrality measures corresponding to the data and find the most central cities as shown below.

	Cities	Degree		Cities	EigenVector		Cities	Betweenness
1	Chicago, IL	0.62	1	Chicago, IL	0.13	1	Chicago, IL	0.1
2	Denver, CO	0.55	2	Dallas/Fort Worth, TX	0.12	2	Denver, CO	0.07

3	Dallas/Fort Worth, TX	0.52	3	Atlanta, GA	0.12	3	Anchorage, AK	0.05
4	Atlanta, GA	0.5	4	Houston, TX	0.12	4	Los Angeles, CA	0.05
5	Houston, TX	0.48	5	Denver, CO	0.12	5	Atlanta, GA	0.04
6	Los Angeles, CA	0.46	6	Washington, DC	0.11	6	Minneapolis, MN	0.04
7	Minneapolis, MN	0.45	7	Detroit, MI	0.11	7	San Francisco, CA	0.04
8	Washington, DC	0.44	8	New York, NY	0.11	8	Dallas/Fort Worth, TX	0.04
9	New York, NY	0.43	9	Los Angeles, CA	0.11	9	Seattle, WA	0.04
10	Detroit, MI	0.43	10	Minneapolis, MN	0.11	10	Washington, DC	0.03

Table 7: Top 10 Cities for every Centrality Measure

The three main metrics show Chicago as the city with the most relevance in the U.S. Network, being placed first in the three centrality measures. It can be observed that 62% of all the U.S. flights have a connection through Chicago. Interestingly, Chicago is only 3rd in the U.S. for population after New York, and Los Angeles, however the relevance of these two cities lag behind in importance in comparison to other less populated cities.

There is a cluster of cities that are consistent in all centrality measures and connected to one another, therefore the eigenvalue measures are so narrowly close by, while the cities relevance order remains almost unchanged. These city airports are all within the top 15 of the BTS 2022 based on unplanned passengers, or included in the Top 3 tiers of Ryerson (2013) hub airports classification. The presence of such a dense interconnected core of cities within the network represents a solution to disruptions in one of the cities and a dispersion on the dependency of the country in a single airport.

This is also true even if the networks is weighted for other parameters, Chicago tops the list for passenger, seats, distance, departures performed, and then evidently PPF and SPF, the only exception is the Occupancy Rate which is second only after Denver Colorado and the ASM where comes second after Los Angeles. But the cities Los Angeles, Dallas/ Fort Worth Tx, Atlanta, GA, Houston, TX, Denver and New York, NY are the cities who are represented at the top five places of all the weighted metrics. As mentioned earlier, it is clear that these cities are known to be airport hubs of the most important Airlines of the U.S.

Analyzing the betweenness centrality, the one city that stands out is Anchorage, which scores low in the rest of the metrics. Given its relative location to other cities in the network within its region, it serves as the distribution center to other smaller cities in the province and it connects its people to the rest of the U.S.. But given its size, weather conditions and distance to the mainland, its positioning serves as a feeder to remote locations. This is also in line with finding that Guimera et al. (2005) encountered studying the global network, they argue that this is due to the geographical and political scene, that requires the peninsula to be tied to mainland, and it is this reason why it lacks connections to closer cities in Canada.

If we allow the betweenness centrality to be weighted for distance, the nodes that show the highest metrics are Portland, Oregon and Seattle, Washington. These are two cities that are geographically closer to Alaska and also serve as connecting airports for Alaska Airlines, Therefore the score in betweenness centrality can be derived from two reasons, they represent the main cities for Alaska Airlines, 83% of the flights of Alaska network have a connection with Seattle, and 44% with Portland.

Plotting the above mentioned metrics on a map, it is visible how the degree centrality is highly correlated to the Eigenvector centrality and a bit skewed to the East. Whereas in terms of Betweenness centrality seems to follow a more balanced distribution across the whole territory, including Alaska and Hawaii given the presence of important regional airports.

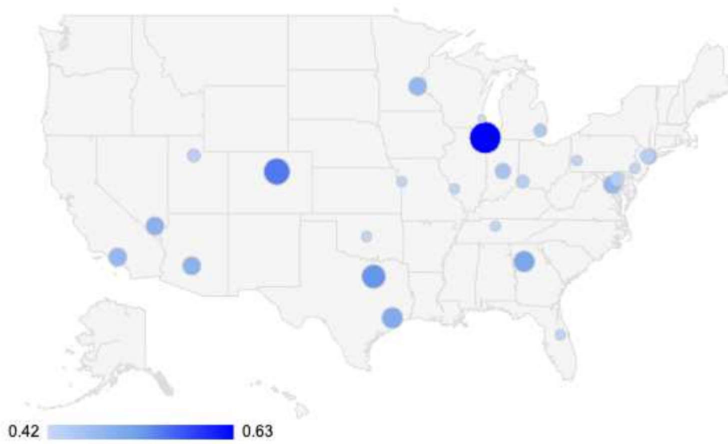


Figure 11: Degree Centrality; Top 25 Cities

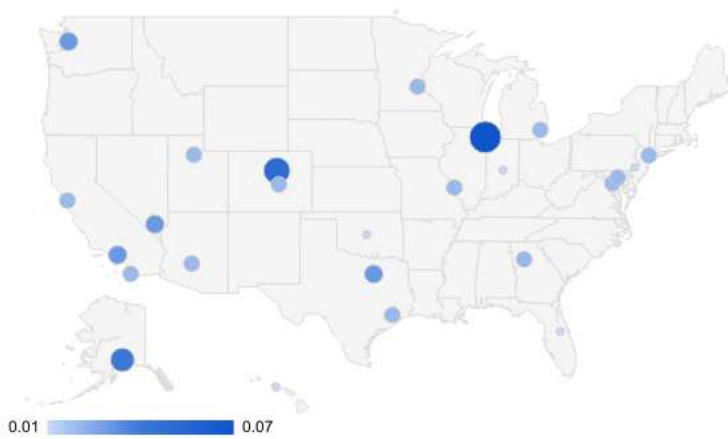


Figure 12: Betweenness Centrality; Top 25 Cities

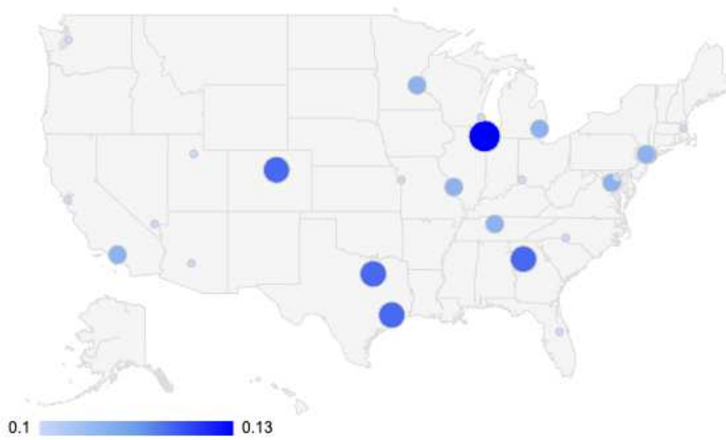


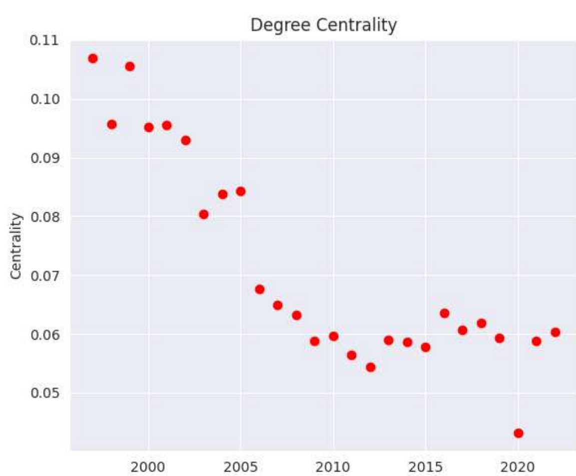
Figure 13: Eigenvector Centrality; Top 25 Cities

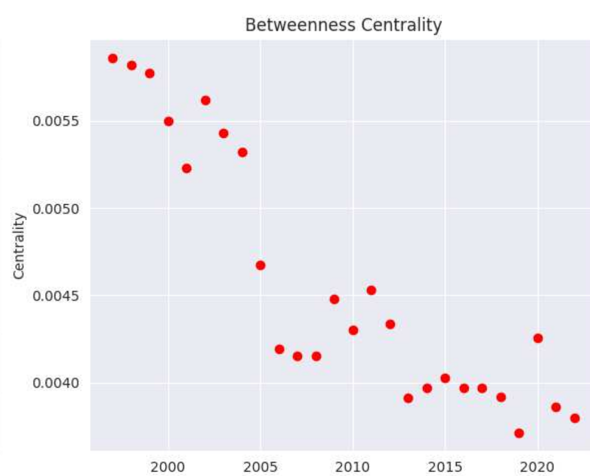
3. NETWORK EVOLUTION

To avoid the issue of misrepresenting the network dynamics, the evolution of the node's centrality measures were plotted and analyzed. The data shows a consolidation of the network that was impacted by the mergers and restructurings that pushed for economies of density.

Specifically, the average values of all three metrics show a drop that was exacerbated in 2005, when fuel prices rose 48% impacting airlines' operations, increasing costs and forcing them to mainstream their networks in the light of a decrease in profit and a rise of competition. Throughout that period, the market share of Low Cost Carriers (LCCs) rose to 45% impacting the centralization measures, given that most of those airlines operate in a point-to-point network scheme; while 3 legacy carriers were operating under Chapter 11 FAA (2005), forcing them to eliminate redundant routes, and improve its costs structure to ensure profit.

This consolidation, however, represents a more balanced network, with a robust, and diversified system that is less concentrated among one single node or airline. This also gives the network greater shock tolerance in the long-term (Lin & Ban 2013).





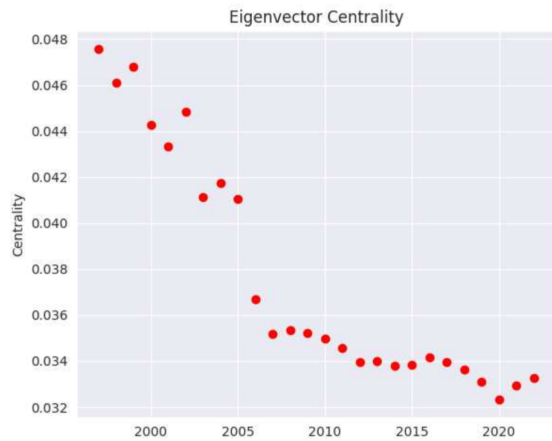


Figure 14: Evolution of Average Degree Centralities in the Network

On an individual level, cities also evolved in their influence in the network as seen in Fig. 15. One example is Atlanta that, in from 2006 to 2009, had a significant increase in all centrality metrics driven mainly by the efforts of Delta Airlines who in the brink of bankruptcy pushed for a network restructuring, adding 100 more flights departing from Atlanta to be the city with the "most destinations than any other Hub" (Delta Airlines 2016).

Another example is Denver, CO, that shows a steady increase in its centrality metrics, to place itself among the top 3 cities in the last five years. Denver's airport has been growing in size and the city has an ideal mid-country location, which makes it favorable to a central airport. In 2003, an extra runway allowed bigger aircraft to land, and in 2006, Southwest airlines decided to add Denver as a major hub (Southwest Airlines 2006). Nowadays, it also serves as a major airport for United, Southwest, and Frontier Airlines.

This city view level, is important because it show some polarization of the metrics, while the majority of the top 10 cities loose importance in the network over the last five years, the centrality measures consolidates in three cities Chicago, Dallas and Denver, who even show some resilience during the Covid-19 period. This evolution also shows the hub dynamics throughout time and how difficult it is for a hub to gain influence.

The consolidation of the centrality measures to a smaller group of cities (Fig. 15) can be partially explained by what Amaral et al. (2000) call preferential attachment. In which new nodes in an expanding network tend to connect to nodes with highly connected degrees. Or even more, explained by the Banconi Barbas model of complex networks, where nodes are assigned a *fitness* quantitative measure, that explains why some nodes are better at acquiring new nodes through their evolution than others in spite of having the same number of connections (Bianconi & Barabási, 2001). In the airline network, and specially in one that is so robust as is the American one, adding a new node implies that the city is small and therefore, demand needs to be sourced from multiple cities. It is easier, then, for an airline to cover operating costs connecting it to a high density city and sourcing demand, allowing them to exploit density economics. Taking Atlanta again as an example, it is easier for an airline to take it as a hub, if it is already the hub of another airline, the infrastructure is already in place, demand is constant, and a cross traffic between airline can source more passenger volume maximizing the probability of a route to be a success.

Figure 15: Evolution of Degree Centrality Measures

4. COMMUNITY STRUCTURE

The evolution of the metrics, the consolidation into fewer nodes, and the distribution of the betweenness centrality raises the question of how geography plays a role in the system? Are those cities that are most connected, the most central? What patterns can be seen in the data?

4.1. Anomalies in the network

It is expected from a random graph to have strong correlation between centrality and its degrees, those nodes with more connections would also possess a high betweenness centrality. However, this is not necessarily the case for the U.S. airline network. To prove this, I followed the normalization of the betweenness centrality weighted by distance used by Guimera et al (2005), as $Bc\,norm = \frac{bi}{avg(Bi)}$ where bi is node i, to correlate both metrics. The data shows both; cities with very high degrees that have lower than average betweenness centralities or not very central, suggesting the existence of nodes with direct flights but far away, while also showing some cities with higher than average betweenness centrality with fewer connections. This is further evidenced when factoring the amount of passengers or departures. Data shows that some important cities score low on betweenness centrality, making the most connected cities, not necessarily the most geographically central.

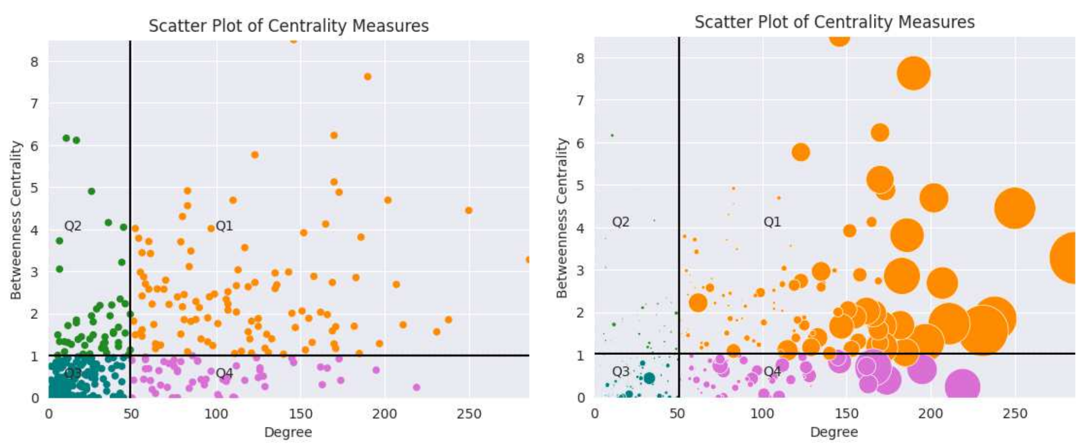


Figure 16: Betweenness Centrality weighted by Distance (Average)

Figure 17: Betweenness Centrality weighted by Distance (Top 25)

This can be further explained if instead of average, we control for the top 25 cities as shown in Fig. 17. The four quadrants are divided as before but with the cut up value being the number of the 25th most connected city on the x axis, and the 25th city with the highest centrality on the y axis. It would be expected that the top 25 most connected cities are also the most central, graphically the expectation is to have the majority, if not all, of the data placed in the first quadrant, however the data shows interesting results.

There are 18 cities that are within the top 25 in betweenness that are not the most connected. There are only 7 cities who make up both, the highest score in centrality and degree: Denver, CO, Washington, DC, Seattle, WA, San Francisco CA, St. Louis MO, Baltimore, MD, and Kansas City, MO. However, there exist some cities in the *Quadrant 2* that present very high centrality, with very few nodes that are anomalous, and vice versa. Implying that the most central connected cities are not the most geographically central. This is related to the existence of regions with a high density of airports but few connections outside its region, and confirms the existence of regional centers.

	Top Betweenness (Q3)	Top Degree (Q4)
1	Portland, OR	Chicago, IL
2	Sitka, AK	Dallas/Fort Worth, TX
3	Everett, WA	Atlanta, GA
4	Sacramento, CA	Houston, TX
5	Rapid City, SD	Los Angeles, CA
6	Lewiston, ID	Minneapolis, MN
7	Sioux Falls, SD	New York, NY
8	Champaign/Urbana, IL	Detroit, MI
9	Duluth, MN	Salt Lake City, UT
10	Fayetteville, NC	Phoenix, AZ
11	Oklahoma City, OK	Newark, NJ
12	Provo, UT	Charlotte, NC
13	Laramie, WY	Milwaukee, WI
14	Columbus, GA	Boston, MA
15	Memphis, TN	Orlando, FL
16	Missoula, MT	Cincinnati, OH
17	Cordova, AK	Nashville, TN
18	Billings, MT	Colorado Springs, CO

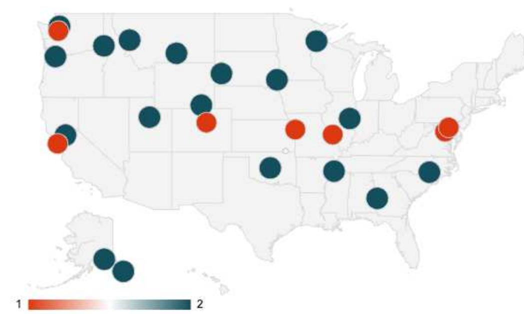


Figure 18: Top Betweenness

Figure 19: Top Degree Map

4.2 Unweighted network

Analyzing the unweighted network, shows a better correlation of the Degree to the Betweenness centrality, and few things stand out, first, the better linear correlation of the data, second, the mismatch between the weighted network (by distance) and the unweighted one in number of cities that lie within each quadrant, and finally, the fewer cities have small degrees but high betweenness centrality and can be translate as non-hub cities. Accounting for passengers traveled, there is an obvious correlation in the network between the variables.

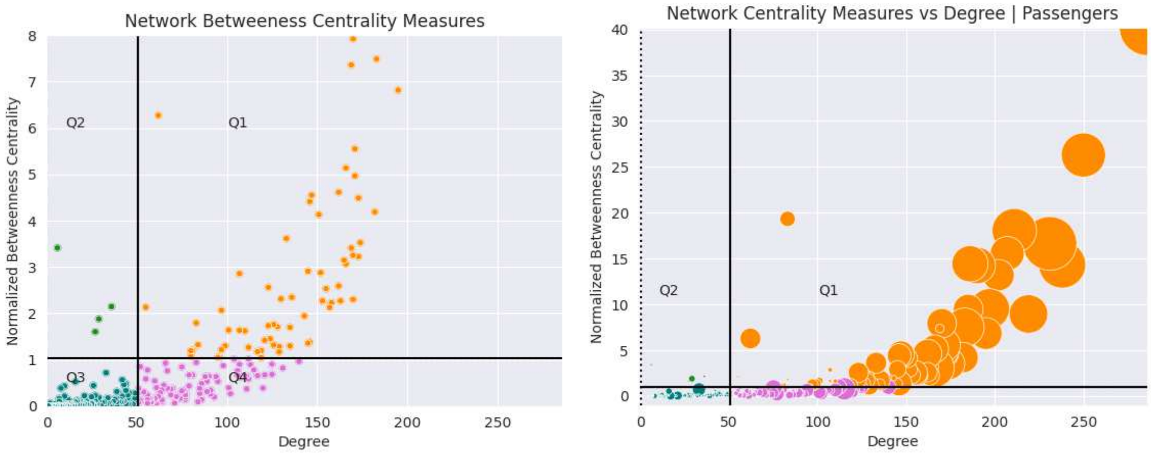


Figure 20: Betweenness Centrality unweighted (Average)

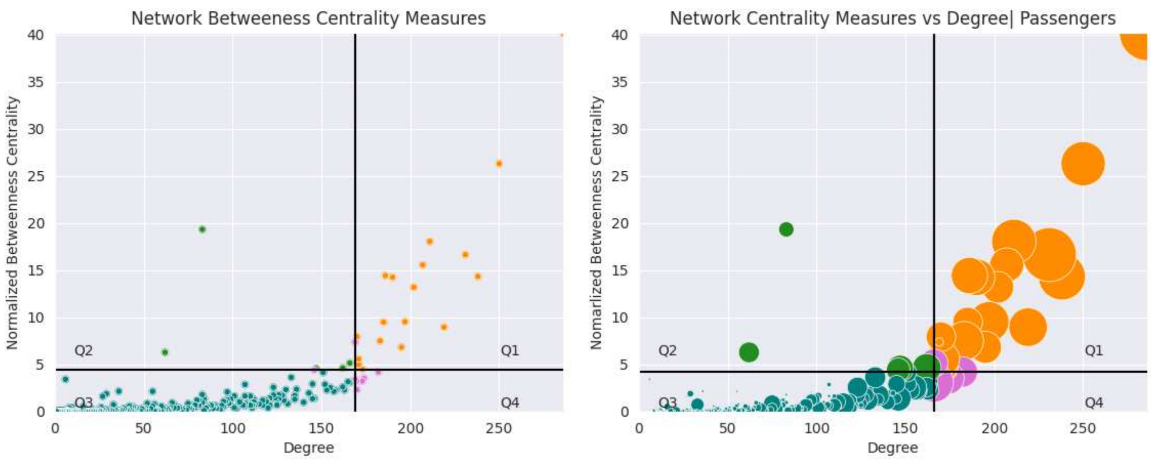


Figure 21: Betweenness Centrality unweighted (Top 25)

However, two cities clearly stand out having high Betweenness centrality and very low degrees: Anchorage, and Honolulu, and relates to the first centrality measures. These nodes are the most important cities in non-continental regions, and "play a role in congestion and diffusion and the cohesiveness of complex networks". Guimera et al (2005) highlight geopolitics as responsible for these anomalies, however, in the case of a country wide network, where language, geography, historical connections and political affinity should not play such a stronger role, the hypothesis of what causes these differences is linked to regional community structures.

4.3 Community Structures and their interactions with one another

To deep dive in these data outliers, it is interesting to understand what are the community structures within the network and delve into their influence in regional cohesion among some nodes to explain local behavior. The Louvain Community algorithm, developed by Blondel et al (2008), groups the nodes that share some characteristics in a network to find "families" in the data or data that shares similar structures or patterns to understand group behavior and dynamics (see Annex).

Using an unweighted graph, the algorithm found 6 communities as depicted in Fig. 22. Similar tests were performed with weighted networks for seats, passengers and departures, which yielded similar community results.

For reference the regions will be refered as following:
1. Alaska
2. Midwest
3. South
4. West (including Hawaii)
5. East
6. Pacific

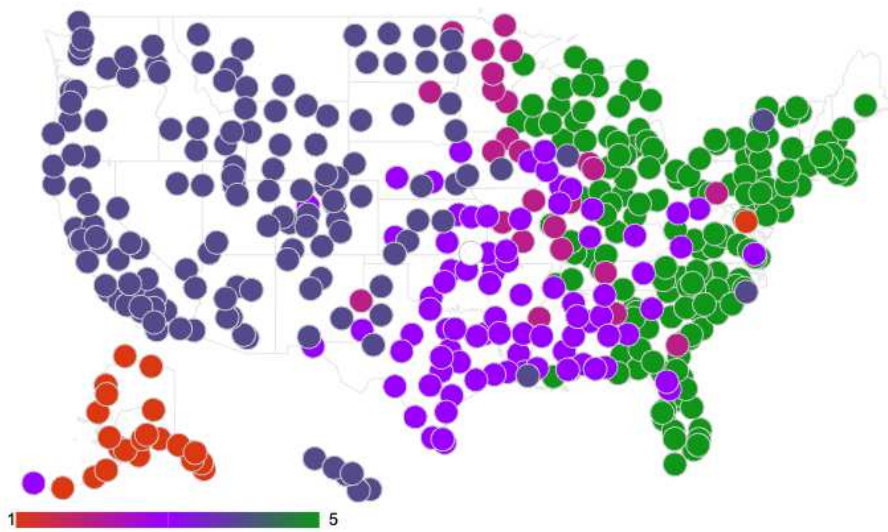


Figure 22: U.S. Community Classification

Note: the 6th region is composed of 3 cities: Guam, Rota and Saipan, located in islands in the North Pacific Ocean and are not depicted above. It can be seen that the cities are distributed geographically.

The classification shows a geographical pattern. This distribution can be explained given that modularity[5] is assigned by the edge's connections and as discussed before, the nodes with the highest degrees are: Chicago, IL (Midwest), Denver, CO (West), Dallas/Fort Worth, TX (South), Atlanta, GA (East), Houston, TX (South). The reason why some cities in the center of the country are classified as South, instead of Midwest, for example, is due to the increment in modularity. Chicago is so dense, that the marginal gain of modularity with a small node, is minimal, in comparison to the southern region; a denser node would have less gain, with the addition of another node to its community.

[5] Modularity measures the density of the connections within clusters versus the density of the connections between clusters.

4.4 Regional Role of Cities

With these structures and the previous information on centrality measures, it is useful to understand what is the role of each city within and outside their community. Guimera and Amaral (2005) develop a methodology to measure the intra community and intercommunity. The intra-community score measures how well connected a node is to other nodes within its community, and is computed by applying a Z-Score while a Participation score measures the node's connectivity outside its community. Both are computed as follows:

$$Intracommunity\ score:\ \ z_i = \frac{\kappa_i - \overline{\kappa_{si}}}{\sigma_{si}}.$$

Where κ_i measures the number of links to all other nodes within the community, $\overline{\kappa_{si}}$ the average of κ over the nodes in the community (s_i) and σ_{si} its standard deviation, and the appearance of hubs and non hub cities are seen.

$$Participation\ score:\ \ P_i = 1 - \sum_{s=1}^{N} \left(\frac{\kappa_{is}}{k_i}\right)^2.$$

Where κ_{is} is the number of links of the node *i* to all nodes in the community s and k_i is the total degree of the node. The result will be closer to 1 if the links are uniformly distributed among all communities and zero if all its links are within its own community. This is the case for example, with Saipan, in the Pacific who has two nodes connecting to the other two cities within the Pacific (Rota and Guam), and no other connections. Therefore the only κ_{is} is 2 which and the nodes in the complete network k_i=2 as well, therefore P_i =0.

These two metrics allow each city to be classified into discrete categories. Firstly, if the Z-score is greater than 2.5, a node is defined as a hub given its high interaction and influence among its region. Then, measuring the relationship between the participation coefficient a node can be classified as follows:

For non hubs:

- R1:Ultra -peripheral nodes- all the nodes are within its module. (P≤0.05)
- R2: Peripheral- nodes- at least 60% of the links are within the module. (0.05<P≤0.62)
- R3: Non-hub connectors- has half of its links within the module. (0.62<P≤0.80)
- R4: Non-hub kinless nodes- has 35% of its links within the module. (P>0.80)

For hubs:

- R5: Provincial hubs- has ⅚ of its links within the module (P≤0.30).
- R6: Connector hubs- has at least half of its links within the module (0.30<P≤0.75)
- R7: Kinless hub- has less than half its links within the module (P>0.75).

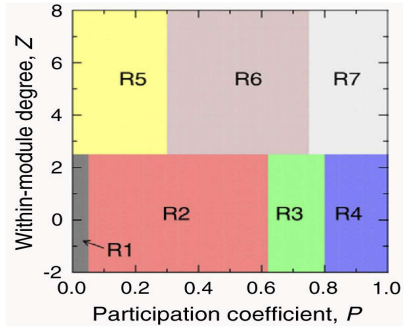


Figure 23: Intra and Inter Community Score Categorization
Source: Guimera & Amaral, 2005.

The intracommunity scores, showed 12 cities as hubs: Anchorage, AK, Atlanta, GA, Chicago, IL, Dallas Fort Worth, TX, Denver, CO, Houston, Tx, Los Angeles, CA, Salt Lake City, UT, San Francisco, CA, Seattle WA, Phoenix, AZ, and Washington, DC.

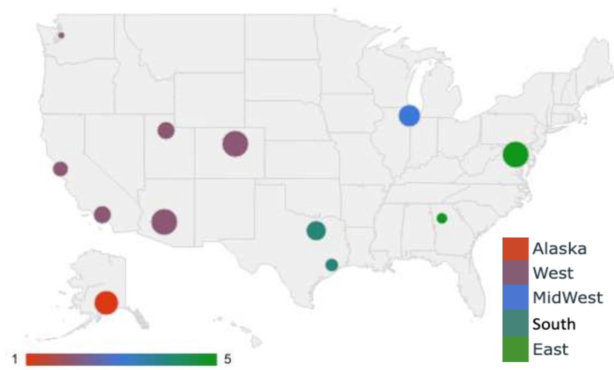


Figure 24: Hubs per Region

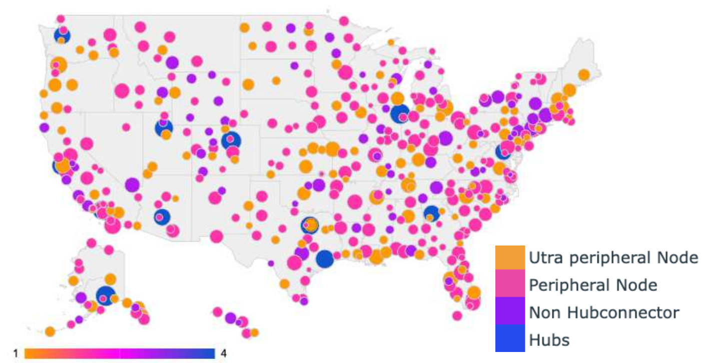


Figure 25: Nodes Categorization

This classification shows that each region, with the exception of the Pacific, has at least one connector hub, however, the methodology for the U.S. Network fails to categorize different types of hubs. All 11 cities are classified as “connector hubs” without finding any provincial or kinless hub. It is also important to note that this hub classification differs from its traditional definition. The FAA classifies hubs by the proportion of passengers the node carries per year, while other authors establish a more robust definition, for example Ryerson and Kim (2013) mentioned earlier who incorporate hub hierarchy and frequency to classify nodes as hubs.

The data also shows that there are 3% of cities that are hubs in the country where 50% are located in the West region. More surprisingly, is that 81% of the nodes are either ultra peripheral (27%) or peripheral nodes (55%). iEven looking within each region, this number ranges from 75%, in the Midwest, to 100% in the Pacific. Finally, the only region that holds the most non-hub connector nodes is the Midwest, with 21%, but on average, only 14% of the cities present this structure, showcasing that in general, within a region, if the city is not a hub, it has few interactions with cities outside of its regional boundaries (See Annex).

This classification also highlights some differences with the classification found through the centrality measures. In spite of the majority of the cities matching in both the extent or influence of a city varies greatly. One clear example is Anchorage, that shows a high influence both inter and intra community making it one of the most important cities in the network. But in real life it tells a different story. It is a city that has a massive importance to its region, but it is not likely for someone in continental U.S. to use it as a hub to reach any destination outside Alaska. Another example is Salt Lake that is not present with the centrality measures, however, has even a higher score than Washington, DC. Most importantly, there is a correlation of the behavior of each regional group in terms of how they score in both metrics as seen in Fig. 26.

City	Community	P-Score	Z-score
Seattle, WA	West	0.678	2.716
Atlanta, GA	East	0.633	2.896
Houston, TX	South	0.676	3.039
San Francisco, CA	West	0.642	3.290
Salt Lake City, UT	West	0.628	3.385
Los Angeles, CA	West	0.647	3.433
Dallas/Fort Worth, TX	South	0.676	3.664
Chicago, IL	Midwest	0.692	3.933
Anchorage, AK	Alaska	0.714	4.078
Denver, CO	West	0.647	4.390
Phoenix, AZ	West	0.656	2.572
Washington, DC	East	0.604	2.624

Table 8: Z-scores and Participation Scores

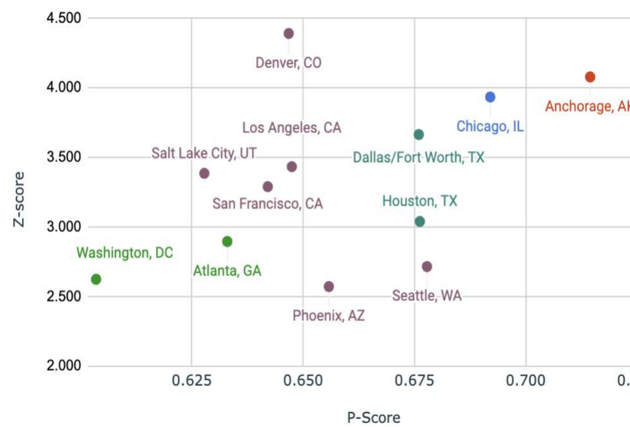


Figure 26: P-S Scores Scatter Plot

It can be concluded that even within country networks, regional community structure appears evident and important. In spite of having no geopolitical boundaries per se, and being governed under the same political system, there is a stronger attachment to creating links within the region strengthening the regional core and hub cities, and thus distance plays a strong role. This finding is important, because it highlights that geography does play a role and again preferential attachment has a regional influence too.

This yields an important confirmation about the structure on the network and the importance of the peripheral cities. However, this presents only half the picture, since the regional structure does not convey transit patterns but only ties among nodes. In the data, there is a big mismatch between the structure of the network and the amount of passengers traveling per modality of nodes and region. Those 3% hub cities transport 41% of the passengers in the United States, and the ultra peripheral and peripheral nodes only carry 18%, and 32% of the passengers respectively. And more impressively, 14% of the travels are between the 12 cities considered hubs. The good news, however, is that there is some capacity for building flexible networks given the existing structure of the network and rely more on peripheral existent nodes.

The regional differences become more stark when adding this factor into play. In the Midwest, for example, 73% of the passengers travel through the hub city (Chicago, IL) which amounts to almost 230 million passengers or 6% of the U.S. Passenger Volume in the period. The South, Alaska and the West show similar trends, where 56%, 58% and 54% of the passengers travel in their hub airports respectively, with the difference that the South has 2 hubs both in Texas, and the West has 6, including los Angeles, CA, Denver, CO and Seattle, WA. The big difference in regions is the East where only 19% of all the travels happen in hubs, there is confirmation then that communities behave differently from one another, and some internal politics and dynamics can also play a part.

The possible explanation for the distinct behavior is the type of cities that are within the regions and its economic activities. The East Region which carries 43% of the passengers is composed of cities like New York and Washington D.C, that holds the political and economic power, therefore, travel patterns have been strong and contained between both cities without the need of transiting from one region to the next. Whereas, the Northwest, composed mainly of one big city, Chicago IL, and smaller cities or rural areas and tied by geographical boundaries of the Big Lakes and Canada, depends on flying outside of the region to perform economic activity.

VI CONCLUSION

After analyzing the topology of the U.S. airline network for the past 25 years and the centrality and evolution of its nodes, the study confirms that with the newest data available, the network still behaves as a small-world one, with a high clustering coefficient and a small betweenness centrality, as it had been proved before. The most important cities which score highest in centrality networks and that remained relevant in spite of external shocks, are Chicago, Denver, and Dallas. However, there exists some nodes that are the most central in the network but are not the most connected, and this is shown with both, an unweighted network, or weighting by distance, showing some anomalies in the system.

Following the methodology of multi community structure by Guimera et al, I find that even within a country, regional connectivity is present and important, and plays a role in how the people move and travel. It is shown as well how some cities are connected only within their regional boundaries. The network structure highly relies on the mix between ultra-peripheral nodes, peripheral nodes, non-hub connectors and hubs. Where all regions present about half of its nodes as peripheral nodes. However, in spite of these classifications there is a big mismatch between regional structure and the volume of people that travel through them. Some regions are highly dependent on their hub cities to function, like the Midwest, in spite of having only one regional airport as hub, while some others have a more balanced structure both as a network and as passenger flows like the West region, and its hypotheses that the reason is the types of cities that are within each region.

These findings also highlight that certain regions are more prone to risk, both internally within the region, and externally among the others. For example, the Midwest heavily relies on its hub, and a failure of the airport could imply important economic losses and an absence of flexible routes or airports to carry on operations, while the East region has much more flexibility and a balanced network structure and dispersion of passengers being more robust to external shocks.

Some discussion may arise from these findings, the first is that the regional classification was given by the Louvain algorithm which may impact how regions are determined. Some other algorithms have been found to also be efficient in their partition and could provide different network communities and yield different results (Mandal, 2022). The second is that the weights of the network could change the centrality of the cities and yield different results. If instead of distance, one measured air-time or flight path, to account for a real travel distance and not as "as the crow flies", and even to which extent should only the contiguous U.S. be only considered given that Alaska, Hawaii and the Pacific Islands are geographically farther away, and can bias the results.

However, the findings are still relevant for policy. If we are to ask how a network could be better planned, both for efficiency, that is to say to reduce fuel costs and passenger times, as well as to build a more robust system, it would make sense to have the most central cities as the most connected ones. The fact that networks have a regional imbalance, highlights how the economic development and migration patterns also occur and how some cities are the center of gravity for economic activity and growth. Delving into the underlying reasons for the existence of these nodes of cities that power the country, and why they evolve as they do in the first place, is an interesting topic for further research.

VII. BIBLIOGRAPHY

Aldous, J. M., & Wilson, R. J. (2000). Graphs and applications: An Introductory Approach. In *Internet Archive* (1st ed., pp. XI, 444). Springer. https://archive.org/details/graphsapplicatio0000aldo/page/30/mode/2up

Amaral, L. A. N., Scala, A., Barthelemy, M., & Stanley, H. E. (2000). Classes of small-world networks. *Proceedings of the National Academy of Sciences*, *97*(21), 11149–11152. https://doi.org/10.1073/pnas.200327197

Bagler, G. (2008). Analysis of the airport network of India as a complex weighted network. *Physica A: Statistical Mechanics and Its Applications*, *387*(12), 2972–2980. https://doi.org/10.1016/j.physa.2008.01.077

Bianconi, G., & Barabási, A.-L. (2001). Competition and multiscaling in evolving networks. *Europhysics Letters (EPL)*, *54*(4), 436–442. https://doi.org/10.1209/epl/i2001-00260-6

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/p10008

Broido, A. D., & Clauset, A. (2019). Scale-free networks are rare. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-019-08746-5

BTS. (2022). *Airport Rankings 2022 | Bureau of Transportation Statistics*. Www.bts.gov. https://www.bts.gov/topics/airlines-and-airports/airport-rankings-2022

Cheung, D. P., & Gunes, M. H. (2012, August 1). *A Complex Network Analysis of the United States Air Transportation*. IEEE Xplore. https://doi.org/10.1109/ASONAM.2012.116

Civil Aviation Authority. (2016). No-Frills Carriers: Revolution or Evolution? In *Civil Aviation Authority* (pp. 1–98). https://publicapps.caa.co.uk/docs/33/CAP770.pdf

Cook, G. (1996). A Review of History, Structure, and Competition in the U.S. Airline Industry. *Journal of Aviation/Aerospace Education & Research*, *7*(1). https://doi.org/10.15394/jaaer.1996.1183

da Rocha, L. E. C. (2009). Structural evolution of the Brazilian airport network. *Journal of Statistical Mechanics: Theory and Experiment*, *2009*(04), P04020. https://doi.org/10.1088/1742-5468/2009/04/p04020

Dai, L., Derudder, B., & Liu, X. (2018). The evolving structure of the Southeast Asian air transport network through the lens of complex networks, 1979–2012. *Journal of Transport Geography*, *68*, 67–77. https://doi.org/10.1016/j.jtrangeo.2018.02.010

De Paula , A. (2017). Econometrics of Network Models. In *Advances in Economics and Econometrics: Volume 1* (Cambridge University Press, pp. 268–323.). Cambridge University Press.

Delta Airlines. (2006). Delta Air Lines – Plan of Reorganization. In *DeltaAirlines*. https://web.archive.org/web/20070104050922/http://images.delta.com.edgesuite.net/delta/pdfs/speeches_pres/investor.pdf

Erdős, P., & Rényi, A. (1959). On the Evolution of Random Graphs. *Publicationes Mathematicae Debrecen*, *6*(290). https://static.renyi.hu/~p_erdos/1960-10.pdf

FAA. (2005). FAA Aerospace Forecasts Fiscal Year 2006-2017. In *Federal Aviation Administration* . U.S. Department of Transportation Federal Aviation Administration. https://www.faa.gov/data_research/aviation/aerospace_forecasts/2006-2017/media/Review%20of%202005.pdf

Golbeck, J. (2013, January 1). *Chapter 3 - Network Structure and Measures*. ScienceDirect; Morgan Kaufmann. https://www.sciencedirect.com/science/article/abs/pii/B9780124055315000031

Guida, M., & Maria, F. (2007). Topology of the Italian airport network: A scale-free small-world network with a fractal structure? *Chaos, Solitons & Fractals*, *31*(3), 527–536. https://doi.org/10.1016/j.chaos.2006.02.007

Guimerà, R., & Amaral, N. (2005). Cartography of complex networks: modules and universal roles. *J Stat Mech*. https://doi.org/10.1088/1742-5468/2005/02/

Guimera, R., Mossa, S., Turtschi, A., & Nunes Amaral, L. A. (2005). The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *Proceedings of the National Academy of Sciences*, *102*(22), 7794–7799. https://doi.org/10.1073/pnas.0407994102

Hansen, D. L., Shneiderman, B., Smith, M. A., & Himelboim, I. (2020, January 1). *Chapter 6 - Calculating and visualizing network metrics* (D. L. Hansen, B. Shneiderman, M. A. Smith, & I. Himelboim, Eds.). ScienceDirect; Morgan Kaufmann. https://www.sciencedirect.com/science/article/abs/pii/B9780128177563000066

Hossain, Md. M., & Alam, S. (2017). A complex network approach towards modeling and analysis of the Australian Airport Network. *Journal of Air Transport Management*, *60*, 1–9. https://doi.org/10.1016/j.jairtraman.2016.12.008

ICAO. (2017). *List of Low Cost Carriers* (pp. 1–7). https://www.icao.int/sustainability/documents/lcc-list.pdf

Jia, T., Qin, K., & Shan, J. (2014). An exploratory analysis on the evolution of the US airport network. *Physica A: Statistical Mechanics and Its Applications*, *413*, 266–279. https://doi.org/10.1016/j.physa.2014.06.067

Lederer, P. J., & Nambimadom, R. S. (1998). Airline Network Design. *Operations Research*, *46*(6), 785–804. https://doi.org/10.1287/opre.46.6.785

Li, W., & Cai, X. (2004). Statistical analysis of airport network of China. *Physical Review E*, *69*(4). https://doi.org/10.1103/physreve.69.046106

Li-Ping, C., Ru, W., Hang, S., Xin-Ping, X., Jin-Song, Z., Wei, L., & Xu, C. (2003). Structural Properties of U.S. Flight Network. *Chinese Physics Letters*, *20*(8), 1393–1396. https://doi.org/10.1088/0256-307x/20/8/362

Lin, J., & Ban, Y. (2014). The evolving network structure of U.S. airline system during 1990–2010. *Physica A: Statistical Mechanics and Its Applications*, *410*, 302–312. https://doi.org/10.1016/j.physa.2014.05.040

Mandal, T. (2022, August 12). *A Comparative Analysis of Community Detection Algorithms* -. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2022/08/a-comparative-analysis-of-community-detection-algorithms/

Markusen, A. R., Noponen, H., & Driessen, K. (1991). International Trade, Productivity, and U.S. Regional Job Growth: A Shift-Share Interpretation. *International Regional Science Review*, *14*(1), 15–39. https://doi.org/10.1177/016001769101400102

Marti, L., Puertas, R., & Calafat, C. (2015). Efficiency of airlines: Hub and Spoke versus Point-to-Point. *Journal of Economic Studies*, *42*(1), 157–166. https://doi.org/10.1108/jes-07-2013-0095

McKnight, W. (2013). *Information ManagementStrategies for Gaining a Competitive Advantage with Data*. Elsevier Science .

Milo, R., Shen-Or, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., & Alon, U. (2002). Network Motifs: Simple Building Blocks of Complex Networks. *Science*, *298*(5594), 824–827. https://doi.org/10.1126/science.298.5594.824

Roucolle, C., Seregina, T., & Urdanoz, M. (2020). Measuring the development of airline networks: Comprehensive indicators. *Transportation Research Part A: Policy and Practice*, *133*(1), 303–324. https://doi.org/10.1016/j.tra.2019.12.010

Ryerson, M. S., & Kim, H. (2013). Integrating airline operational practices into passenger airline hub definition. *Journal of Transport Geography*, *31*, 84–93. https://doi.org/10.1016/j.jtrangeo.2013.05.013

SouthWest Airlines. (2006, January 3). *Southwest Airlines Begins Service in Denver, Announces Additional Flights and Destinations*. Investor Relations ; Southwest Airlines. http://investors.southwest.com/news-and-events/news-releases/2006/03-01-2006

Statista. (2022, July). *Number of business and leisure travelers US 2008-2022* . Statista; Statista. https://www.statista.com/statistics/207103/forecasted-number-of-domestic-trips-in-the-us/

Sun, X., Wandelt, S., & Linke, F. (2014). Temporal evolution analysis of the European air transportation system: air navigation route network and airport network. *Transportmetrica B: Transport Dynamics*, *3*(2), 153–168. https://doi.org/10.1080/21680566.2014.960504

U.S. Department of Transportation. (2000, April 11). *Federal Aviation Administration SORNs* . Www.transportation.gov. https://www.transportation.gov/individuals/privacy/privacy-act-system-records-notices/DOT-FAA

U.S. Department of Transportation. (2022). *Domestic Airline Consumer Airfare Report | US Department of Transportation*. Www.transportation.gov. https://www.transportation.gov/policy/aviation-policy/domestic-airline-consumer-airfare-report

U.S.Code House of Rep. (2004). *CHAPTER 441—Registration and Recordation of Aircraft*. Uscode.house.gov; Office of the Law Revision Counsel of the United States House of Representatives. https://uscode.house.gov/view.xhtml?req=(title:49%20chapter:441%20edition:prelim)%20OR%20(granuleid:USC-prelim-title49-chapter441)&f=treesort&num=0&edition=prelim

United States Department of Transportation. (2023, March 16). *Full Year 2022 U.S. Airline Traffic Data | Bureau of Transportation Statistics*. Www.bts.gov; Bureau of

Transportation Statistics. https://www.bts.gov/newsroom/full-year-2022-us-airline-traffic-data#:~:text=U.S.%20airlines%20carried%20194%20million

Verma, T., Araújo, N. A. M., & Herrmann, H. J. (2014). Revealing the structure of the world airline network. *Scientific Reports*, *4*(1). https://doi.org/10.1038/srep05638

Wang, J., Mo, H., Wang, F., & Jin, F. (2011). Exploring the network structure and nodal centrality of China's air transport network: A complex network approach. *Journal of Transport Geography*, *19*(4), 712–721. https://doi.org/10.1016/j.jtrangeo.2010.08.012

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of "small-world" networks. *Nature*, *393*(6684), 440–442. https://doi.org/10.1038/30918

Xu, Z., & Harriss, R. (2008). Exploring the structure of the U.S. intercity passenger air transportation network: a weighted complex network approach. *GeoJournal*, *73*(2), 87–102. https://doi.org/10.1007/s10708-008-9173-5

Yang, H., Le, M., & Wang, D. (2021). Airline Network Structure: Motifs and Airports' Role in Cliques. *Sustainability*, *13*(17), 9573. https://doi.org/10.3390/su13179573

Zhang, L., Zhao, Y., Chen, D., & Zhang, X. (2021). Analysis of Network Robustness in Weighted and Unweighted Approaches: A Case Study of the Air Transport Network in the Belt and Road Region. *Journal of Advanced Transportation*, *2021*, 1–13. https://doi.org/10.1155/2021/8810254

VIII. ANNEXES

Louvain Community Algorithm (Bondel et al 2008):

It works in a two step process: first, assigns every node to a single community, so there are as many communities as there are nodes, and then evaluates what would happen to the neighbors' community if it the node were to be added, if there is a gain in modularity by removing the node from its community and adding it to another, then it is placed in the new community in which the gain was the maximum. The process is repeated until there are no further improvements. The second; constructs a new network whose communities are those found in the first phase and assigns the weight of the edges as the the sum of the weight of the connections between the nodes of the corresponding clusters.

Modularity measures the density of the connections within clusters versus the density of the connections between clusters..

Tables

I. Data Snapshot

DEPARTURES_PERFORMED	SEATS	PAX	DISTANCE	RAMP_TO_RAMP	UNIQUE_CARRIER	CARRIER_NAME	CARRIER_GROUP	CARRIER_GROUP_NEW	ORIGIN_CITY_MARKET_ID	ORIGIN	ORIGIN_WAC	DEST	DEST_COUNTRY	DEST_WAC	AIRCRAFT_GROUP	AIRCRAFT_TYPE	AIRCRAFT_CONFIG	YEAR	MONTH	DISTANCE_GROUP	CLASS	DATA_SOURCE
1	139	26	1655	217	AA	American Airlines Inc.	3	3	32575	LAX	91	ACA	MX	148	6	655	1	1997	1	4	L	IU
1	139	78	1861	313	AA	American Airlines Inc.	3	3	30977	ORD	41	ZIH	MX	148	6	655	1	1997	1	4	L	IU
1	188	124	1179	183	AA	American Airlines Inc.	3	3	30194	DFW	74	HUX	MX	148	6	622	1	1997	1	3	L	IU
1	263	261	2204	305	AA	American Airlines Inc.	3	3	31703	JFK	22	POS	TT	280	6	691	1	1997	1	5	L	IU
3	417	238	1872	819	AA	American Airlines Inc.	3	3	30977	ORD	41	HUX	MX	148	6	655	1	1997	1	4	L	IU
3	450	218	1179	518	AA	American Airlines Inc.	3	3	30194	DFW	74	HUX	MX	148	7	715	1	1997	1	3	L	IU
4	556	419	1537	845	AA	American Airlines Inc.	3	3	32575	LAX	91	ZIH	MX	148	6	655	1	1997	1	4	L	IU
4	556	537	1124	700	AA	American Airlines Inc.	3	3	30194	DFW	74	ACA	MX	148	6	655	1	1997	1	3	L	IU
1	97	21	17	68	AA	American Airlines Inc.	3	3	31703	EWR	21	LGA	US	22	6	603	1	1997	1	1	F	DU

In this snippet of the year January 1997, American Airlines has data for each city pair where they performed or scheduled a flight.. For example, for the route LAX - ACA contained 139 seats, 1 departure performed, carrying 139 passengers. Similarly, for the route ORD- ZIH, performed 1 departure with 139 seats and transferred 78 passengers.

II. Tables II. Top 10 city pairs for different parameters.

	ORIGIN	DESTINATION	PAX
1	Los Angeles, CA	Chicago, IL	9,706,076
2	Chicago, IL	Los Angeles, CA	9,524,572
3	New York, NY	Chicago, IL	9,258,890
4	Chicago, IL	New York, NY	9,176,461
5	Chicago, IL	Denver, CO	8,629,687
6	Denver, CO	Chicago, IL	8,525,974
7	Los Angeles, CA	New York, NY	8,051,676
8	New York, NY	Los Angeles, CA	7,988,678
9	Las Vegas, NV	Los Angeles, CA	7,948,265
10	Los Angeles, CA	Las Vegas, NV	7,898,195

	ORIGIN	DESTINATION	DEP
1	New York, NY	Chicago, IL	82,585
2	Chicago, IL	New York, NY	81,587
3	Boston, MA	New York, NY	79,604
4	New York, NY	Boston, MA	79,442
5	San Francisco, CA	Los Angeles, CA	79,137
6	Los Angeles, CA	Las Vegas, NV	78,460
7	Los Angeles, CA	San Francisco, CA	78,428
8	Las Vegas, NV	Los Angeles, CA	77,842
9	Chicago, IL	Minneapolis, MN	77,155
10	Minneapolis, MN	Chicago, IL	76,181

	ORIGIN	DESTINATION	SEATS
1	New York, NY	Chicago, IL	11,357,264
2	Los Angeles, CA	Chicago, IL	11,329,762
3	Chicago, IL	Los Angeles, CA	11,292,444
4	Chicago, IL	New York, NY	11,291,215
5	Los Angeles, CA	Las Vegas, NV	10,774,833
6	Las Vegas, NV	Los Angeles, CA	10,645,314
7	San Francisco, CA	Los Angeles, CA	10,606,746
8	Los Angeles, CA	San Francisco, CA	10,539,327
9	Washington, DC	Chicago, IL	10,268,606
10	Chicago, IL	Denver, CO	10,100,440

	ORIGIN	DESTINATION	PPF
1	Honolulu, HI	Los Angeles, CA	172,517
2	Los Angeles, CA	Honolulu, HI	170,133
3	San Francisco, CA	Chicago, IL	164,314
4	Los Angeles, CA	Chicago, IL	162,536
5	Chicago, IL	San Francisco, CA	160,567
6	Chicago, IL	Denver, CO	160,118
7	Denver, CO	Chicago, IL	158,760
8	New York, NY	Atlanta, GA	157,823
9	Minneapolis, MN	Chicago, IL	155,615
10	Atlanta, GA	New York, NY	154,768

	ORIGIN	DESTINATION	DISTANCE
1	Honolulu, HI	Los Angeles, CA	2,152,152
2	Los Angeles, CA	Honolulu, HI	2,147,040
3	San Francisco, CA	Chicago, IL	1,876,445
4	Chicago, IL	San Francisco, CA	1,861,551
5	Los Angeles, CA	Chicago, IL	1,837,854
6	Los Angeles, CA	Washington, DC	1,803,876
7	Chicago, IL	Los Angeles, CA	1,746,928
8	Washington, DC	Los Angeles, CA	1,668,769
9	Seattle, WA	Chicago, IL	1,590,225
10	Newark, NJ	Los Angeles, CA	1,546,020

	ORIGIN	DESTINATION	SPF
1	Minneapolis, MN	Chicago, IL	211,101
2	Chicago, IL	Washington, DC	206,062
3	San Francisco, CA	Chicago, IL	203,226
4	New York, NY	Atlanta, GA	202,599
5	Chicago, IL	San Francisco, CA	201,047
6	Los Angeles, CA	Honolulu, HI	200,916
7	Honolulu, HI	Los Angeles, CA	200,533
8	Chicago, IL	Minneapolis, MN	198,901
9	Detroit, MI	Chicago, IL	198,886
10	Washington, DC	Chicago, IL	198,865

	ORIGIN	DESTINATION	OR
1	San Juan, PR	Fort Dix, NJ	100%
2	South Bend, IN	Sioux Falls, SD	100%
3	Corpus Christi, TX	Beaumont/Port Arthur, TX	100%
4	Seattle, WA	Bloomington, IN	100%
5	Knoxville, TN	Columbus, GA	100%
6	King Salmon, AK	Ketchikan, AK	100%
7	Atlantic City, NJ	Richmond, VA	100%
8	Killeen, TX	Raleigh/Durham, NC	100%
9	Seattle, WA	Killeen, TX	100%
10	Ketchikan, AK	Portland, OR	100%

III. Top 50 most connected cities

City	Connections	City	Connections

1	Chicago, IL	286	26	Colorado Springs, CO	169
2	Denver, CO	250	27	Las Vegas, NV	166
3	Dallas/Fort Worth, TX	238	28	San Diego, CA	166
4	Atlanta, GA	231	29	Oklahoma City, OK	165
5	Houston, TX	219	30	Austin, TX	163
6	Los Angeles, CA	211	31	Philadelphia, PA	162
7	Minneapolis, MN	207	32	Pittsburgh, PA	162
8	Washington, DC	202	33	Indianapolis, IN	158
9	New York, NY	197	34	Cleveland, OH	157
10	Detroit, MI	195	35	Tampa, FL	155
11	Seattle, WA	190	36	Raleigh/Durham, NC	153
12	San Francisco, CA	186	37	Memphis, TN	152
13	Salt Lake City, UT	185	38	New Orleans, LA	151
14	Phoenix, AZ	183	39	Miami, FL	147
15	Newark, NJ	182	40	Fort Lauderdale, FL	146
16	Charlotte, NC	174	41	Portland, OR	146
17	Milwaukee, WI	173	42	Birmingham, AL	145
18	St. Louis, MO	173	43	San Antonio, TX	145
19	Boston, MA	171	44	Knoxville, TN	143
20	Orlando, FL	171	45	Columbus, OH	140
21	Baltimore, MD	170	46	Madison, WI	136
22	Cincinnati, OH	170	47	Dallas, TX	135
23	Kansas City, MO	170	48	Tulsa, OK	135
24	Nashville, TN	169	49	San Jose, CA	133
25	Colorado Springs, CO	169	50	Grand Rapids, MI	131

IV. Airport Classification per node and passengers

Characteristics per node classification (count)

Regional	Midwest	South	Alaska	East	West	Pacific	Totals	%
Non hub connector nodes	5	15	3	25	23	0	71	16%
Peripheral nodes	13	40	12	98	79	2	244	55%
Ultra peripheral nodes	5	25	8	43	38	1	120	27%
Connector hubs	1	2	1	2	6	0	12	3%
Totals	24	82	24	168	146	3	447	
%	5%	18%	5%	38%	33%	1%		

% per region	Midwest	South	Alaska	East	West	Pacific
Non hub connector	21%	18%	13%	15%	16%	0%
Peripheral nodes	54%	49%	50%	58%	54%	67%
Ultra peripheral	21%	30%	33%	26%	26%	33%
Connector hubs	4%	2%	4%	1%	4%	0%

% per type	Midwest	South	Alaska	East	West	Pacific
Non hub connector	7%	21%	4%	35%	32%	0%
Peripheral nodes	5%	16%	5%	40%	32%	1%
Ultra peripheral	4%	21%	7%	36%	32%	1%
Connector hubs	8%	17%	8%	17%	50%	0%

Characteristics per node classification (Passenger Flow)

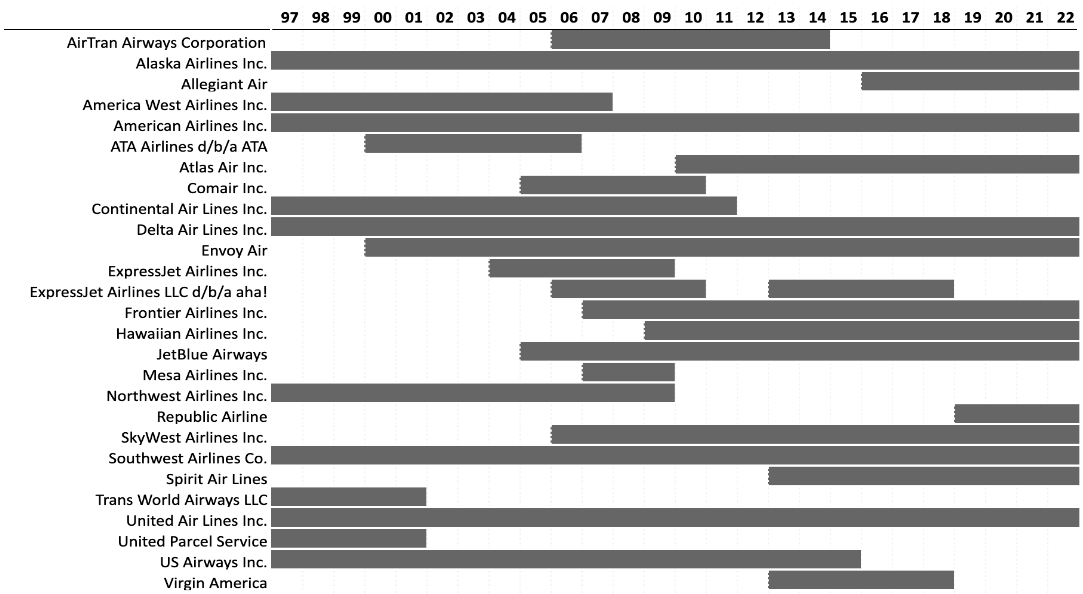
Regional	Midwest	South	Alaska	East	West	Pacific	Total	%
Non hub connector	219,139	46,462,891	408,301	130,215,191	153,793,136	-	331,098,658	9%
Peripheral nodes	85,844,791	64,889,604	4,386,972	660,512,822	316,298,524	60,315	1,131,993,028	32%
Ultra peripheral nodes	24,551	105,349,823	7,654,049	451,290,869	93,227,287	344,831	657,891,410	18%
Connector hubs	227,916,219	273,054,426	17,353,399	289,019,400	664,848,616	-	1,472,192,060	41%
Total Pax	314,004,700	489,756,744	29,802,721	1,531,038,282	1,228,167,563	405,146	3,593,175,156	
% of Pax per region	9%	14%	1%	43%	34%	0%		

% per region	Midwest	South	Alaska	East	West	Pacific
Non hub connector	0%	9%	1%	9%	13%	0%
Peripheral nodes	27%	13%	15%	43%	26%	15%
Ultra Peripheral	0%	22%	26%	29%	8%	85%
Connector hubs	73%	56%	58%	19%	54%	0%

% per type	Midwest	South	Alaska	East	West	Pacific
Non hub connector	0%	14%	0%	39%	46%	0%
Peripheral nodes	8%	6%	0%	58%	28%	0%
Ultra peripheral	0%	16%	1%	69%	14%	0%
Connector hubs	15%	19%	1%	20%	45%	0%

Figures

I. Timeline of the existence of the 27 Major Carriers



II. Visualization of all the nodes and edges of the data

III. Top airline Route Maps