Case study on

Delay Network in Airport

Graph Theory Lab Innovative Work

Shubham Kumar (2K18/MC/111), Vaibhav Jain (2K18/MC/120)

Delhi Technological University

This project was compiled on October 24, 2021

In this work, we study a publicly available airport delay network dataset. By applying different network measures, we observe the importance of the centrality to airport delays by finding a positive correlation between these two measures. We also use correlation coefficients as metrics to compare the correlation between flight delay and centrality score.As observing certain cities having high centrality score and delays, we found news report to verify our observations.

1. Introduction

We have seen a surge of airplane development in recent decades, and according to International Air Transport Association (IATA), by 2035, there will be 7.2 billion air travelers, almost doubling the current 3.8 billion passengers.

With more passengers, airplanes and airports, there are growing problems posed for the proper functioning of air transportation, especially air traffic. Air transportation has special properties that are not shared by road transportations that have been extensively studied. In particular, aircrafts have limited carrying capacity and route stops cannot be flexibly created. The latter is a major concern for solving congestion problem in air transportation, as the airports require great investments and have special requirement that makes temporary airports almost impossible. In such context, route design becomes crucial in solving air transportation congestion problem. To enable better route design that minimizes air traffic, systematic network studies of airport network system becomes important.

Air transportation systems have been previously described as graphs with vertices representing airports and edges direct flights. In this paper, we attempt to apply the same modelling and study air traffic and network pattern using the data of airport transportation from 1987 to 2008 in United States. The air transportation system in United States developed early and has become relatively stable in recent years, which makes it a suitable study project to observe the patterns of the system. We modeled airplane system as a network with airports being the nodes and air routes between the airports as edges. We are interested in identifying vulnerable airports and how the topological structure of the network influence the traffic.

For simplicity of the analysis, we modeled the airport network as simple undirected network and analyze the centrality properties of it. Centrality is intuitively related to the delay pattern. For a node that is more connected with other nodes, the delay that happens at such node matters more to the overall delay situation. We also run the network against clustering coefficient, which measure the transitivity of networks. Clustering coefficient is intuitively important as well because for a perfect transitive network, the delay that happens at one place would be transmitted to other parts of the network. On the other hand, for a not so transitive network, the delay pattern turns to stay more local.

In this work, we conducted different network statistics, such as component, clustering coefficient, cliques, degree centrality, closeness centrality, betweenness centrality. We observe the importance of the centrality to airport delays by finding a positive correlation between these two measures. We then use Pearson correlation coefficient and Spearman correlation coefficient to compare the performance of different centrality measures. As observing certain cities having high centrality score and delays, we found news report to verify our results.

Significance Statement

We have been witnessing flight delays ever since the first day the first airport was constructed. Survey shows delay is one of the most important aspects that influence flight quality of experience. Our studies on network can help flight managers to better design flight routes and avoid long flight delays.

2. Dataset

A. Dataset Description. For this project, we utilize the dataset provided by Statistical Computing Statistical Graphics (http://stat-computing.org/dataexpo/2009/the-data.html). The transportation data is arranged into individual tables in CSV form according to year. As a result, we download 17 tables from the year 1987 to 2008. Each table contains 29 variables, such as year, month, day of month, origin, destination, distance, cancellation status. In each table, every row indicates a flight route from one original city to another destination city in one month. Also, the number of rows increases gradually from around 1,310,000 in 1987 to 7,000,000 in 2008.

B. Dataset Preprocessing. Since much of the data is incomplete and has null value fields that are essential for our organization, thus we choose to use the dataset to be year after 2000. This provides a reasonably sized dataset(around 7 GB). We developed scripts to parse the original dataset, the airport codes and also airport coordinates. Since the original dataset does not have the coordinates for every city and airport code, we wrote scripts to look up the coordinates of the airport. Additionally, each flight's departure time is only stored as a timestamp in the local time of the departure city. This causes issues when flights cross time-zone borders. This is common to our daily experience. When we take a flight from west to east, our departure time and arrival time is all for local time of the airport. To solve this problem, we write a script that collects time-zone information for all airports, and converts flight times into a UTC timestamp when given the year, month, day, time, and airport code.

Then we build our network based on a Python library **NetworkX**. We model the US airport network as an undirected, unweighted network. Moreover, we denote the cities which airports locate as the nodes in the network, the routes of the flight between any individual pairs of two cities as the edges, and the total number of trips between airports within a year as the weights of the network.

Since the inherent limitation of coordination between the library that we use and the dataset, instead of making all edges to be weighted, we choose keep a separate data structure to store the delay time, which can be accessed based on the airport. However, we can still make two copies of data structure, arrival delay and departure delay respectively.

C. Basic Exploration. As our initial step, we look at the relationship to delay time.

In order to understand the network by finding out the detailed information about the network. We explore the dataset by plotting the basic statistics about the network constructed in the previous section. The statistics can be found in Table 1. The node is the airports, which we map with coordinates preprocessed in previous section, and the edges are the airlines that are found in the dataset. We can see there are some nodes that have very high centrality, as there are many lines going to or from the airport, and some of the states are highly clustered. The plot with coordinates can be found in Figure 10.

3. Measurements

Without loss of generality, before presenting our computations and results, several measurements and mathematical definitions are introduced to study the US airport network in the following sections.

A. Degree Centrality. Degree centrality k_i of node i measures how many neighbors of the current node has. In our study, k_i measures how many airports connect to the current airport i by undirected edges.

B. Shortest Distance and closeness Centrality. A shortest distance d_{ij} between node i and node j in a network is the number of edges along the shortest path between node i and node j. The mean shortest distance l_i from a node i to every node in the network is define as $l_i = \frac{1}{n}\sum_j d_{ij}$. From the aspect of the definition of shortest distance, we can define closeness centrality.

Closeness centrality, however, is used to measure the mean distance from a node to other nodes in a network. Thus closeness

centrality C_i of a node i is defined as

$$C_i = \frac{1}{l_i} = \frac{n}{\sum_j d_{ij}} \quad (5)$$

Measure	Statistics
Components	1
Number of Nodes	261
Number of Edges	2241
Average Degree	17.17
Total Triangles	46326
Average Clustering Coefficient	0.673
Number of Cliques	6645

Fig. 1. Basic Statistics about the Network

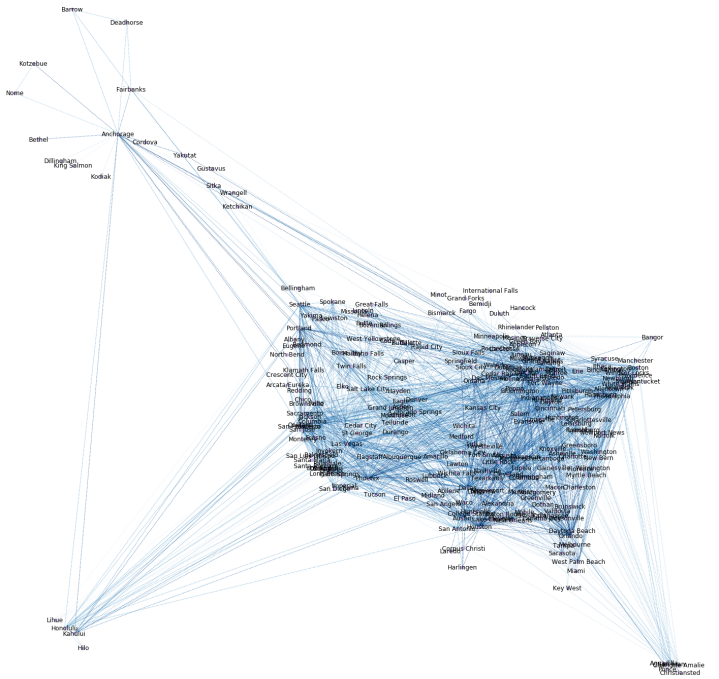

Fig. 2. Map for Airports in Dataset

In this project, the closeness centrality of an airport in the US airport network can be used to measure how efficient an airport is.

C. Betweenness Centrality. Betweenness centrality measures the number of shortest paths between any given pair of nodes in the network that passes through the current node. We define betweenness centrality of node i in a network

$$B_i = \sum_{st} \frac{n_{st}^i}{g_{st}}$$

where n_{st}^i is the number of shortest path from node s to node t that pass through node, and g_{st} is the total number of shortest paths from node s to node t (5). Moreover, if both n_{st}^i and g_{st} are zero, then $\frac{n_{st}^i}{g_{st}} = 0$.

In the project, the betweenness centrality of an airport measures how important an airport is in terms of incoming and outgoing domestic flights. Besides, combing closeness centrality and betweenness centrality together, we can measure the efficiencies of the airports in the US airport network; in other words, both centrality measurements rank which airport is a better transit center in terms of traveling distance between cities.

D. Vulnerability Index. Invulnerability index measures how susceptible an airport in terms of the importance of the airport. In the study, we define vulnerability index V_i as the ratio of bad performance of an airport i in terms of domestic flights to measure how important an airport is in the US airport network. An airport with a more significant centrality score is more important. Furthermore, an airport with longer delay time might have a smaller value of invulnerability index if it only maintains a few flight routes, while an airport with shorter delay time might still maintain a large value of invulnerability index if it has more flight routes than other airports have.

Mathematically, the vulnerability index of an airport i is

$$V_i = \frac{D_i}{C_i,}$$

where D_i is the average delay of the airport and C_i is the centrality of airport i. We will use the strongest correlated airport with delay as the centrality score in equation D.

E. Correlation Coefficient. We use both the Perarson correlation coefficient and the Spearman's order coefficient in the project. Pearson correlation coefficient ρ measures the linear dependence of two variables.Pearson correlation for any given paired data $(x_1, y_1), ..., (x_n, y_n)$, the Pearson correlation coefficient is defined as

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where n is the size of the sample, x_i and y_i are the individual data points, and $\bar{x}$ and $\bar{y}$ are the mean values of the data points x_i and y_i $(i = 1, ..., n)$ particularly. Furthermore, the Cauchy-Schwarz ineuqality states

$$| < \vec{u}, \vec{v} > |^2 \leq (\vec{u}, \vec{u}) \cdot (\vec{v}, \vec{v})$$

where $\vec{u}$ and $\vec{v}$ are vectors; as a consequence, the Pearson correlation coefficient ρ between two variables must be between +1 and -1. Moreover, +1 is defined as the total positive linear correlation; 0 means no linear correlation between the two variables; -1 is defined as the total negative correlation.

Spearman's order coefficient r_s measures the dependence of the rankings of two variables. r_s is also considered as the Pearson correlation coefficient between the rank variables. Moreover, we define Spearman's order coefficient as

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where d_i is the difference between the ranks of two corresponding variables for each observation, and n is the number of observations. The difference between Spearman's order coefficient and Pearson correlation coefficient is that Spearman's order coefficient assesses the monotonic relationship between two variables, while the latter determines linear relationships between two corresponding variables.

In this paper, we compute the Pearson correlation coefficients and the Spearman's order coefficient between each centrality measurement and delay time for every airport to study the dependent relationships between centrality scores and delay time.

F. Clustering Coefficient. Clustering coefficient C measure the degree of transitivity of a given network. We measured the clustering coefficient C of the whole network by the formula

$$C = \frac{\text{(number of closed paths of length two)}}{\text{(number of paths of length two)}}$$

where C=1 indicates the perfect centrality and C=0 indicates no centrality. (6)

4. Results

Based on the measurements introduced in the previous section, we present our results in this section. We first show the different correlations between centrality scores and delay time. We second offer the computational results of the vulnerability index.

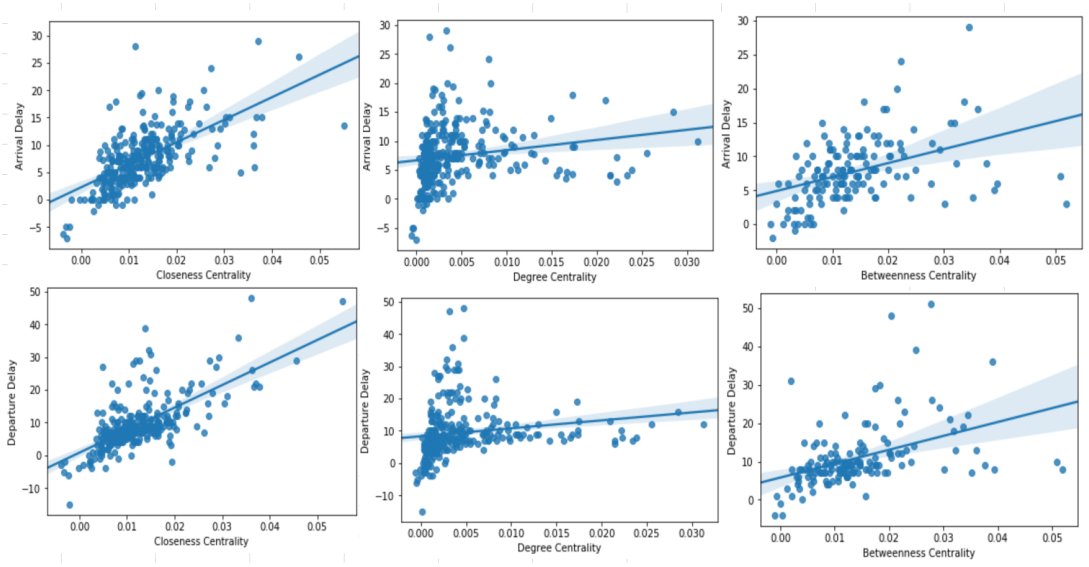

Fig. 3. Different Centrality Measurements vs. Delay Time

A. Correlation Between Centrality Scores and Delay Time. We compute different centrality scores of the US airport network in the period of 1987-2008. Based on our computational results, the betweenness centrality scores and the closeness centrality scores have good correlations with the delay time, while the degree centrality scores have poor dependence relationships with the delay time. As shown in Fig.3, we can see linear dependent relationships between both the betweenness centrality scores and the closeness centrality scores and delay time for the arrival delay case and the departure delay case.

For the arrival delay case, the value of the Pearson correlation coefficient between the delay time and the degree centrality scores is around 0.179, which is not strongly correlated with the delay time, while the values of the Pearson correlation coefficient

Table 1. Airport Ranking Based on Different Centrality Measurements

Rank	Degree Centrality	Betweenness Centrality	Closeness Centrality
1	Atlanta	Atlanta	Atlanta
2	Chicago	Salt Lake City	Chicago
3	Dallas	Dallas	Dallas
4	Denver	Minneapolis	Denver
5	Minneapolis	Anchorage	Minneapolis
6	Detroit	Chicago	Salt Lake City
7	Salt Lake City	Denver	Detroit
8	Houston	Houston	Houston
9	Cincinnati	Detroit	Cincinnati
10	New York	San Francisco	Las Vegas

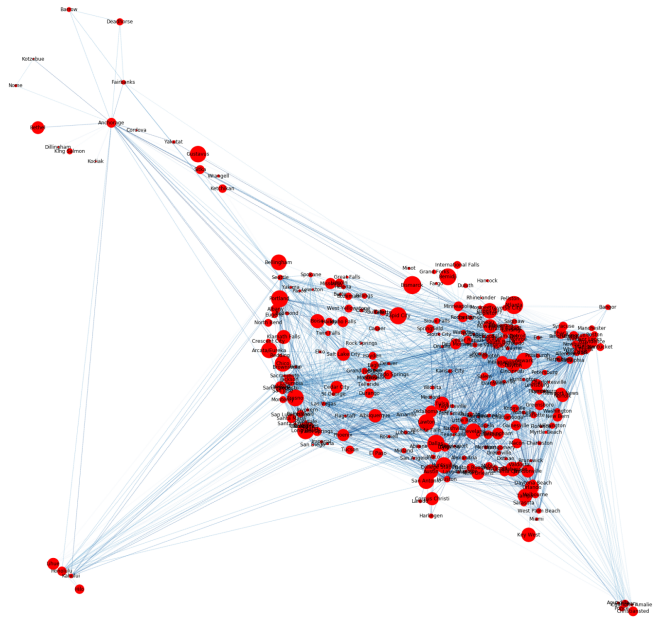

Fig. 4. Visualization of the Vulenrability index for each Airport

Fig. 5. Top Ten Vulnerable Airports for Delay

Cities	Vulerablility Index
1. Atlanta	0.160344
2. Salt Lake City	0.128662
3. Minneapolis	0.127824
4. Dallas	0.126783
5. Detroit	0.101268
6. Denver	0.076429
7. Chicago	0.062064
8. Houston	0.059642
9. Anchorage	0.057610
10. Phoenix	0.055963

for the betweenness centrality scores and the closeness centrality scores are around 0.414 and 0.634 particularly. Moreover, the computational results for the Spearman's order coefficients strongly support the results of the Pearson correlation coefficients as the betweenness centrality scores have around 0.489 correlation value with the arrival delay time, and the closeness centrality scores have approximately 0.644 correlation value with the arrival delay time. Similarly, for the departure delay case, the value of the Pearson correlation coefficient between the betweenness centrality scores and the departure delay time is around 0.536, while the value of the Pearson correlation coefficient between the closeness centrality scores and the departure delay time is about 0.634. The results discussed previously suggest that betweenness centrality measurement and closeness centrality measurement are proper tools to analyze the delay time of airports.

B. Vulnerability Index. We calculate the vulnerability index for each airport in the US airport network. As shown in Table. 5, we can find the top ten vulnerable airports for the delay. Furthermore, from Table.1 and Table.5, we notice that the airports with high centrality scores tend to have large vulnerability index values. As shown in the tables, Atlanta, Salt Lake City, Minneapolis, Dallas, Chicago, Denver, Detroit, and Huston maintain not only high values of the vulnerability index but also large values of each centrality measurement.

5. Analysis

A. Altanta. Surprisingly, we find that the airport that has the highest vulnerability and several centrality values is Atlanta International Airport, rather than the airports that belong to the most popular and well-known cities. We want to explore the reason behind this surprise.

One hypothesis that we propose is that due to the high ranking of cities like Atlanta such as Chicago and Detroit, the reasons that these city have high ranking is due to the heavy industry history that they belong to and old airplane paths that have developed since then. Also, it coud also be currently the goods that are transported through airplane outweighs humans and so the industrial cities have high ranking due to their importance in creating goods.

Indeed, we found from (7) that the Hartsfield-Jackson International Airport at Atlanta is the connecting hub for United States, and also home to Delta Air Lines which was initially created there. However, as the source suggests, there indeed are large scale of passengers that are connected rather than the goods as hypothesized earlier.

B. Degree centrality. We find that degree centrality is not a strong indicator of delay time, as they do not have strong linear dependence with each other. This result seems surprising at first, but makes sense after more consideration. Degree centrality can indicates that the airport either have many routes to one or a set of few other airports, or have routes to many other airports.

On the other hand, the greater the vulnerability index, the greater the betweeness centrality. Betweeness centrality gives the importance of the airport (node) in the network in terms of its connection to different nodes. If the airport is central to the connection of many other airports, then its delay would affect them greatly. we think these two results convey something realistic about the nature of air transportation and delay, just as noted in [the international network report], degree centrality and betweeness centrality diverge in general.

6. Discussion and Future Work

In this work, we explore flight delay network dataset. By applying different network measures, we observe the importance of the centrality to airport delays by finding a positive correlation between these two measures. We also use correlation coefficients as metrics to compare the correlation between flight delay and centrality score.

We also have very interesting observations, such as the linear dependent relationships between the delay time and centrality measurements. Yet, our exploration has very great potential to inspire future works on this dataset. We list as following:

- For simplicity, we process our whole network as a simple network, but there are far more options than doing so. For example, in related works, they may understand how one single flight influence the other. Such temporal information can not be included in our simple network.
- Our vulnerability index is very simple by dividing delay with centrality score. Future works may come up with more sophisticated formulations for vulnerability index to better understand our results.
- In this work, we consider, with great portion, the influence of centrality to the whole network. However, there are far more network measures than centrality scores, which may also have correlations.

1. DP Cheung, MH Gunes, A complex network analysis of the united states air transportation. *2012 IEEE/ACM International Conference on Advances Social Networks Analysis Mining* (year?).
2. AB Aurelien Gautreau, MB elemy, Microdynamics in stationary complex networks. *Proceedings National Academy Sciences* **106**, 8847–8852 (2009).
3. P Fleurquin, JJ Ramasco, VM Eguiluz, Systemic delay propagation in the us airport network. *Scientific reports* **3**, 1159 (2013).
4. DBYW Cong Wei, Minghua Hu, F Cheng, Empirical analysis of airport network and critical airports. *Chinese Journal Aeronautics* **28**, 512–519 (2016).
5. M Newman, *Networks.* (Oxford University Press), (2018).
6. M Newman, *Networks.* (Oxford University Press), (2018).
7. H Tan, https://www.cnbc.com/2017/12/19/why-the-atlanta-airport-the-busiest-in-the-world.html (2017).

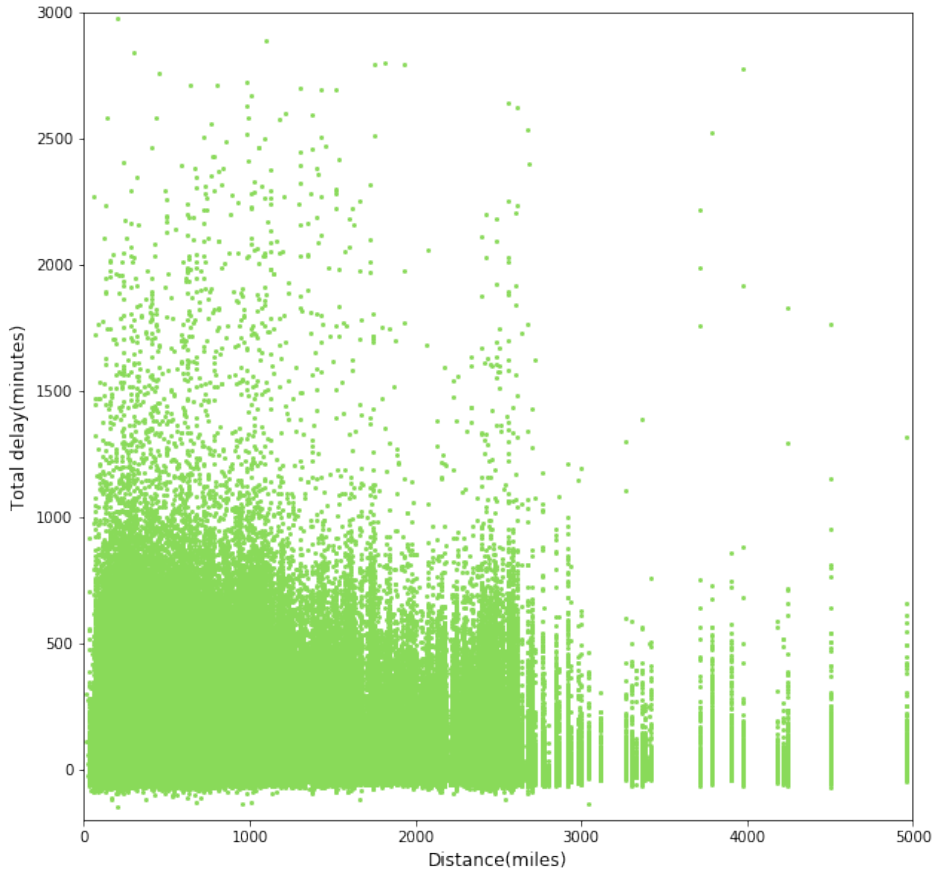

Fig. 6. Relation between distance of flight distance and delay.

A. Appendix: Supplementary Materials

We thereby include supplementary materials associated with with work. We use Python Jupyter Notebook to show our results and visualize our code. We establish all our results on https://colab.research.google.com/drive/1kgZTzQDd7JpRH46Lfkr0jYJ3OV8qBhVL.Reader could run through this notebook to reproduce our results.

We keep track on our progress by Github, an open source code repo. Our code is in https://github.com/ChangyuYan/Mathematical-Networks

B. Additional Figures and Information

A. Basic Exploration Figures. In order to understand the relationship between delay and travel distance of air route. We can see intuitively, most of the points are clustered to be low distance and low total delay time. This intuitively makes sense by our daily experience. We can observe, modelling directly through travel distance, or other metrics not-related to geolocation(such as location of the airport) will not generate meaningful result.

B. Larger Figures for US Maps. In this subsection, we include a larger figure than the paper itself to better show our results.

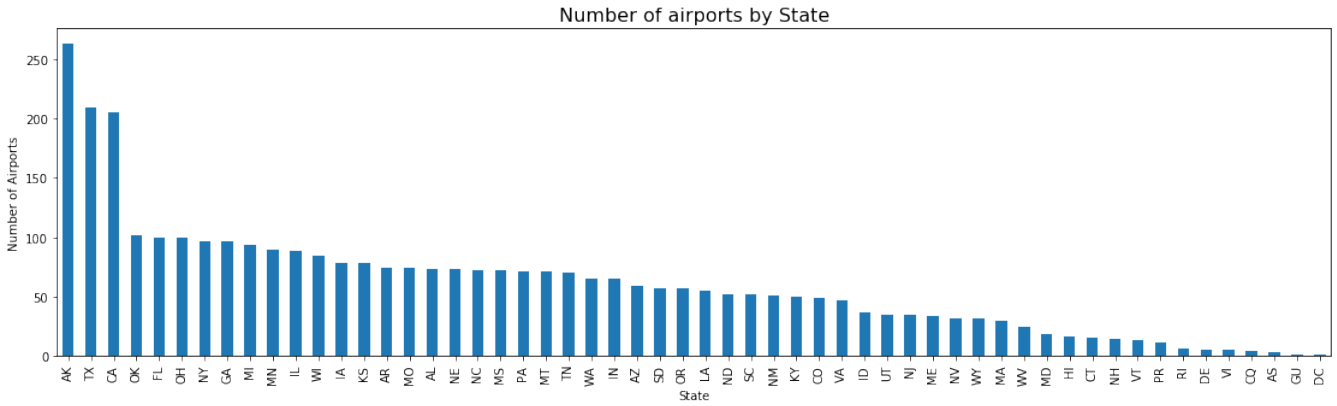

Fig. 7. State Histogram on Airport numbers

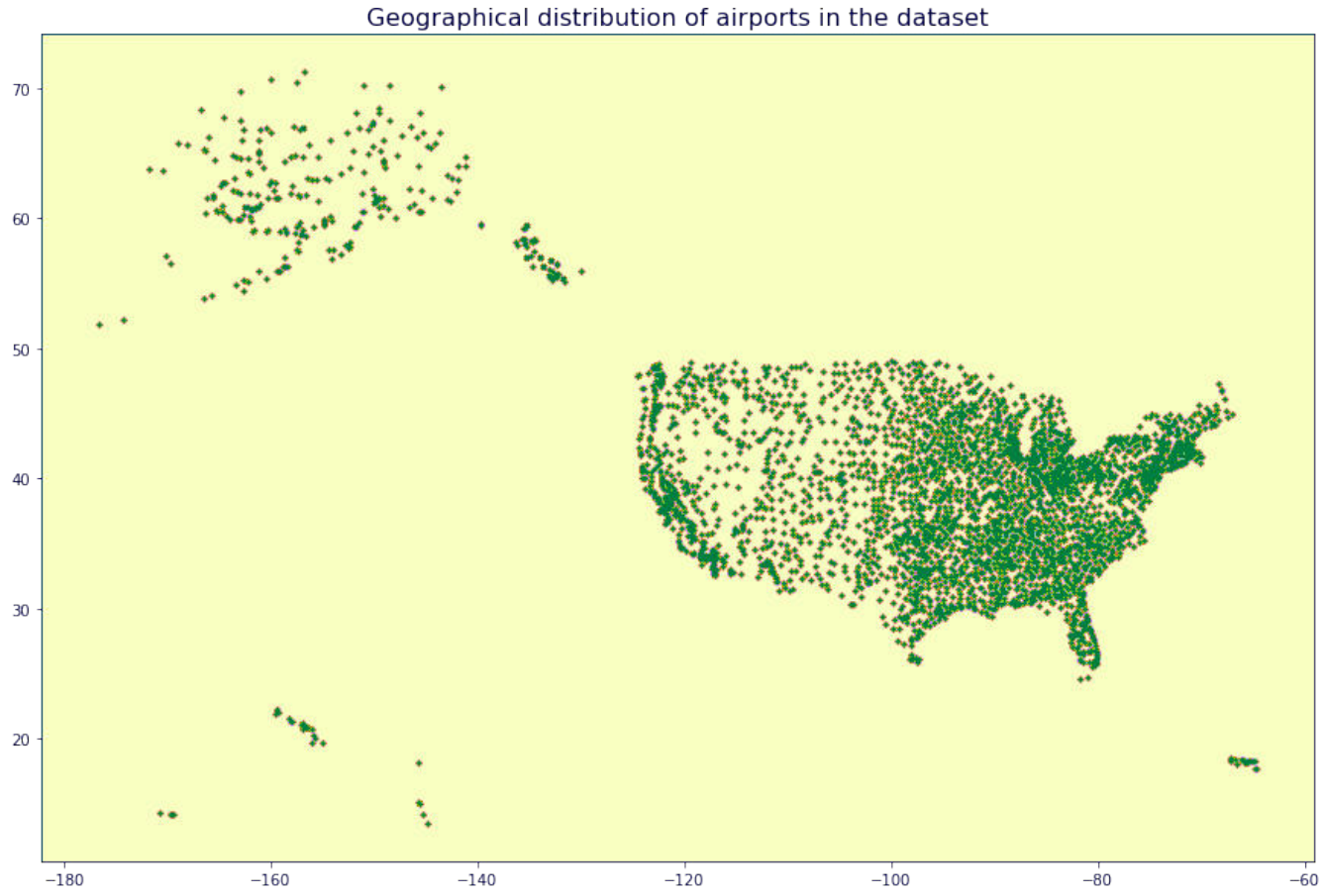

Fig. 8. Plot of airport Geolocations

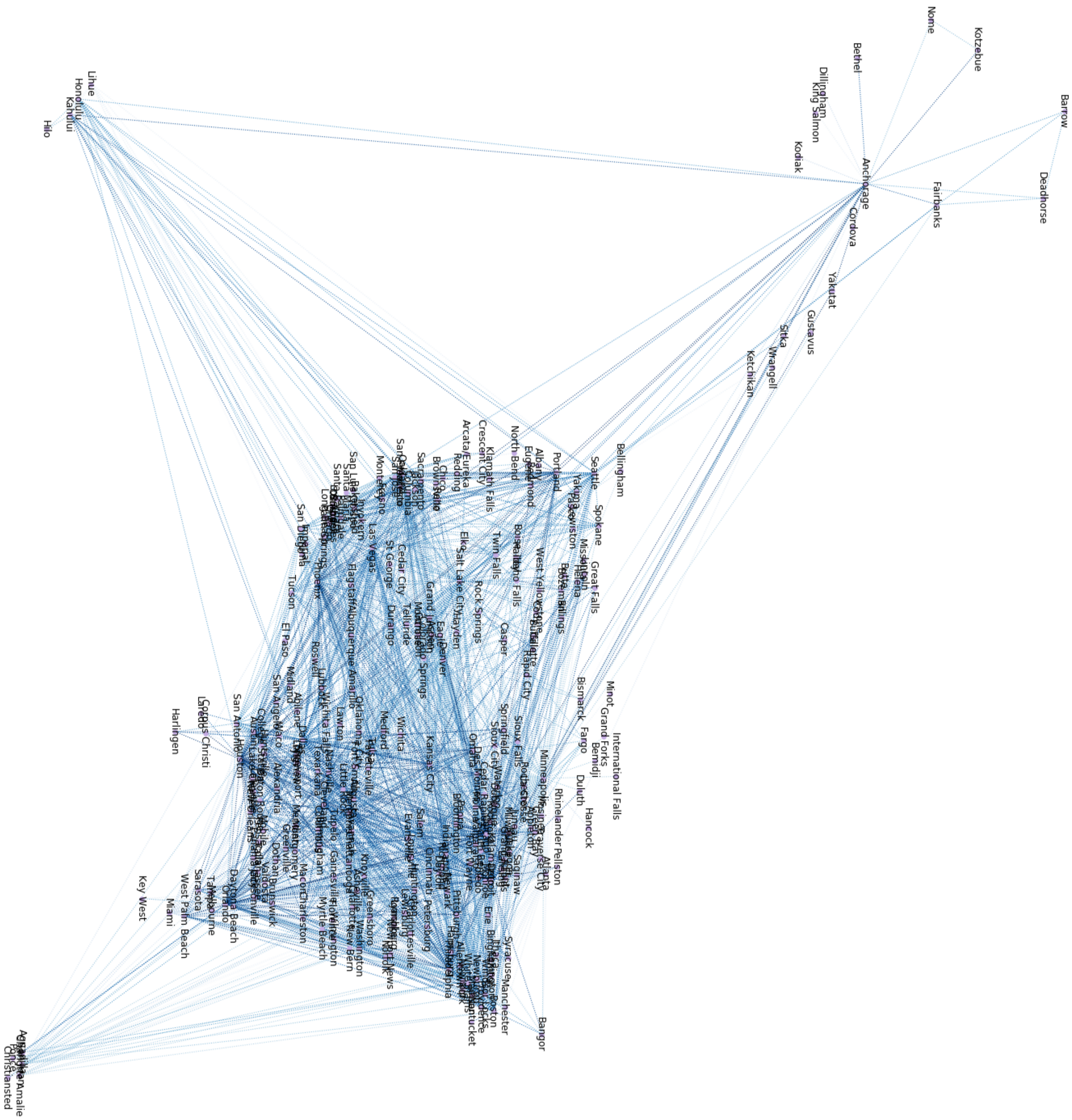

Fig. 9. Larger Figure for Figure 2

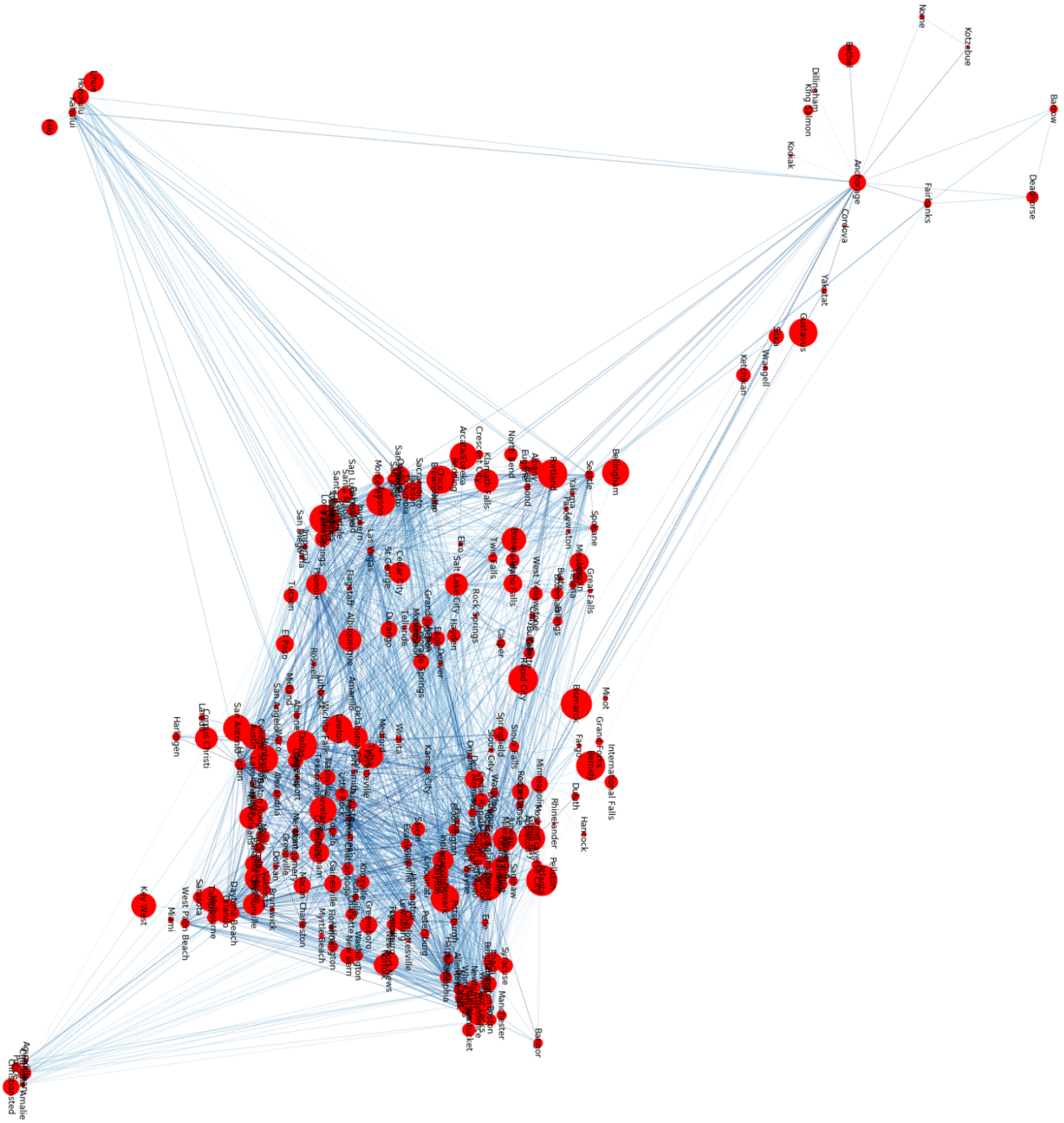

Fig. 10. Larger Figure for Figure 4