American Domestics Airline Analysis

Linxiao Bai, Yuanzheng Du, Keting Lyu,
University of Rochester
Rochester, NY
{lbai2, ydu20, klyu}@ur.rochester.edu

Abstract – Air transportation plays an important role in efforts to improve the efficiency of society. In this paper, we analyze the domestics' airline network in America to better understand its characteristics and distinctions. For this, first we visualize American major airports traffic grid, with respect to geometric location. We measure several complex network features including centralities, degree distribution, assortativity, clustering and resiliency. Then we build devised configuration model to fit the airport network as original distribution. We also simulate the random process to evaluate and build the decision tree to interpret the network deeply. Finally, we provide case analysis of airport JFK and ORD.

Keywords – airline transportation; complex networks; network analysis; airline model

I. INTRODUCTION

General aviation is particularly popular in North America, with over 5,200 airports available for public use. According to the U.S. Aircraft Owners and Pilots Association, general aviation provides more than one percent of the United States' GDP.

Although there are many papers provide aviation network analysis, for example,

- Yang and so on (2015) analyzed the corresponding distance-weighted network, showed the difference from the airline network in features like edge density and average shortest path.
- Dorothy, Mehmet (2012) focus on the evolution of complex network features, which including average shortest path, degree distribution, assortative mixing, clustering coefficient, betweenness centrality, over two decades. They put forward U.S air transportation network exhibits small world characteristics and has a special partial power law degree distribution.
- Guimer and Amaralb(2004) found the most connected cities of world-wide airport network are surprisingly not the most central cities. Their distance model has exponential dependence and power-law dependence, they also think politic plays a role of constraints
- Zengwang(2008), Robert apply and emerging methodology to study weighted network of passenger air transportation, besides small-world network, they also found a rich-club phenomenon.

The focus of these paper is to analyze the complex network features of the air transportation network in America. Including average shortest path, assortative mixing, and clustering coefficient; in addition, analyze how the network has evolved over the past years. And our paper will focus on not only multiple kinds of network features, but also network model to deeper understand air network in America.

II. METHODOLOGY

A. Data Set

The airline data is obtained from Bureau of Transportation Statistics, we use the dataset for the year of 2015. We use the attributes of original airport code and destination airport code, as well as distance group. Airport is uniquely identified as the airport code. Geometric location of airports comes from http://openflights.org/data.html. We use each airport's latitude and longitude data. Airports locations and flight details are merged together to form an integrated data for analysis purposes.

B. Network Generation

We generate our network via the Python Library igrpah. Also *echarts* JavaScript plotting library is used to generate complex graphs for Monte Carlo methods of simulating model. Data is pre-processed and cleaned to allow as igraph input. In the network graph, each node represents an airport, and a directed edge represents an available route from one airport to the other. The weight of an edge represents air-line distance between two airports.

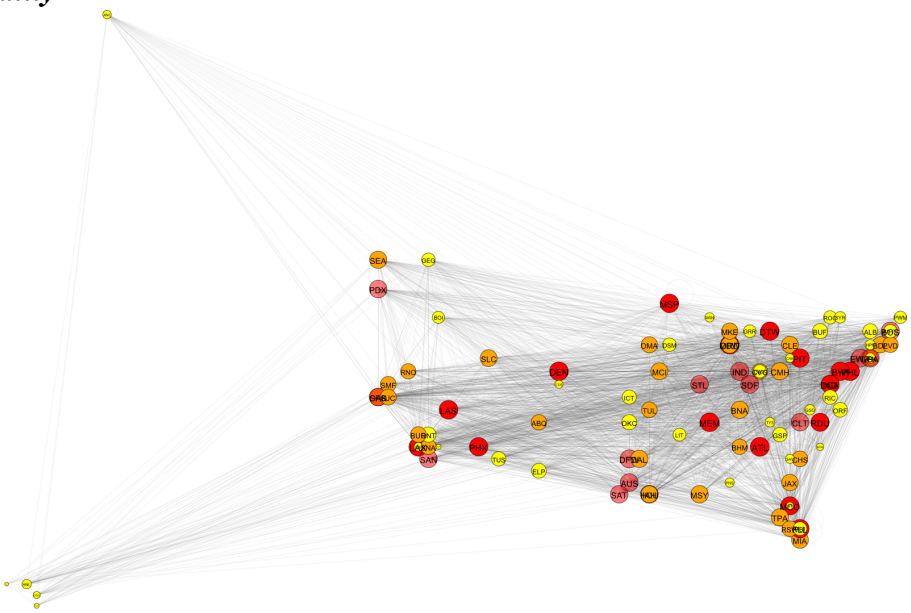
III. ANALYSIS

A. Centrality Analysis

Due to the capacity of showing dots on one graph, we choose only 100 airports in the U.S according to 'Top 100 Busiest Airports in USA'. We then have general ideas how these top 100 airports show their properties.

1. Degree Centrality

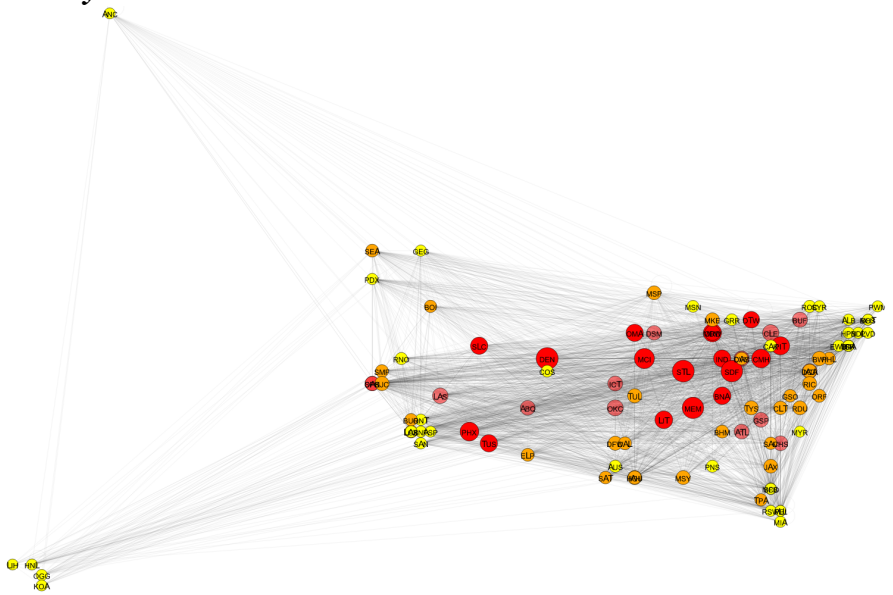
	Airport	Degree
1	ATL	96
2	ORD	95
3	LAX	95
4	MEM	95
5	DEN	95
6	LAS	95
7	PIT	94
8	BWI	94
9	MSP	94
10	PHL	94



High degree airports distribute in relative large and 'important' cities. We have LAX (L.A), ORD (Chicago), JFK (New York), PIT (Pennsylvania), DTW (Detroit) and MEM (Memphis). Most of them are international airports with huge handling capacity for both passengers and cargo. Generally, they are evenly distributed around U.S.

2. *Betweenness Centrality*

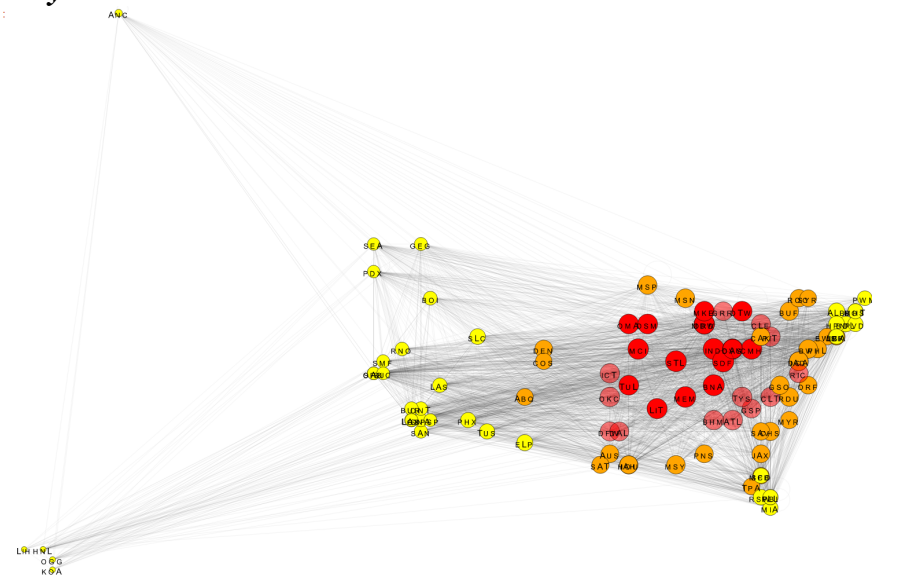
	Airport	Betweeness
1	DEN	183.831713
2	STL	181.396348
3	SDF	176.081077
4	MEM	173.707744
5	MCI	151.300723
6	PHX	129.289050
7	CMH	128.903053
8	PIT	121.038307
9	ORD	119.549887
10	LIT	111.784962



Airports with high betweenness centrality act like bridges. Red dots mostly distribute in the central, which means eastern and western airlines are likely connected by those bridge airports. Interesting at east coast, yellow dots in the middle act as bridges for northern and southern transportations. We also notice that high degree centrality and high betweeness centrality airports are highly overlapping.

3. *Closeness Centrality*

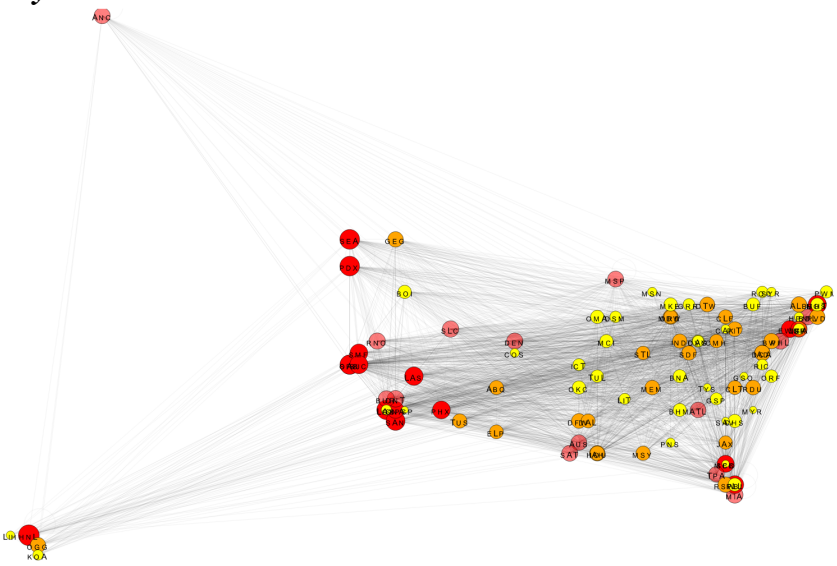
	Airport	Closeness
1	SDF	0.000815
2	IND	0.000805
3	BNA	0.000805
4	STL	0.000795
5	MEM	0.000792
6	MCI	0.000786
7	ORD	0.000783
8	CMH	0.000780
9	MDW	0.000773
10	CVG	0.000773



Airports with closeness centrality show a symmetrical distribution, and high closeness airports gather in center with centrality decreasing gradually as airports getting far away from it. We can assume that people travel from midland usually take shorter average trips to most destinations in U.S.

4. Eigenvector Centrality

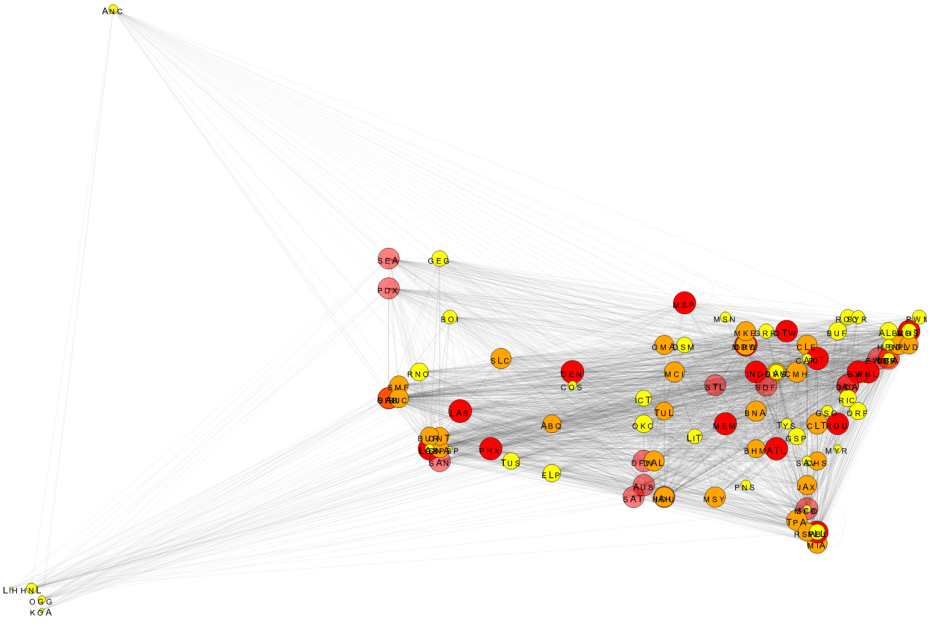
	Airport	Eigenvector
1	HNL	1.000000
2	SEA	0.942037
3	PDX	0.934811
4	OAK	0.921157
5	LAX	0.909372
6	SFO	0.904581
7	SAN	0.892408
8	LAS	0.861544
9	BOS	0.846608
10	SJC	0.839965



Eigenvector centrality shows high self-reinforce pattern. Red dots tend locate at map's corners, and be the center for 'clusters'. We can see surrounding airports decreasing in eigenvector centrality.

5. PageRank Centrality

	Airport	Pagerank
1	LAS	0.012616
2	DEN	0.012603
3	LAX	0.012468
4	ATL	0.012444
5	ORD	0.012422
6	PIT	0.012363
7	PHX	0.012344
8	MEM	0.012302
9	MSP	0.012272
10	PHL	0.012186

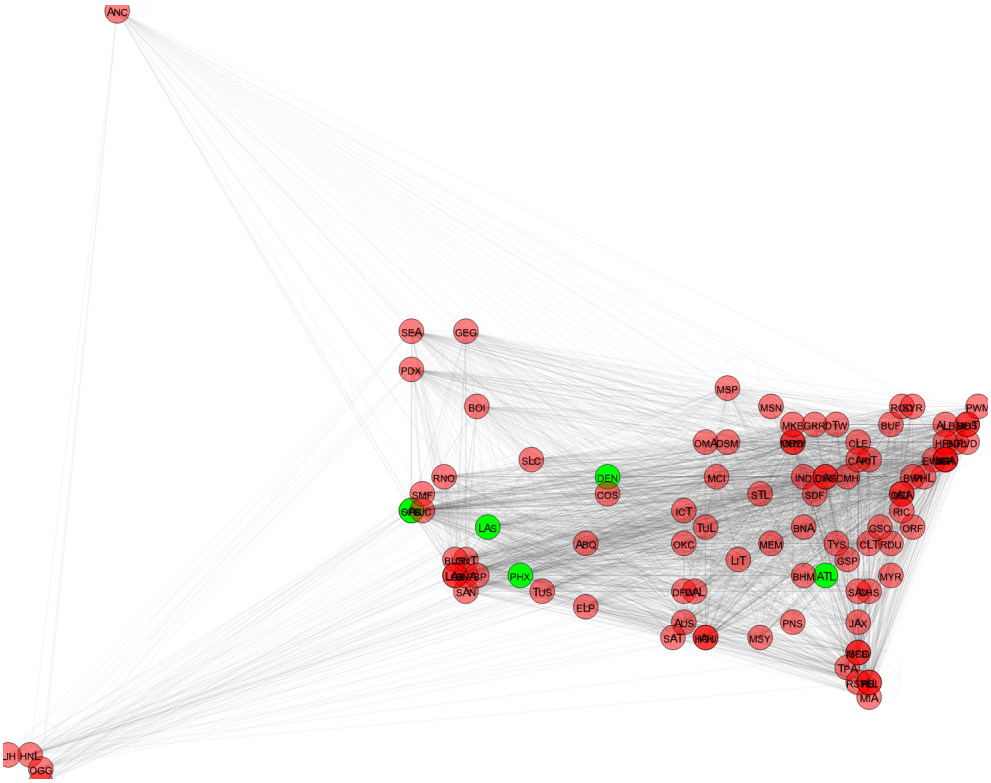


High centrality distributed really close to Eigenvector centrality. However, pagerank pick up some points which are also noted as high centrality in degree one. This is somehow the center of a cluster with also high degree connections.

B. Finding Relative 'Important' Airports

Basing on the analysis above, it is reasonable to make assessments about which airports tend to be 'important'. Three criteria are considered: **Degree** - People have high demands traveling to it. **Betweenness** - Whether it is a transfer hub for cargo and passengers distribution. **Eigenvector Centrality** - High Eigenvector centrality usually shows it somehow a key airport in a cluster and it usually connect to the important airports in other clusters.

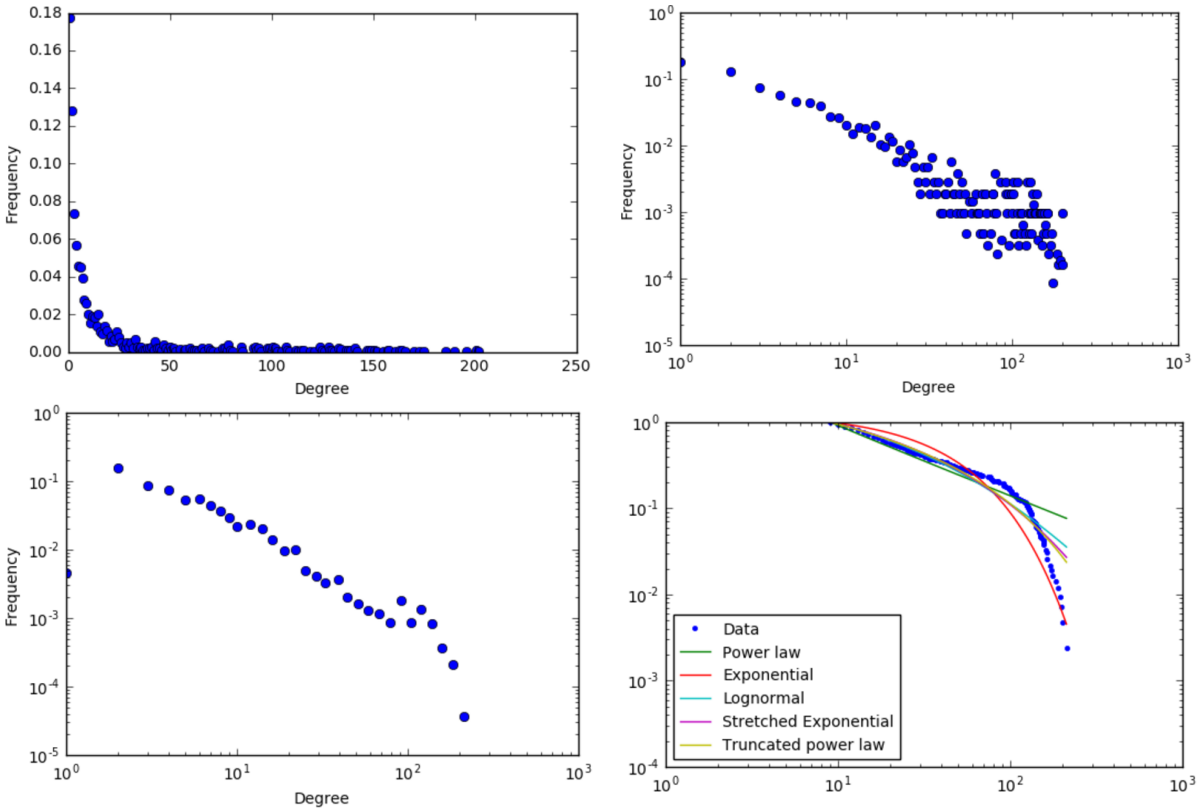
We simply take airports with high rank in degree, betweeness and Eigenvector centrality. Implemented in Python, we have ATL (Georgia), DEN (Denver), SFO (San Francisco), LAS (Las Vegas) and PHX (Arizona) as with actual rank of 2, 6, 7, 8 and 9 in the Top 100 Airports. We can see that an assessment combined multiple criteria gives a pretty good evaluation about importance.



C. Large Network Analysis.

1. Degree Distribution Analysis.

We now focus attentions on airlines transportations data from Bureau of Transportations Statistics for 2016. Graphs of degree distributions and its log scale with noise reduced at right tail are showed.



Several regression methods showed above, we can see that power law fit did a good job for most part, however the right tail drop exponentially which means shows a partial power law degree distribution so that airports with less number of connections follow more closely with a fitted line which has a scale free power law distribution, whereas the airports with more connections follow an exponential decay.

1. Assortative mixing

The assortative coefficient is 0.05718722 which is not a big number, it shows basing on the data of 2016, it is a dissortative network. In that situations, small airports and big airport tend to connect with each other in a mixing way. That also means traveling between two major cities require less layover than rural cities travel because the network tend to like a star structures usually means rural cities need to travel to large high degree cities then transfer to another rural area.

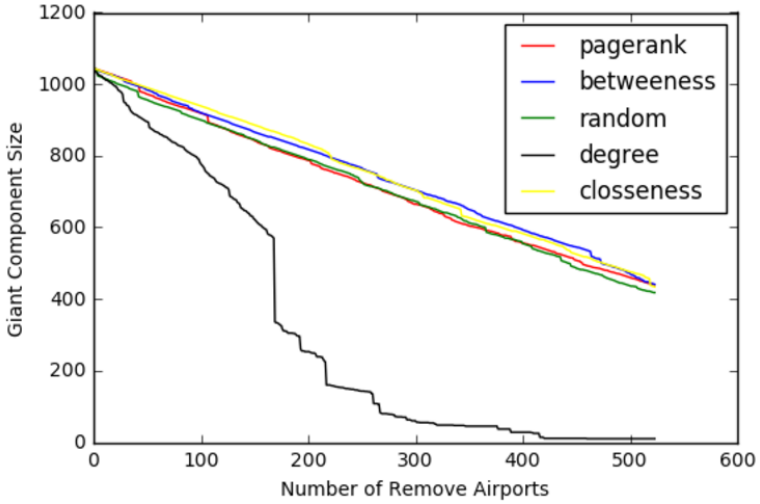
2. Clustering

The clustering coefficient is 0.453900, it is not a very high number. Means that the airports which are close to each other might not be connected with each other and it make sense. Because larger airport usually act like the role of hub to connect small rural airports and due to the expected of flights between airports is made as result of economical optimizations. So that there is no reason to make airlines network a well connected cliques which means big airports neighbors need to communicates each others as well. In that situation, the passagers for those 'unimportant' will be relative small and cause resource waste of flyting spacious airlines.

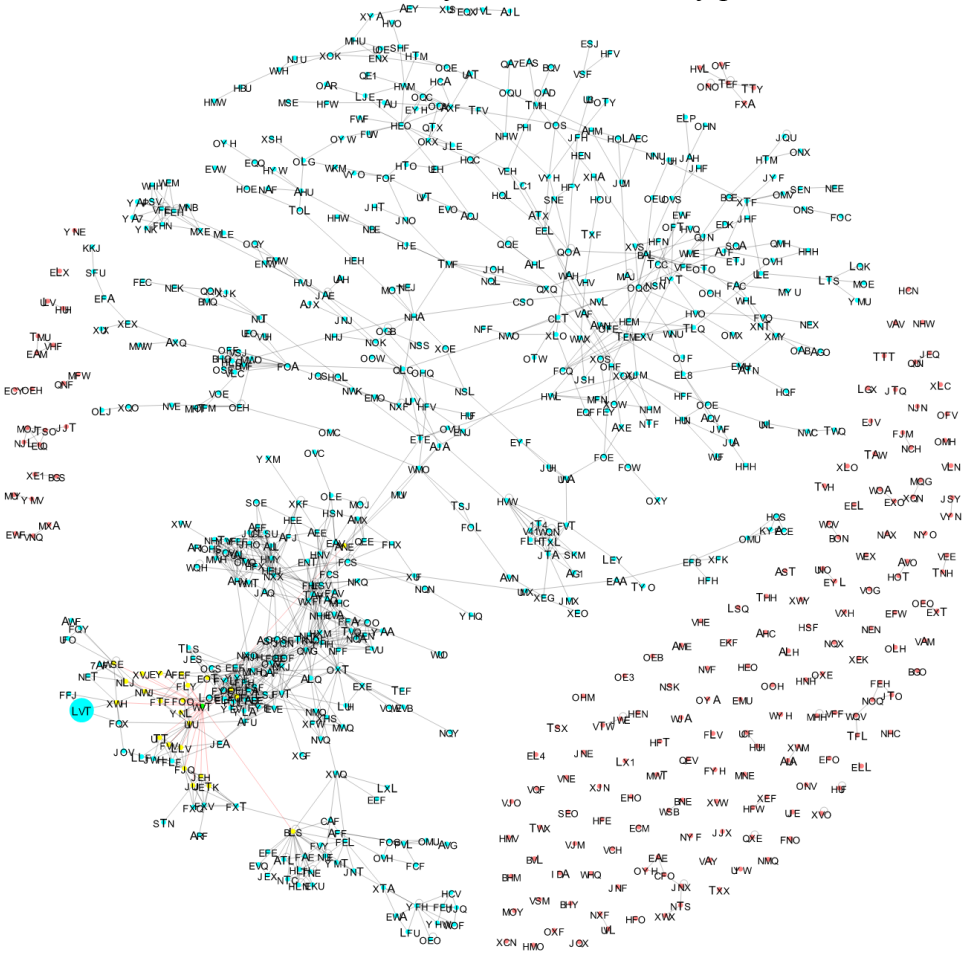
3. Resilience Test

To analyze resiliency, we remove the airports to simulate airports shutdown due to weather issue or even terrorism attack. We remove the airport by 5 ways, first is random, then between, degree, pagerank and eigenvector centrality and remove for those we remove one by one according to their centrality rank. We can see that other than degree centrality, when we delete

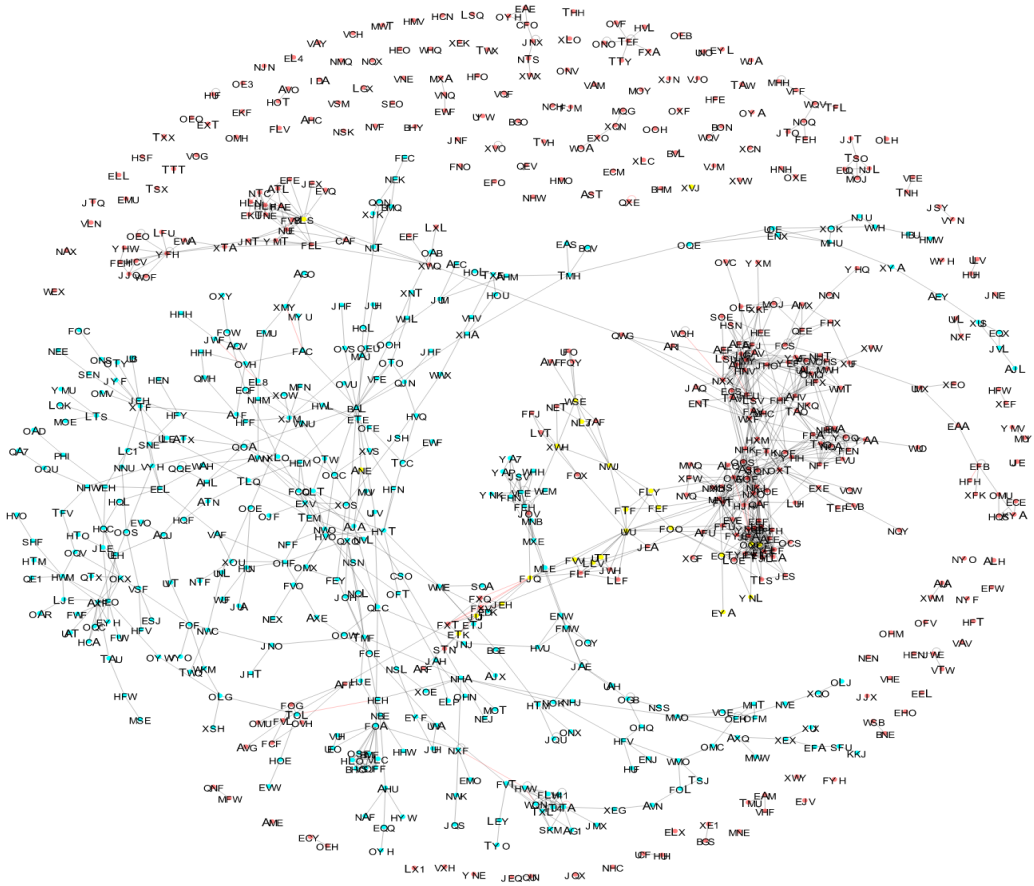
nearly half size then reduce the size of giant components by half. However, in a target attack on highest degree, only around 10% of total airports will significantly reduce the size of giant components.



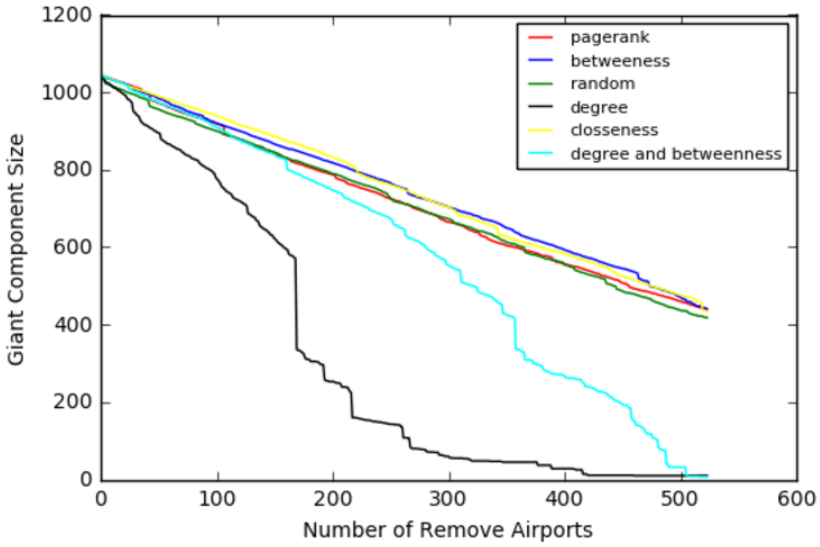
We then take a close look at what key airport been deleted cause the giant components size drop greatly, after we delete ‘DLG’ an airport in Alaska. The size of giant component size reduce from 515 to 311. We want to take look at why it make ‘DLG’ key point for a vulnerable network.



We can see that cyan colored the giant component and light red as non-giant components. Green dot is the DLG before deleted and yellow as the airports connected to DLG. After we deleted it, we can see that yellow dotes fall into the non-giant components and the size of giant decreased a lot. In that situation, DLG as a high betweeness centrality node as the bridge from Alaska to U.S.A mainland. We can see that target at not only on high degree airports but also high centrality airports can cause the air network fall apart efficiently.



From the inspiration from above, we can come up with a new strategy to attack the airports with high degree and betweeness centrality at the same time. After implementation on python we see the results below.

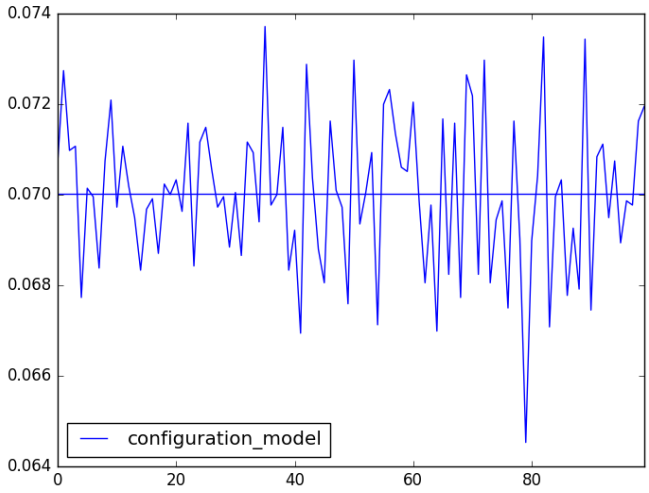


D. Network Modeling

Analysis done before have pointed out specifically that airline network degree follows a power law distribution in general. Our result also shows a great indication of the power law distribution. A natural guess of a network model of power law distribution is a configuration model where the probability of two random airports connected together is: P(A, B) = $\frac{c}{n-1}$.

Given the setting, it is necessary to test the null hypnosis that American airline network is a configuration model. The approach we use is Monte Carle simulation. The goodness of fit of configuration model is quantified based on the simulate data.

Define ***Similarity*** (set_1, set_2) = $\frac{|set1 \cap set2|}{|set2|}$. Then the goodness of k_{th} simulated can be evaluated as: Similarity ($result_k$, TrueData). Where $result_k$ and TrueData are sets of directed edges. We run 100 simulations based on different starting point and random seeds, and keep track of Similarity ($result_k$, TrueData). The result was shown as the plot below:



The mean of Similarity is **0.070**, with standard deviation of **0.0017**. This suggests that only 7.0% of the edges that show up in the actual data was successfully explained by the configuration model, which is a rather bad result.

While configuration model is a good model for this particular data, given other information X of vertices, such as in/out degree, distance between two edges and geographic information, it is possible to construct a better estimator for the probability of two random airports connected together using maximize posterior probability decision. For the sake of convenience and interoperability, logistic regression is used:

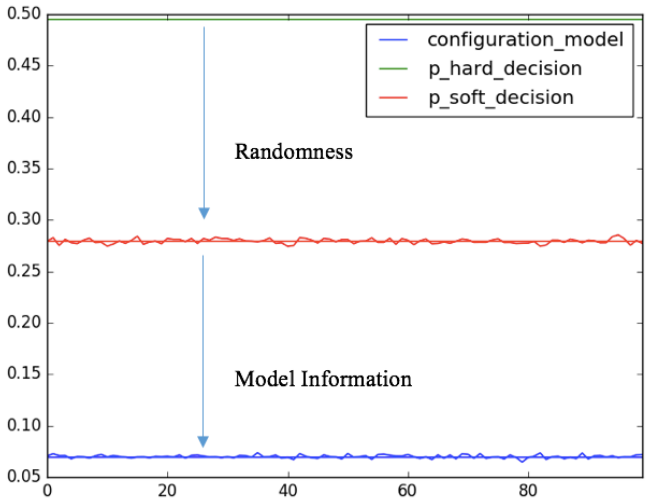
$$P(A, B) = \text{Sigmoid}(\beta_0+\beta_1\text{distance}(A, B) +\beta_2\text{degree}(A)+\beta_3\text{degree}(B) +\beta_4\delta(A, B \text{ are in the same state}))$$

Where $\boldsymbol{\beta} = \arg\max_\beta \text{MSE}(\text{predict}, \text{trueth})$, a trained best decision parameter.

After replacing P (A, B) in the original configuration model, two strategies of generate edge can be made based on P (A, B):

- Hard Decision: Choose top k of P (A, B_i)'s Bs as destination airports leaving from A.
- Soft Decision: weighted random sampling without replacement based on P (A, B).

Then we apply Monte Carlo method just as before, and use the same Similarity measurement to estimate goodness of simulation. Results by comparison are shown as below:



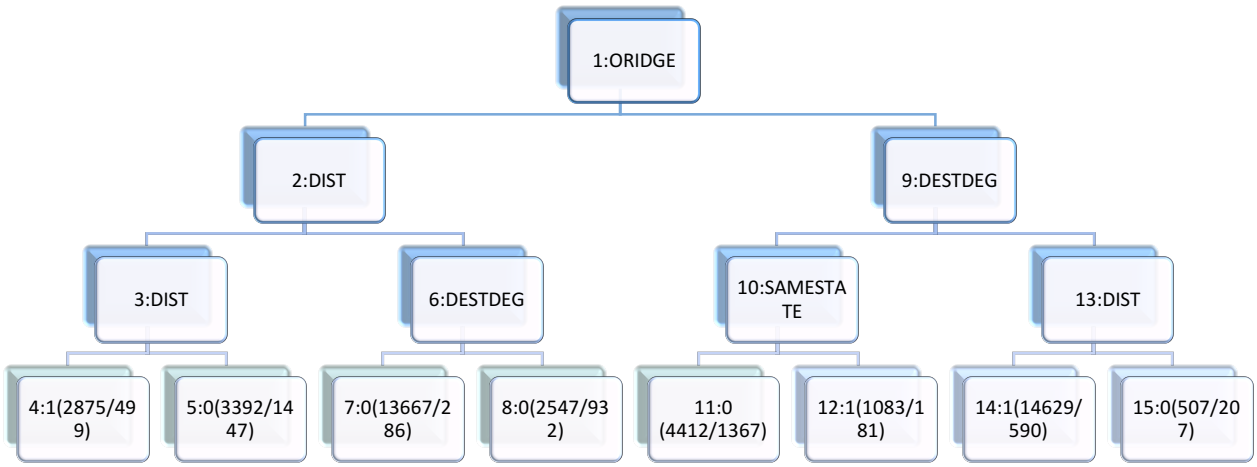
The result shows soft decision is of mean **0.279** and standard deviation of **0.0023**, and hard decision is of constant value **0.495**. Compare to the configuration model, these are great improvements.

Further statistics test rejects the null hypothesis that configuration model results have higher mean than soft decision model (α=0.005). This is a strong evidence that our model is better, and by inferences that there only exists one true model, rejects the hypothesis that true model is a configuration model, because of the existence of a better model.

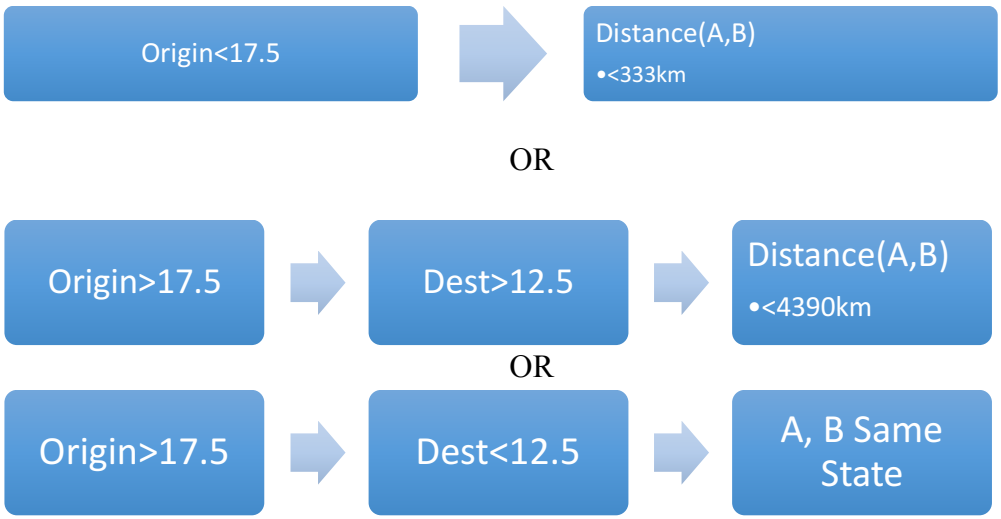
Notice that the differences between soft and hard decision are caused by purely randomness of sampling method, given two decisions are provided with exactly the same information P (A, B). While the difference between soft decision and configuration model is caused by the augmented P (A, B), given two models are both using weighted random without replacement strategy (where configuration model has uniform weights).

E. Model Interpretation

To further interpret the information that provided by the feature carried by the vertices (degrees, locations, and geographic information), a decision tree model is generated to help with rule construction:



By selecting branches that generate 1s (indicator of edge exists), three qualities that Origin, Destination airports must have that are in favor of airline formation is generated:



In conclusion, the American airline connect patterns can be mostly categorized by the following cases:

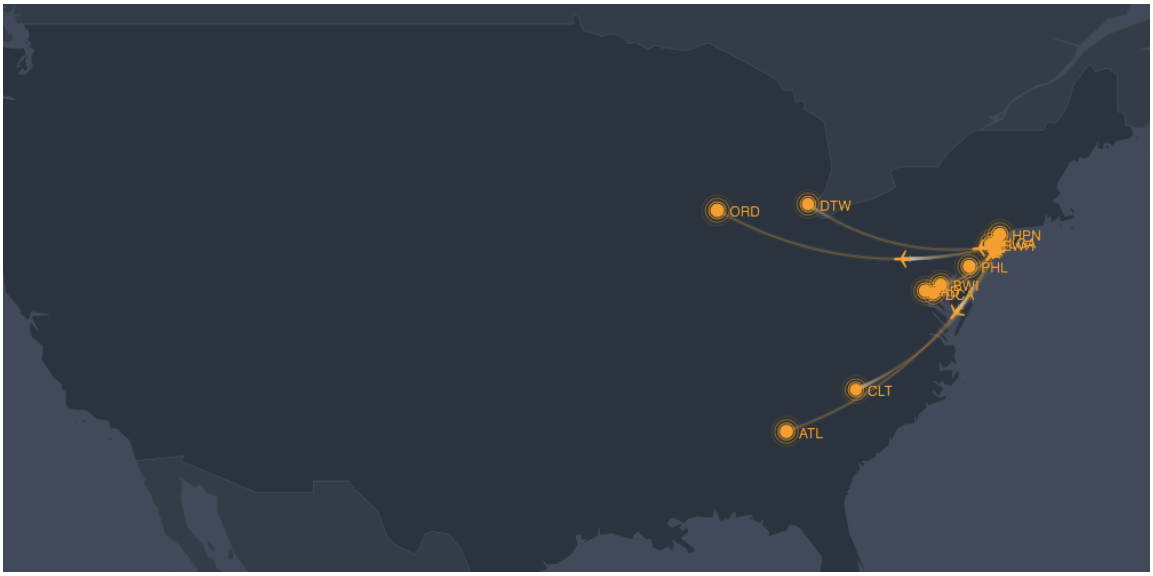
- Small airports that are close to each other.
- Large departure airport and large arrival airports that are within 4390km.
- Large departure airports and small arrival airports that are in the same state.

F. Case Analysis Under Hard Decision Model

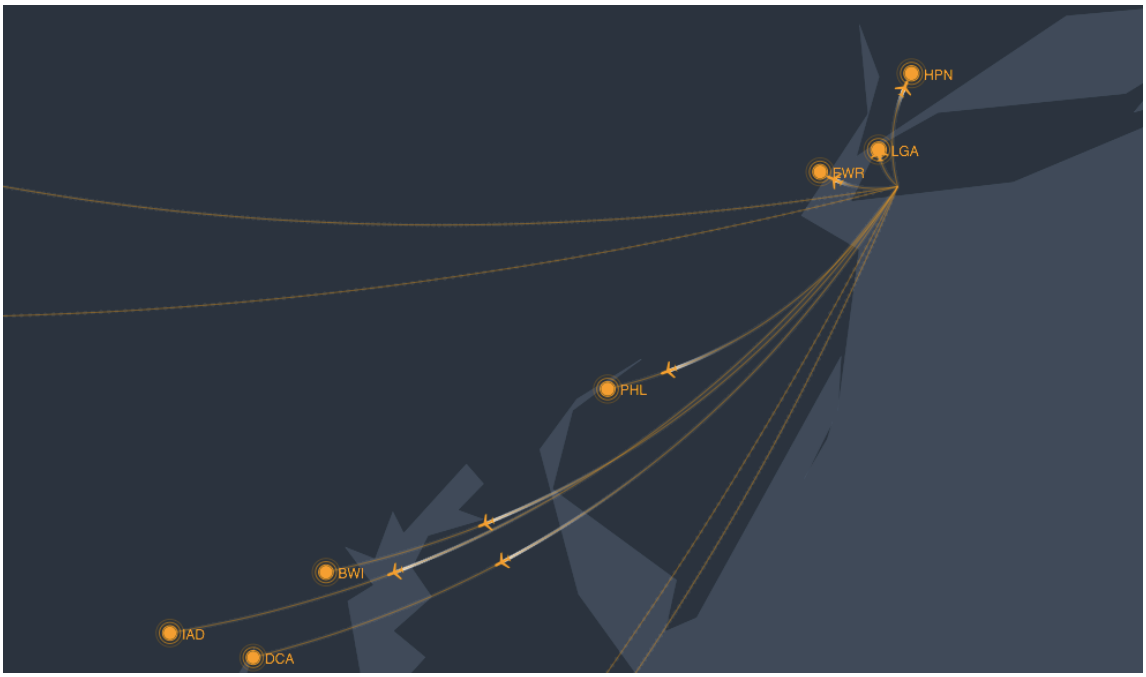
Under Hard Decision Model, the sequence of generated edges that leaving certain major airport (JFK, ORD) can be visualized by tiers of ten.

The key observation is that first 10 edges formed by the model are either leaving for important (high degree) airports of middle range distance, or close middle size airports that are close the origin. Also, as tiers goes down, that is as the process of generating edges goes on, more remote airports are added to the graph, and smaller airports shows up.

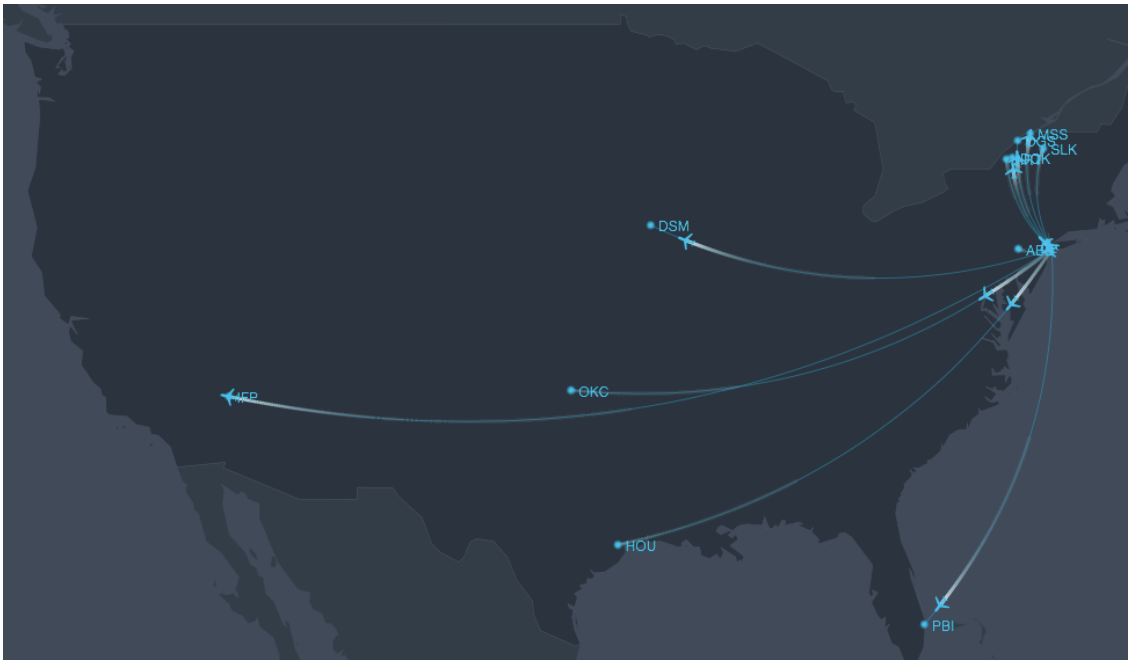
Partial visualization results are shown below:



first tier of edges, connections are closer and to bigger airports.



first tier of edges, zoomed into New England region.



tenth tier of edges, connections reach further, and to smaller airports.

IV. Conclusion

Basing on the centrality analysis, give us an idea that different airports play different role. We can see that the top 100 airports indeed have somehow aspects for degree, betweenness or pagerank. We think the geographic characteristics of U.S and people needs for transportations to important cities. We think this centrality analysis help people to protect do relative important airports not only those with high degree. So that it give us a sense about where area might act like bridge for transportation and people may pay more attention to care about those airports. The partial power law degree distribution may imply that airports in the densely populated area are grow slower rate than less populated areas. Basing on the resiliency test results. Degree focus attack airlines network become vulnerable. After remove certain number of airports some 'key' airport make the network suddenly fall apart and it somehow give us an idea about how to protect our airlines connections. The assortativity and cluster coefficient implies that on average, travelers experience 2 transfer before reaching the destinations. Also travel between big cities usually requires less stop due to economics the profit considerations.

The airline model Monte Carlo experiment rejects the hypothesis that airline model is a configuration model, and suggests that the models we proposed are better. By reconstructing network, at most **49.5**% of the edges can be explained by our model. Also three rules that in favor of a connection are concluded, and visualized.

V. Future Work

In additional simply consider distance as weight, we should put more elements into account such as flight frequency for different routes and passenger compacity for these routes. Also these analysi help us to precisely give us advice about wheather to build new airport in certain area or should we enlarge the shirnk the exsisting airports basing on it's importance and handling capacity.

Also, model generation based on machine learning techniques can be improved by using data from other country to generate training model to prevent over-fitting and self-explaining. Also,

features from community detection results can be added to the model to increase accuracy. Further more, features that rely on observing degree, and other prior information of the network can be replaced with other independent features such as city population, size, and economic situation. In this case, model can be purely prior-free. These improvements will assist with real-life prediction application. For example, optimizing airline connection for airline-companies, contributing to decision making of adding a new airline, and exploring underlying airline patterns.

VI. References:

[1] Hua Yang, Yuchao Nie, Hongbin Zhang, Zengru Di, Ying Fan. *Insight to the express transport network* [J]. Journal of Industrial and Systems Engineering, 2015(8.4): 106-121

[2] Cheung, D. P., & Gunes, M. H. (2012). A Complex Network Analysis of the United States Air Transportation. 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. doi:10.1109/asonam.2012.116

[3] R. Guimer, L.A.N. Amaral. Modeling the world-wide airport network[J]. *The European Physical Journal B*, 2004(38) : 381–385

[4] Zengwang Xu, Robert Harriss. Exploring the structure of the U.S. intercity passenger air transportation network: A weighted complex network approach[J]. *GeoJournal* 73, 2008 (12): 87–102

jose ramasco, IFISC Mallorca