American Domestic Airline Network Analysis

PHY 525 Data Science II : Network Presentations

Yuanzhen Du
Linxiao Bai
Keting Lyu

December 12, 2016

1. OBJECTIVES
2. DATA ACQUISITION
3. NETWORK ANALYSIS
4. AIRLINE MODELS

1

OBJECTIVES

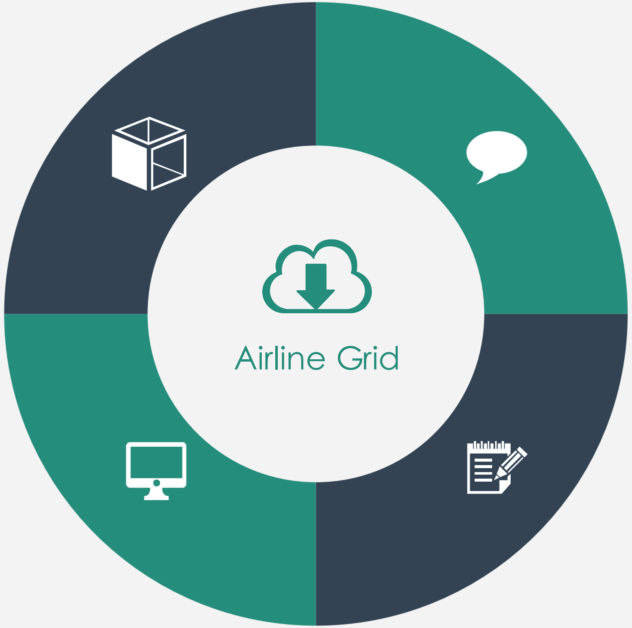
OBJECTIVES

Visualization

American major airports traffic grid, with respect to geometric location

Centrality analysis

Determine the important airports in American.



Resiliency analysis

In case of potential airport shut down, and its impact on the transportation grid.

Network model hypothesis

OBJECTIVES

Air Traffic network model hypothesis test

- Random graph model with degrees of power-law distribution
- Random graph model with degrees of Poisson distribution
- Configuration model with degrees of the original distribution

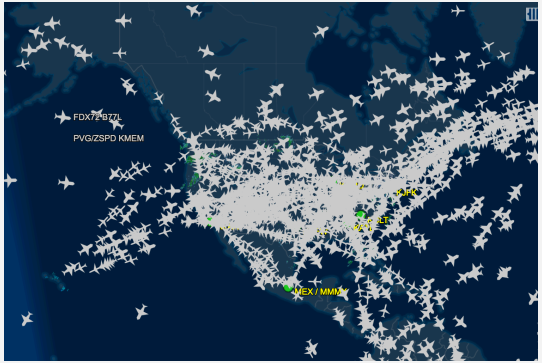
2

DATA ACQUISITION

DATA ACQUISITION

UNIQUE_CARRIER_NAME	ORIGIN	ORIGIN_CITY_NAME	DEST	DEST_CITY_NAME	DISTANCE_GROUP
Tradewind Aviation	ACK	Nantucket, MA	BNA	Nashville, TN	2
Tradewind Aviation	ACK	Nantucket, MA	HPN	White Plains, NY	1
Tradewind Aviation	ACK	Nantucket, MA	SUA	Stuart, FL	3
Tradewind Aviation	BDR	Bridgeport, CT	UBF	Chatham, MA	1
Tradewind Aviation	BNA	Nashville, TN	ACK	Nantucket, MA	2
Tradewind Aviation	BNA	Nashville, TN	TEB	Teterboro, NJ	2
Tradewind Aviation	BOS	Boston, MA	HPN	White Plains, NY	1

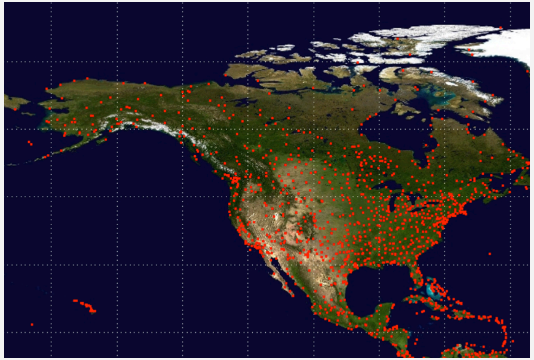
- The airline data is obtained from Bureau of Transportation Statistics.
- Airport is uniquely identified as the airport code.



DATA ACQUISITION

AIRPORT CODE	AIRPORT_NAME	AIRPORT_CITY_NAME	STATE_NAME	LAT_DEGREES	LON_DEGREES
AIA	Alliance Municipal	Alliance, NE	Nebraska	42	102
AIA	Alliance Municipal	Alliance, NE	Nebraska	42	102
AIA	Alliance Municipal	Alliance, NE	Nebraska	42	102
AIZ	Lee C Fine Memorial	Lake of the Ozarks, MO	Missouri	38	92
AIZ	Lee C Fine Memorial	Lake of the Ozarks, MO	Missouri	38	92
B19	Biddeford Municipal	Biddeford, ME	Maine	43	70

- Geometric location of airports comes from http://openflights.org/data.html.



- Airports locations and fligh details are merged together to form an integrated data for analysis purposes.

3 NETWORK ANALYSIS

CENTRALITY

First we analyze the most important top 100 airports

Red

Light Red

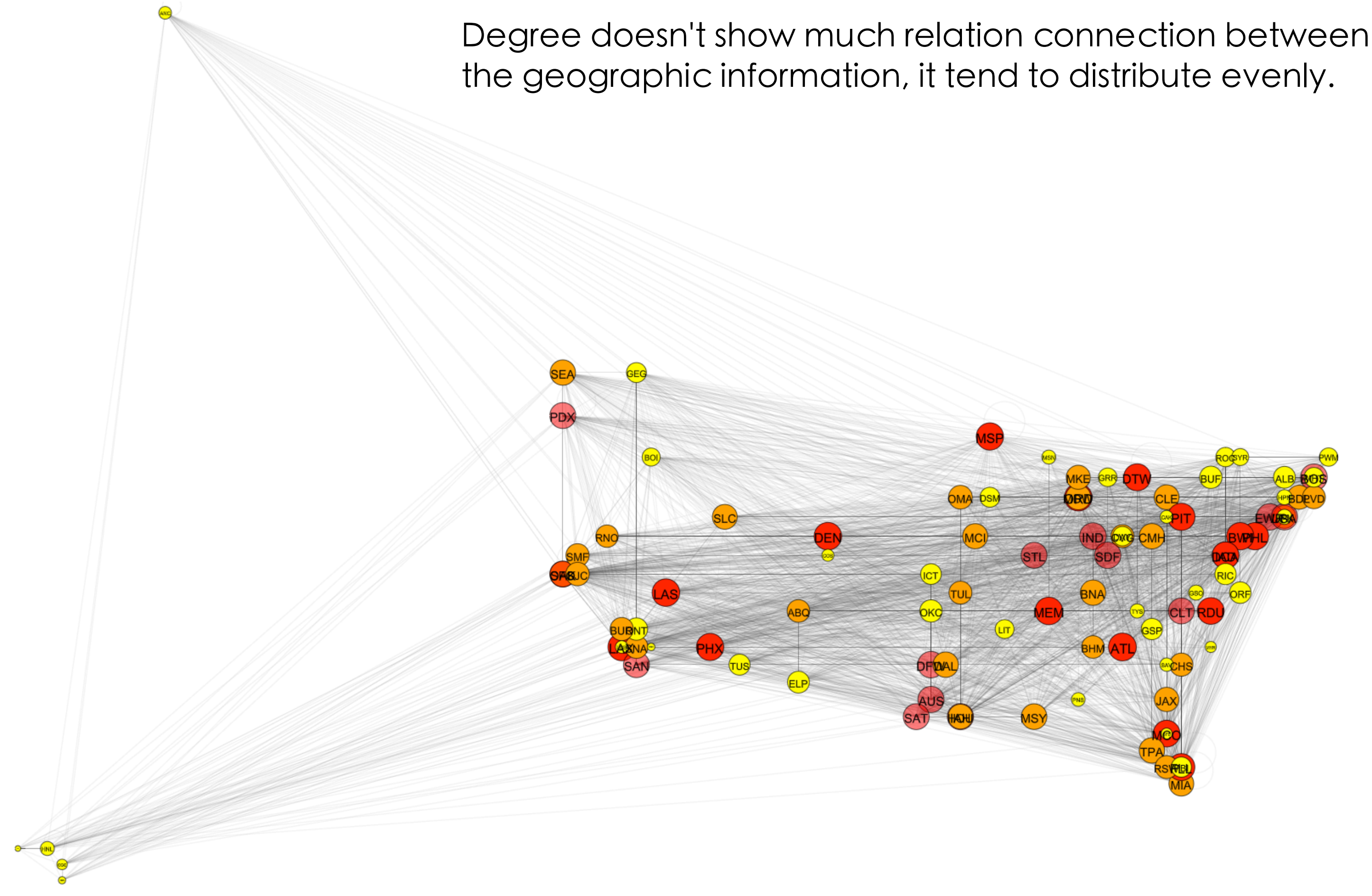
Orange

Yellow

We draw three graph for different centrality and color rank from red, light read, orange to yellow

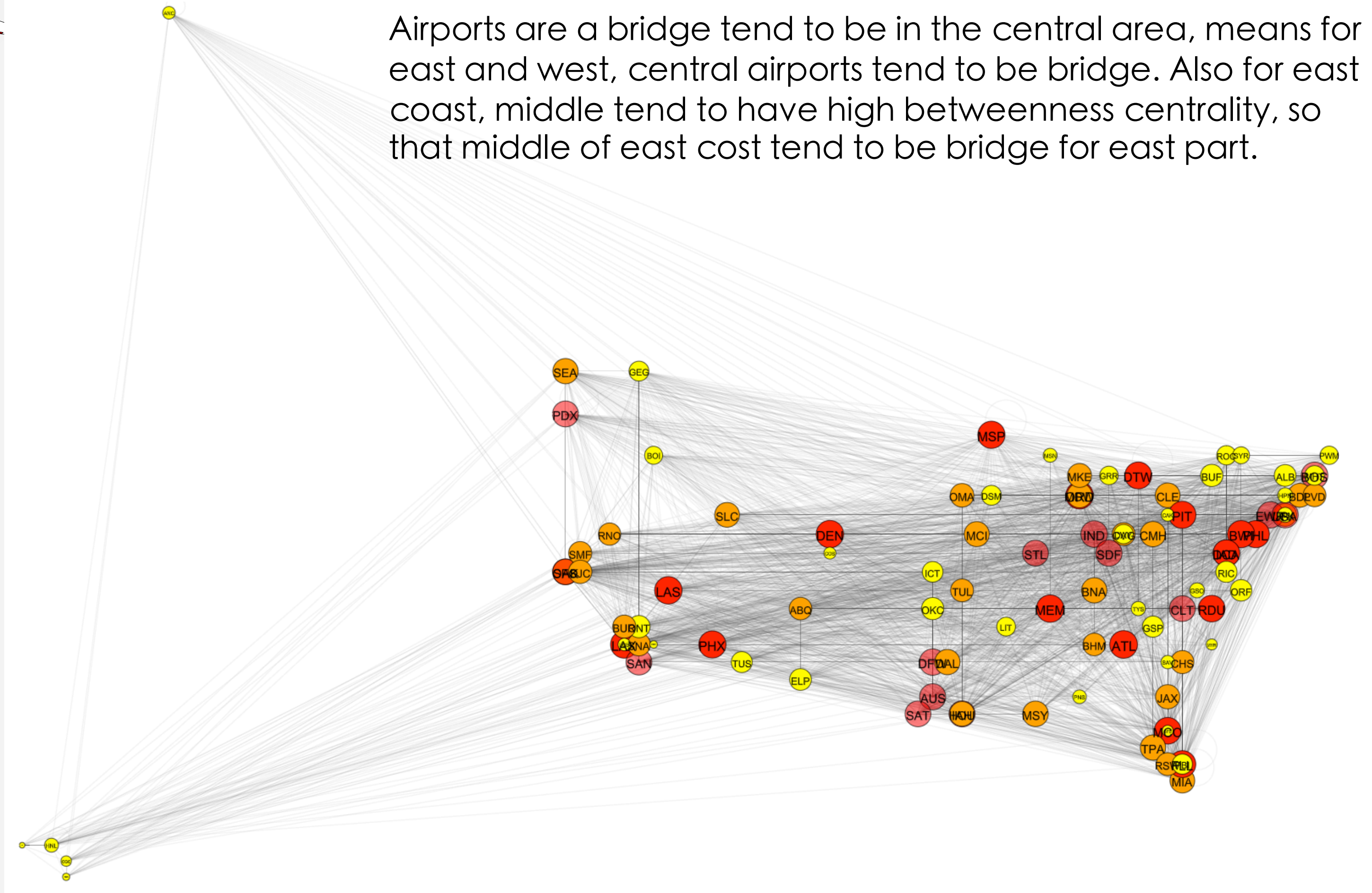
DEGREE CENTRALITY

Degree doesn't show much relation connection between the geographic information, it tend to distribute evenly.

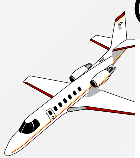


BETWEENNESS CENTRALITY

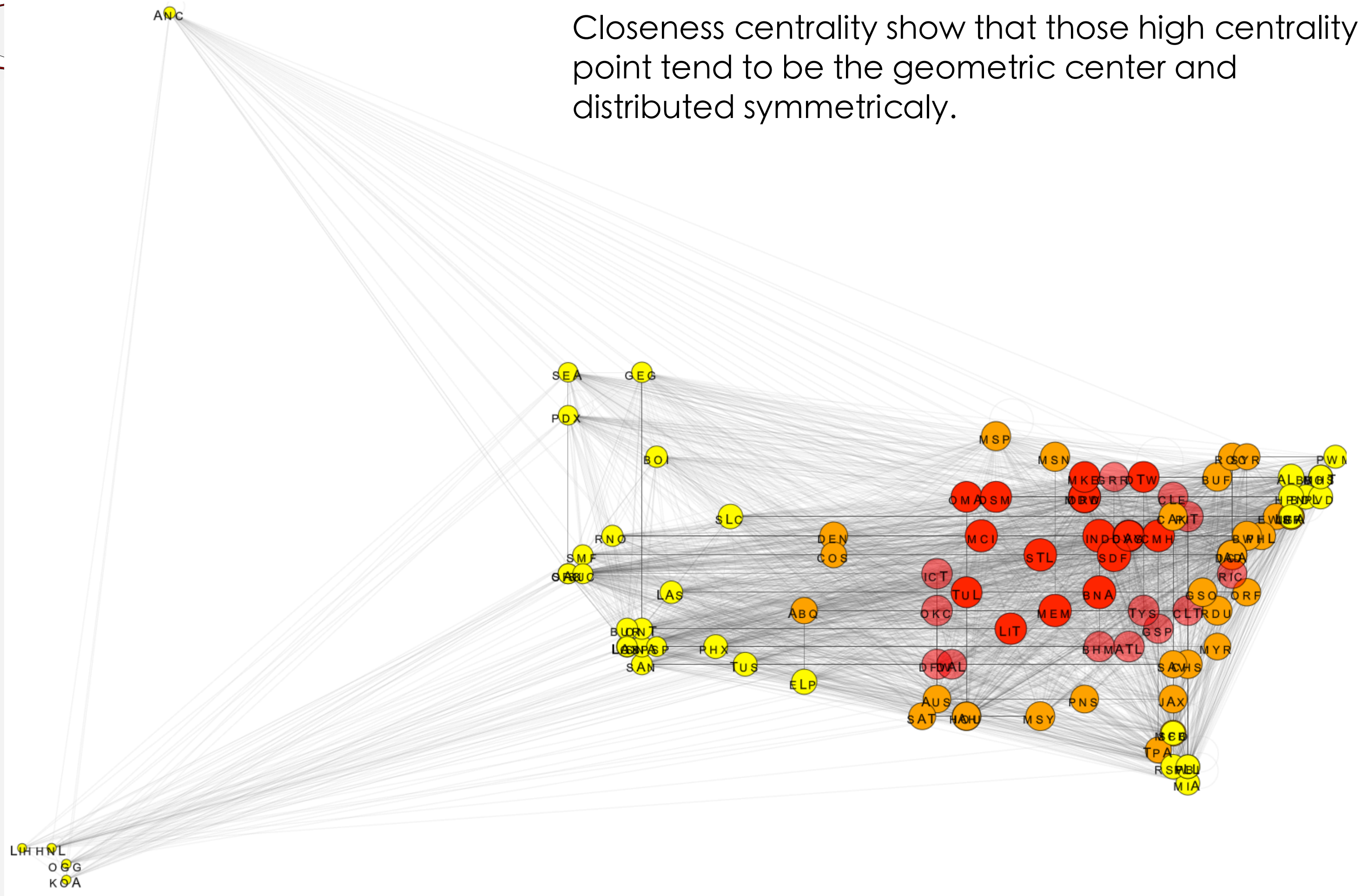
Airports are a bridge tend to be in the central area, means for east and west, central airports tend to be bridge. Also for east coast, middle tend to have high betweenness centrality, so that middle of east cost tend to be bridge for east part.



CLOSENESS CENTRALITY

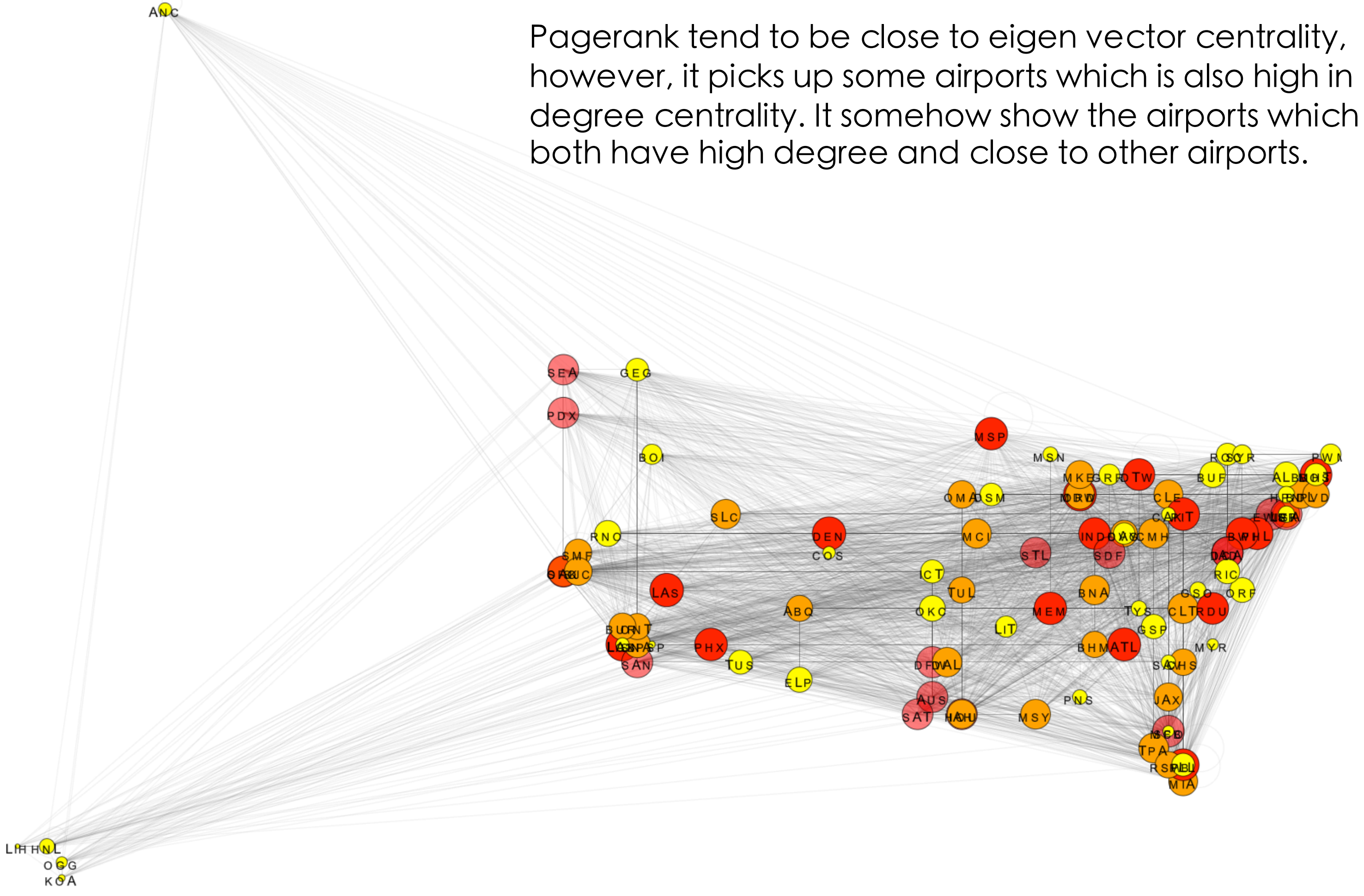


Closeness centrality show that those high centrality point tend to be the geometric center and distributed symmetricaly.



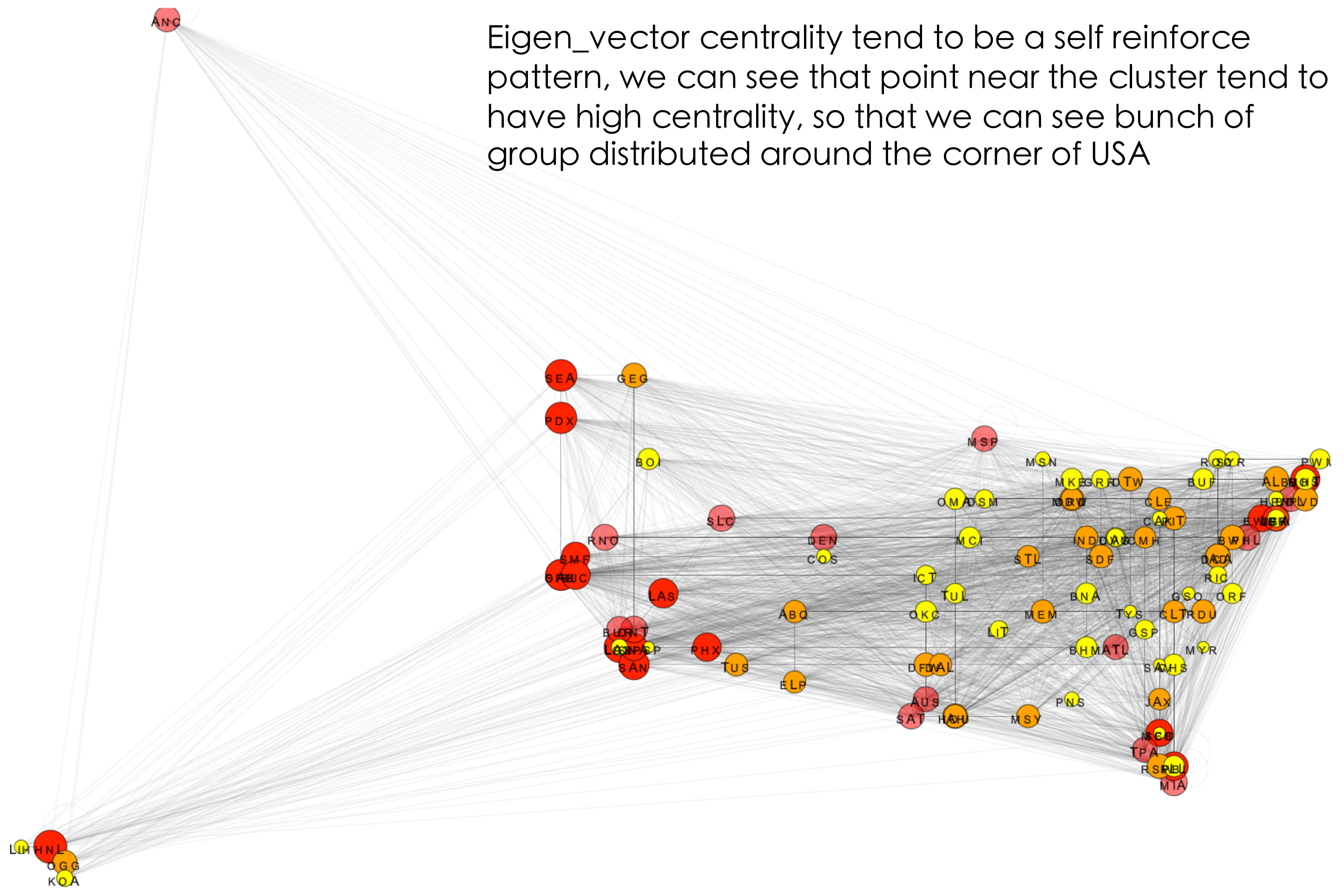
PAGERANK CENTRALITY

Pagerank tend to be close to eigen vector centrality, however, it picks up some airports which is also high in degree centrality. It somehow show the airports which both have high degree and close to other airports.

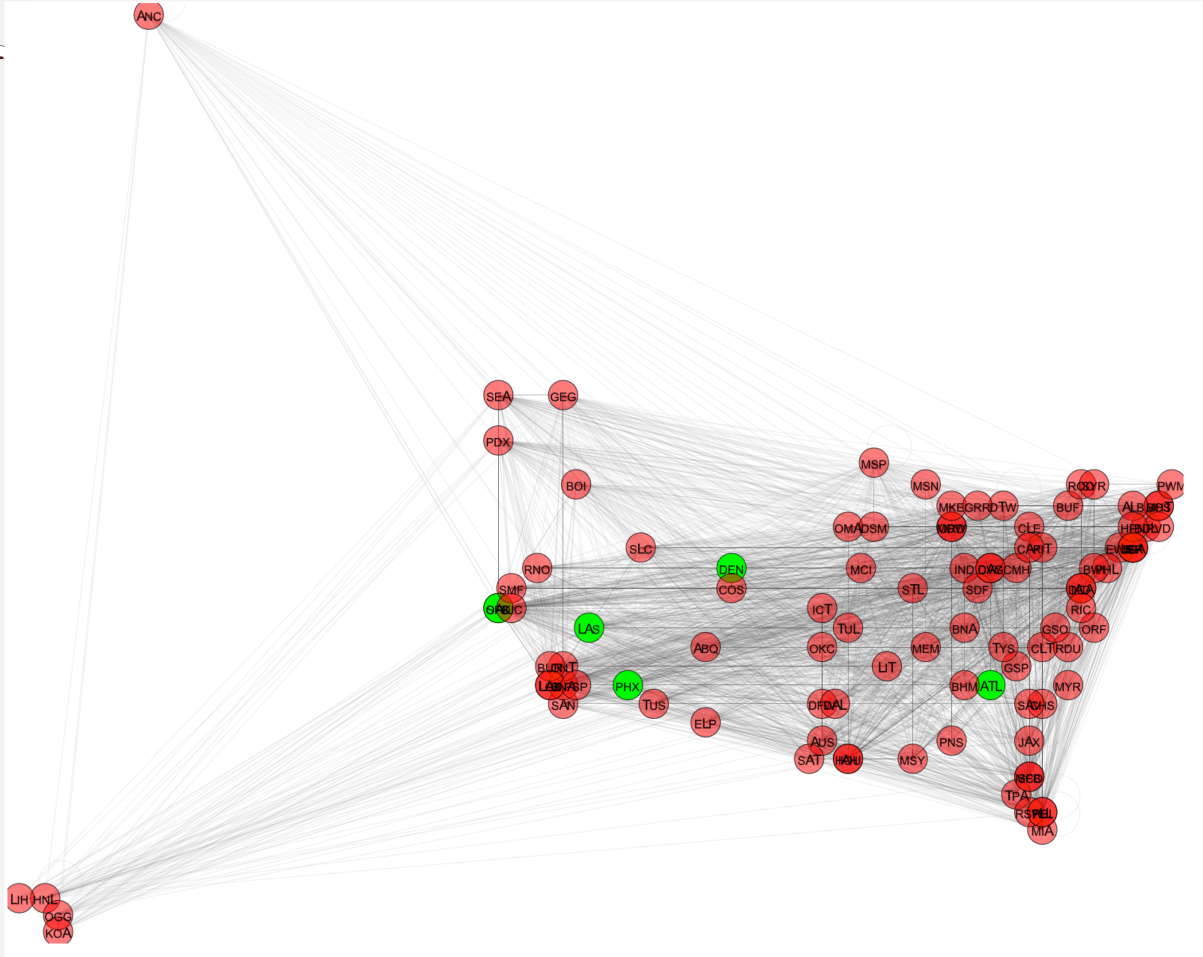


EIGENVECTOR CENTRALITY

Eigen_vector centrality tend to be a self reinforce pattern, we can see that point near the cluster tend to have high centrality, so that we can see bunch of group distributed around the corner of USA

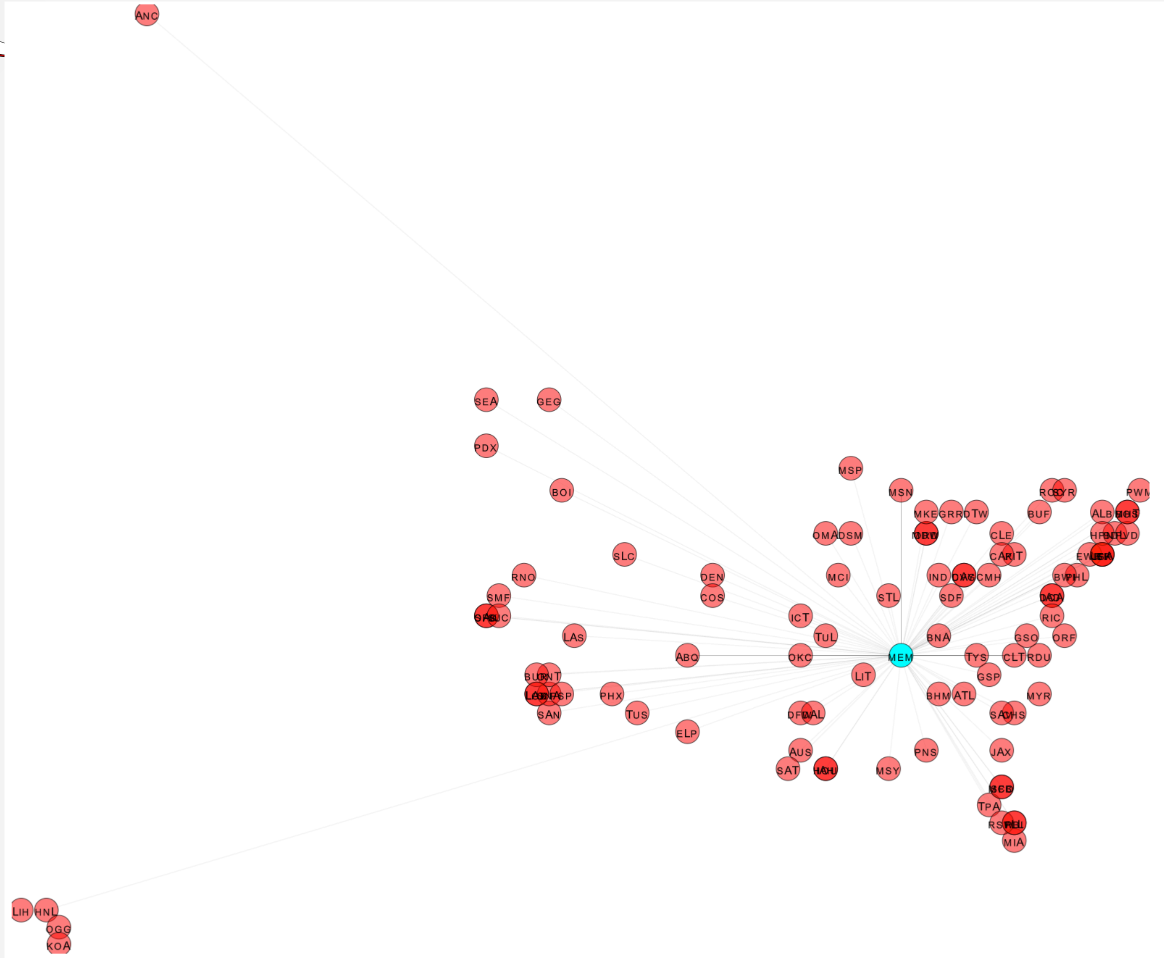


Relative "important" airports distribution



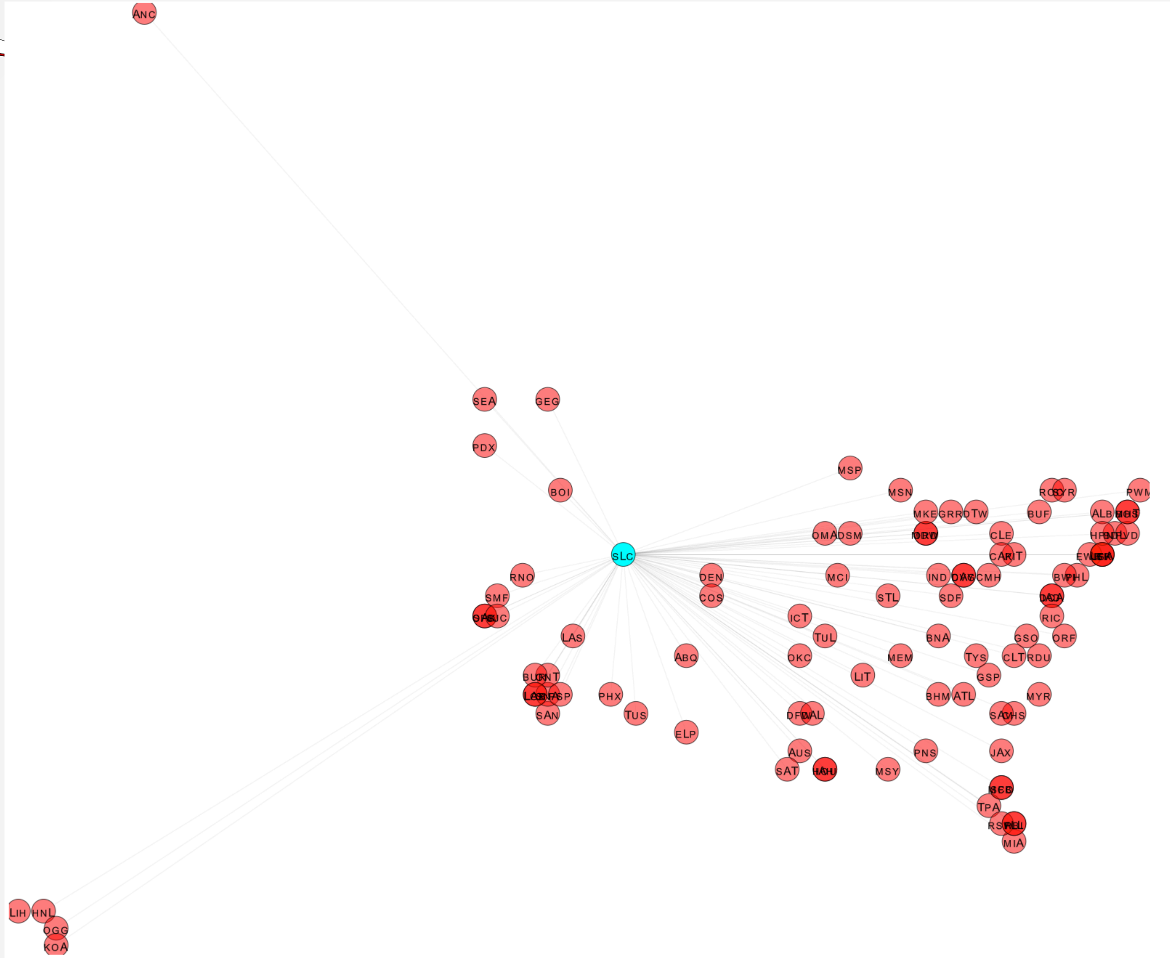
- Here we assume that we need a relative general score for 'important' 100 airports, so that we pick up those airports with high degree, betweeness and eigen_vector centrality.

Single airport properties for different centrality characteristic



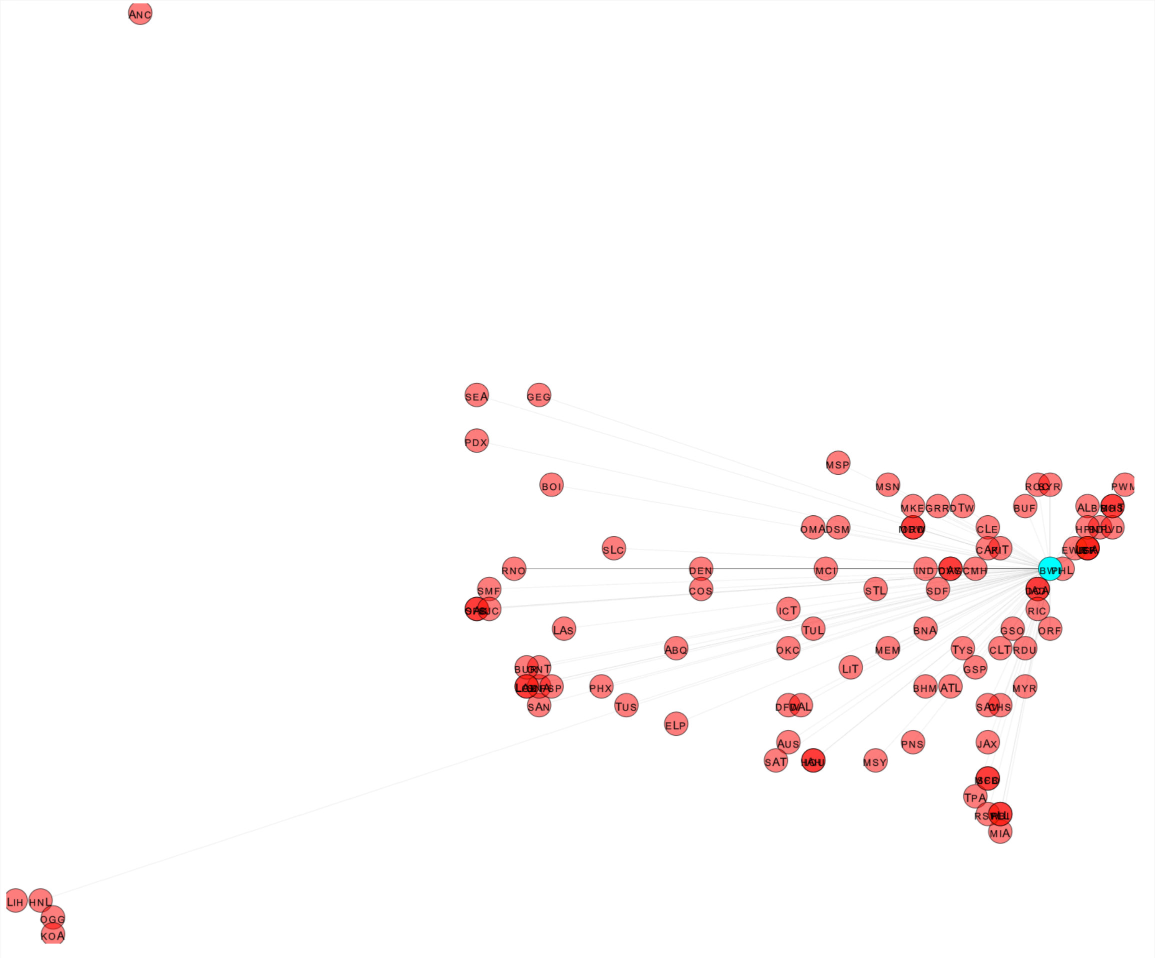
- We choose MKM for high closeness represent
- SLC as represent for betweenness,
- BOS or BWI for eigen_vector centrality

Single airport properties for different centrality characteristic



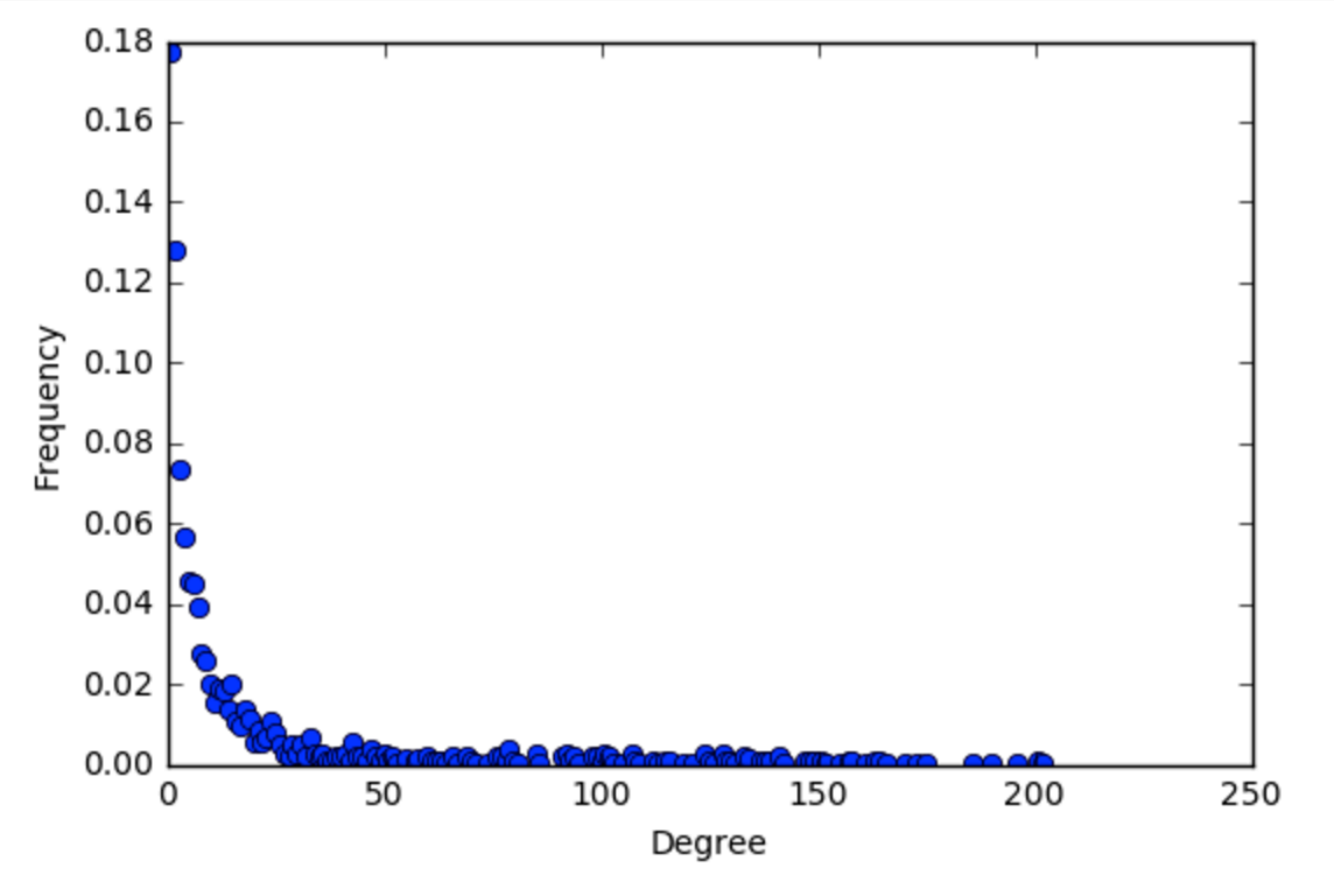
- SLC tend to act like a bridge since we can see not many airports around it we it probabily act like a bridge from east to west coast. Since airport may stop at SLC after long distance.

Single airport properties for different centrality characteristic

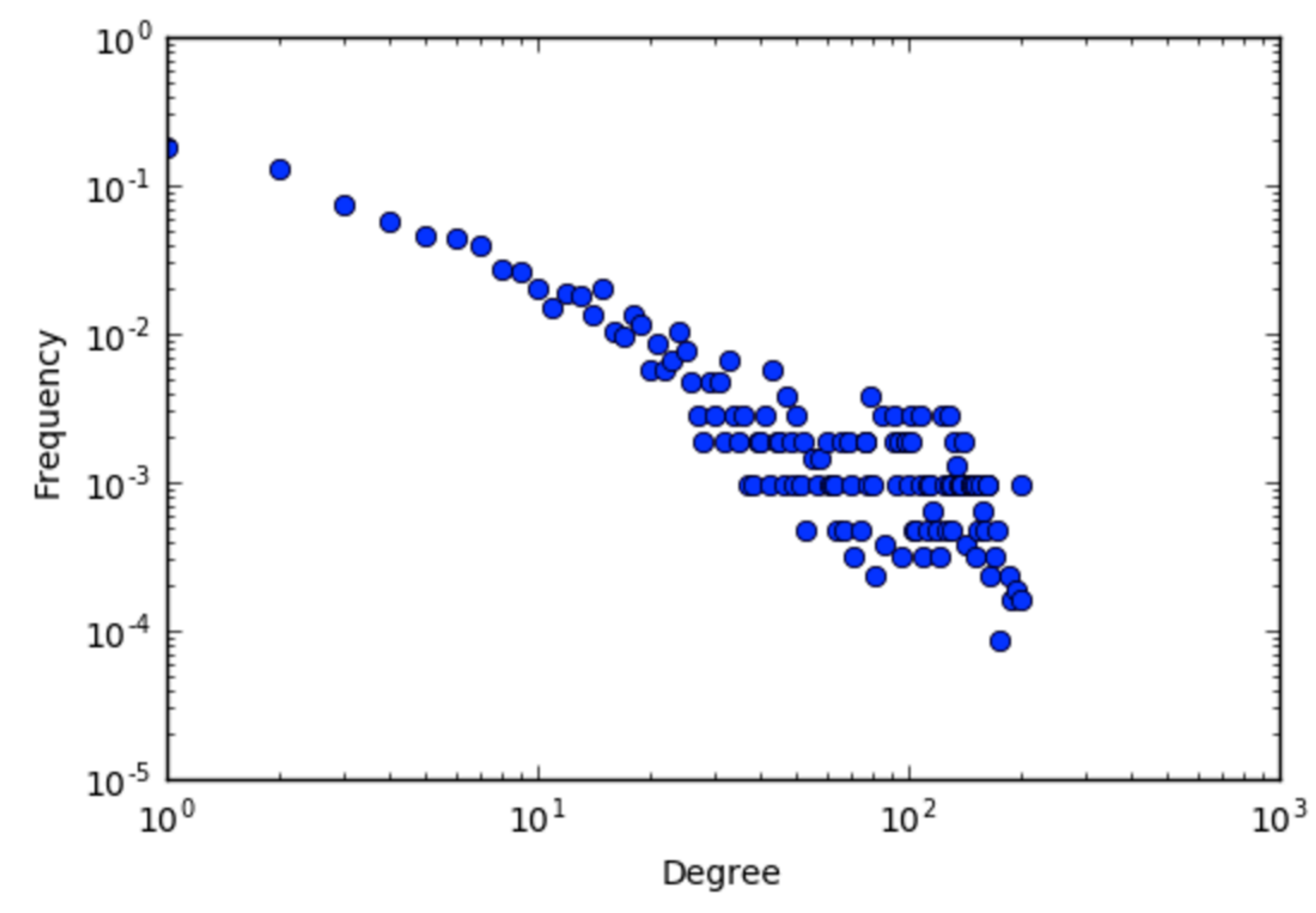


- We can see that airport with high eigen_vector centrality tend to be locate at the corners or coast of the USA, because they to do have similar connections pattern and belong to the same cluster.

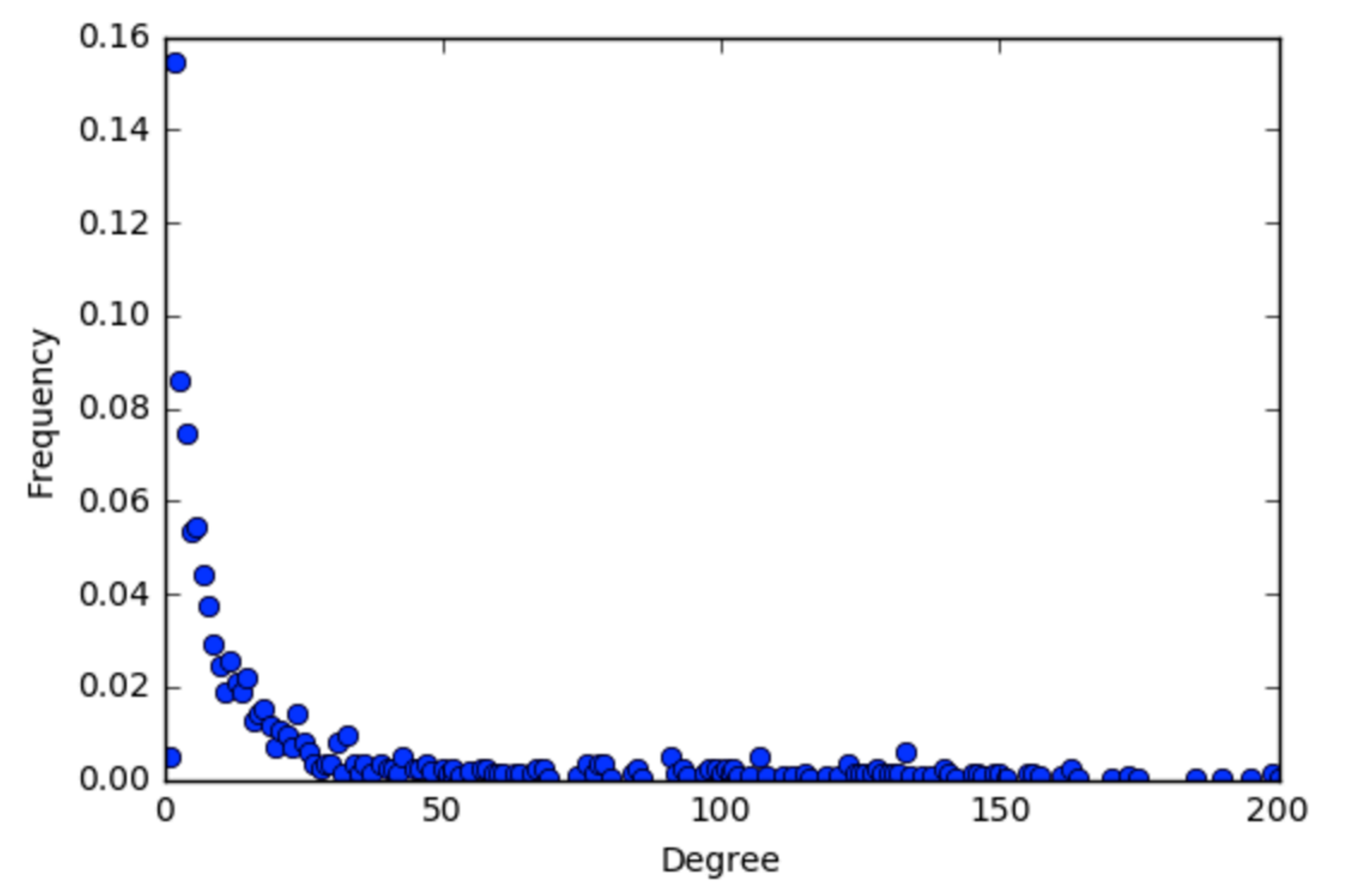
Degree Distribution Analysis



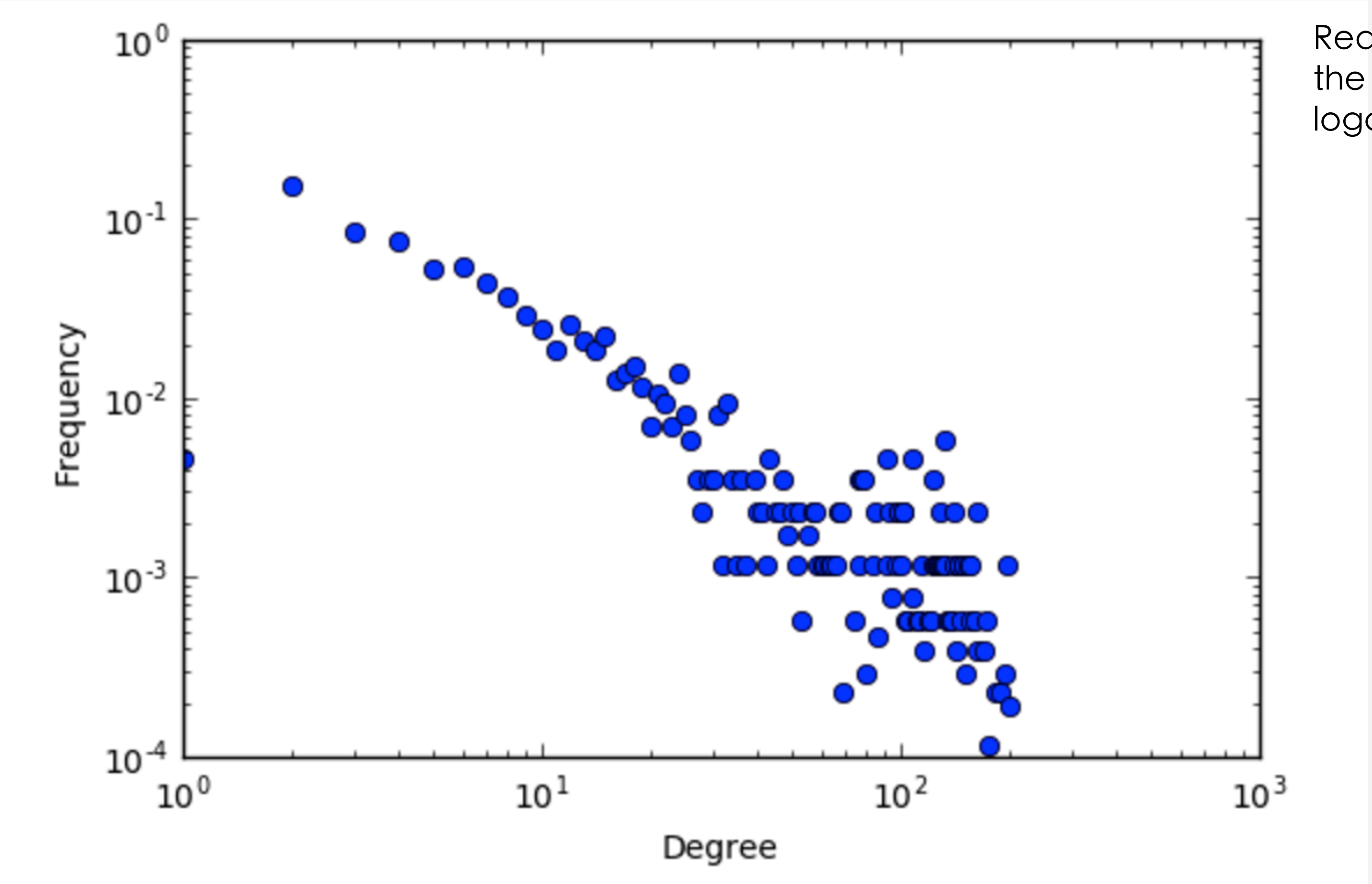
Degree Distribution Analysis



Degree Distribution Analysis

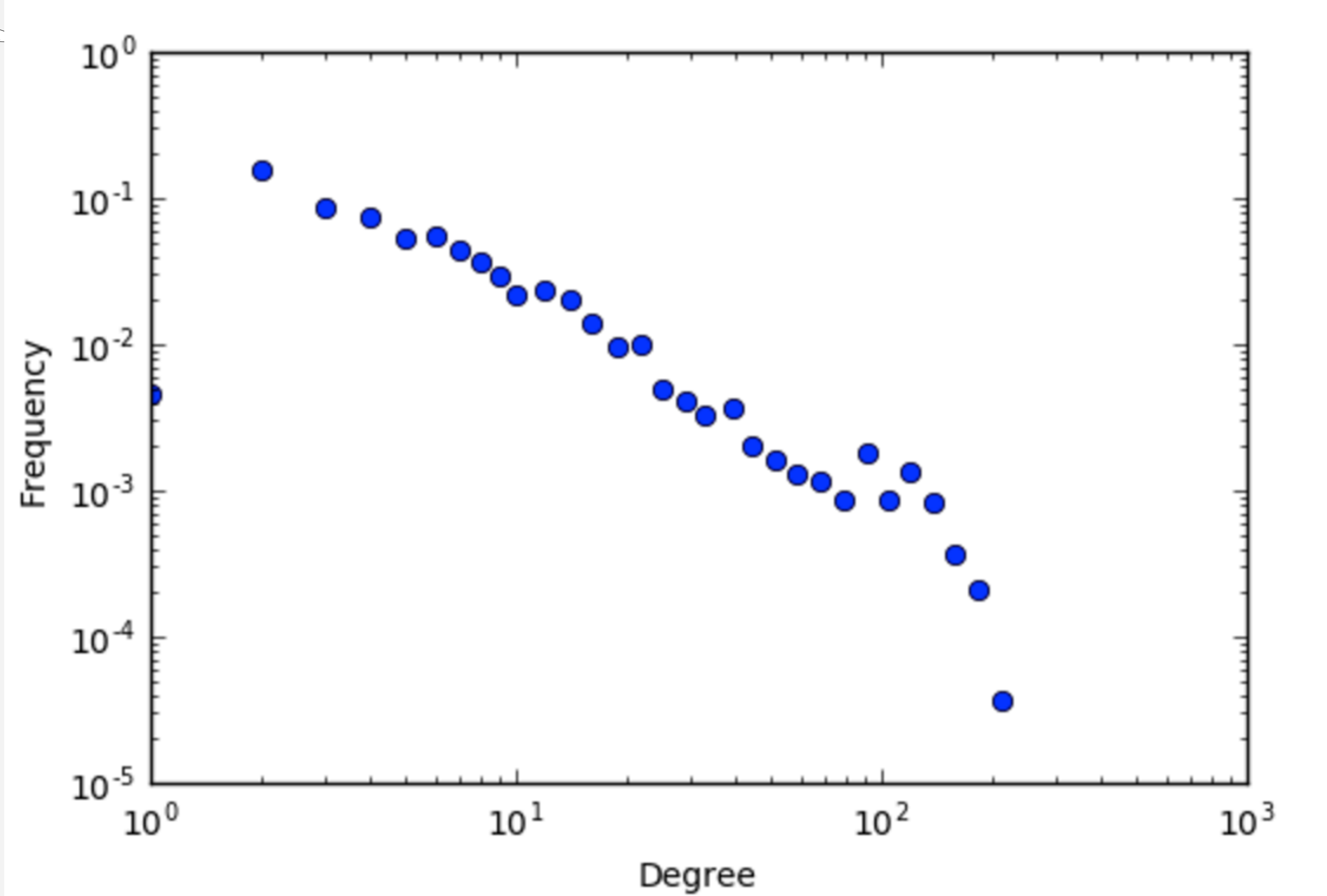


Degree Distribution Analysis



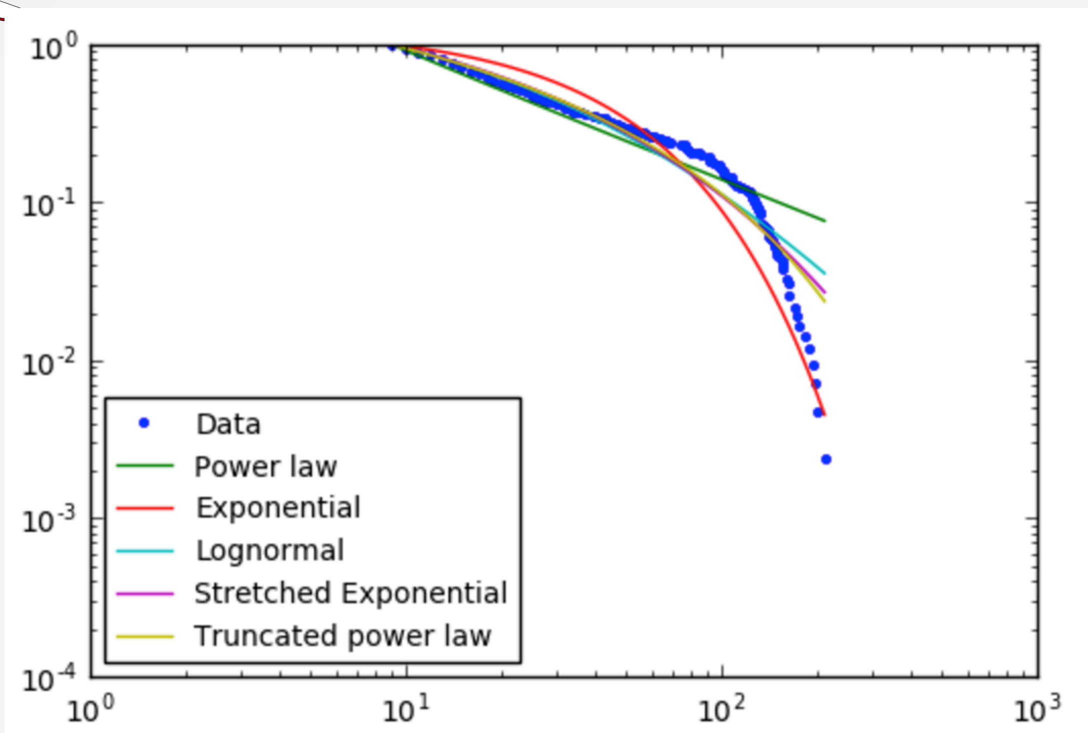
Reduce the noise in the tail by using a logarithmic binning

Degree Distribution Analysis



Try to fit the line

Degree Distribution Analysis



Calculating best minimal value for power law fit

We can see that since the number of airports with high degree drop so quickly so that we can see it goes down exponentially. So that we can see most part can be fit by power law.

Assortativity

0.05718722253490576

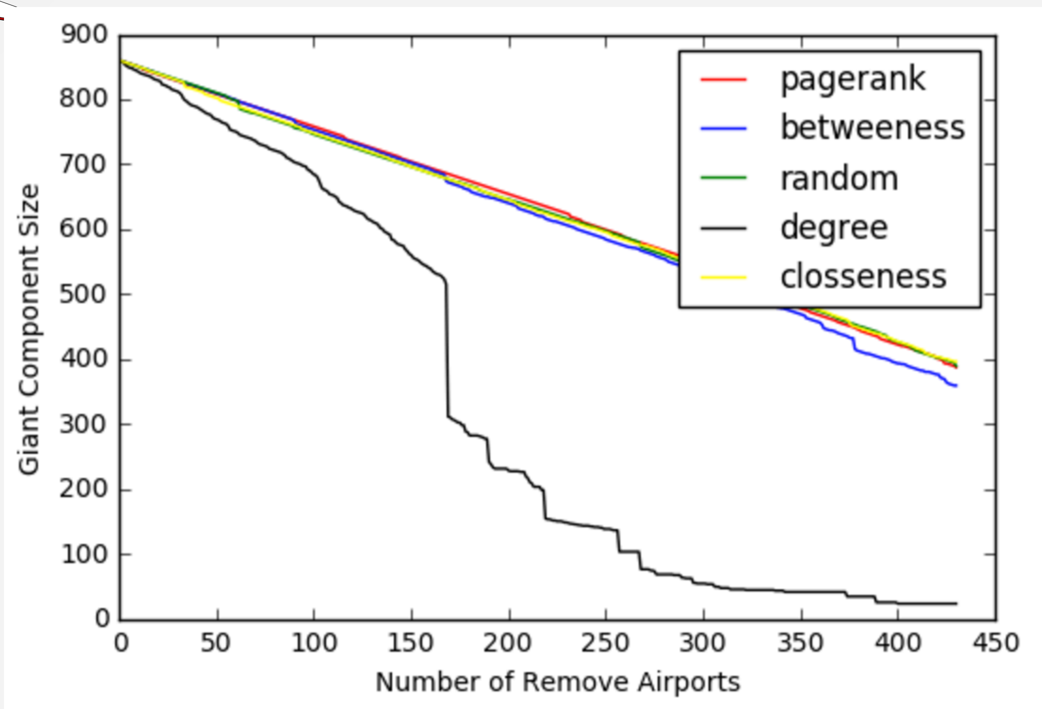
- Assortativity is not that strong, so that we can't say that small airport tend to connect to small and large tend to connect to large airports.

Transitivity

0.4539300395773018

- This is also the cluster coefficient, this value is not that big however since we have many small airports, the airports in US is well connected generally.

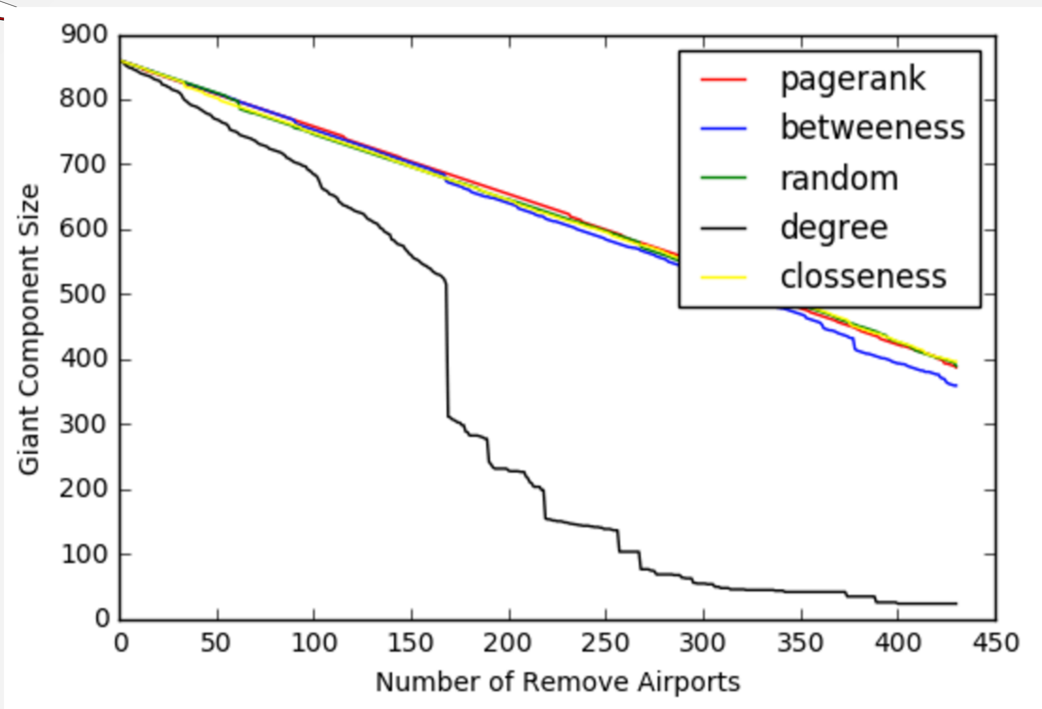
Resilience Test



First, we define a function to remove vertices gradually and record the giant components size change.

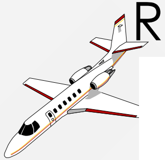
We decide to delete the half of the whole vertices. So that size is the number of airports we decide to attack

Resilience Test

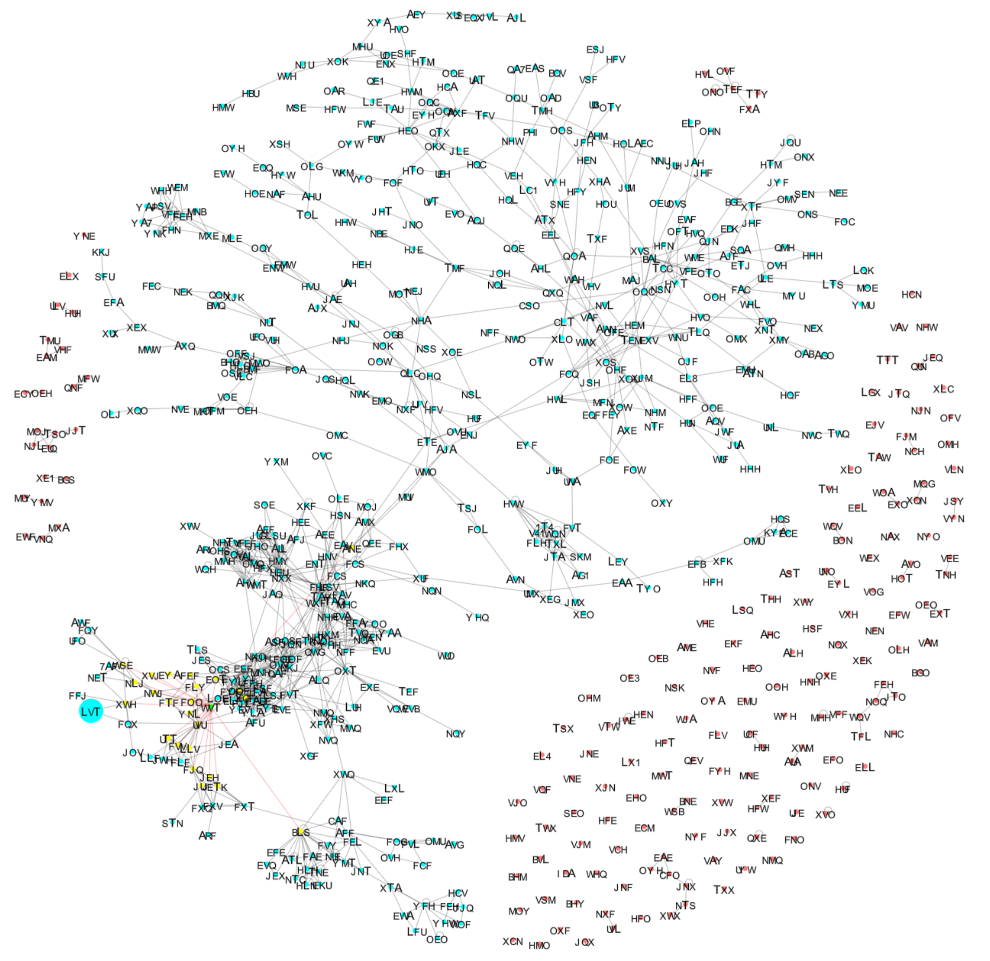


First, we define a function to remove vertices gradually and record the giant components size change.

We decide to delete the half of the whole vertices. So that size is the number of airports we decide to attack

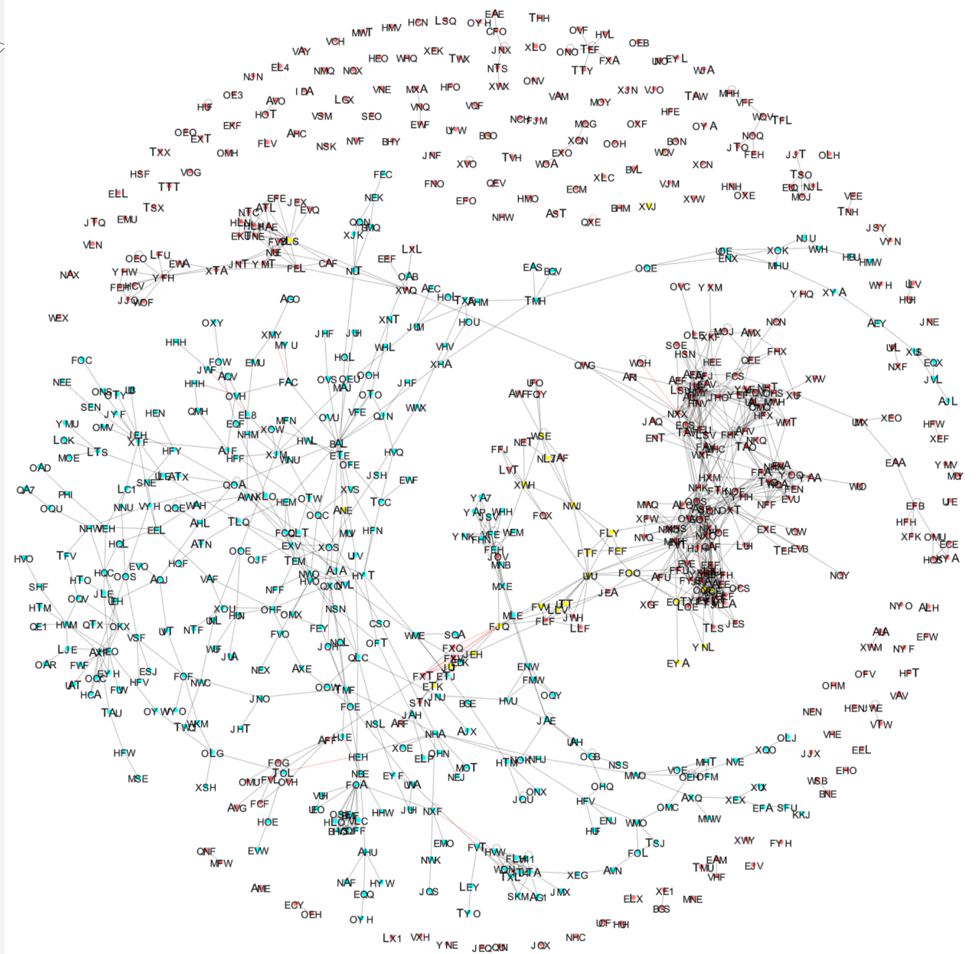


Resilience Test



We can see that yellow is the airport link to DLG the green point and blue points are belong to giant components.

Resilience Test



After we delete the DLG, we can see that the yellow points used belong to giant components fall into the non giant components group

4 AIRLINE NETWORK MODEL

AIRLINE MODEL TEST

Fitting airport degree as original distribution

Configuration Model

C

P(A,B)=1/1046

D

Devised Model

P(A,B)=Sigmoid(beta0+beta1*distance(A,B)+
beta2*degree(A) + beta3*degree(b) +
beta3*delta(A and B are in the same state))/lambda

*Where lambda is a normalizing factor allow probability sums to 1

AIRLINE MODEL TEST

Fitting airport degree as original distribution

P(A,B)=1/1046

Configuration Model

C

Observations

O

- This is a logistic regression model fitted from the training data
- The model estimates the posterior probability that an edge from A to B exists

P(A,B)=Sigmoid(beta0 +beta1*distance(A,B)+ beta2*degree(A)+ beta3*degree(b) + beta3*delta(A and B are in the same state))/lambda

D

Devised Model

O

Observations

- It uses airport degree, location as prior information.

AIRLINE MODEL TEST

Random Simulation Process

Simulation Process

Fix an origin Airport, add an out-edge from origin to 1046 other airports, repeat until reaches out degree, then move on the the next origin.

AIRLINE MODEL TEST

Random Simulation Process

Simulation Process

Fix an origin Airport, add an out-edge from origin to 1046 other airports, repeat until reaches out degree, then move on the the next origin.

Soft Decision

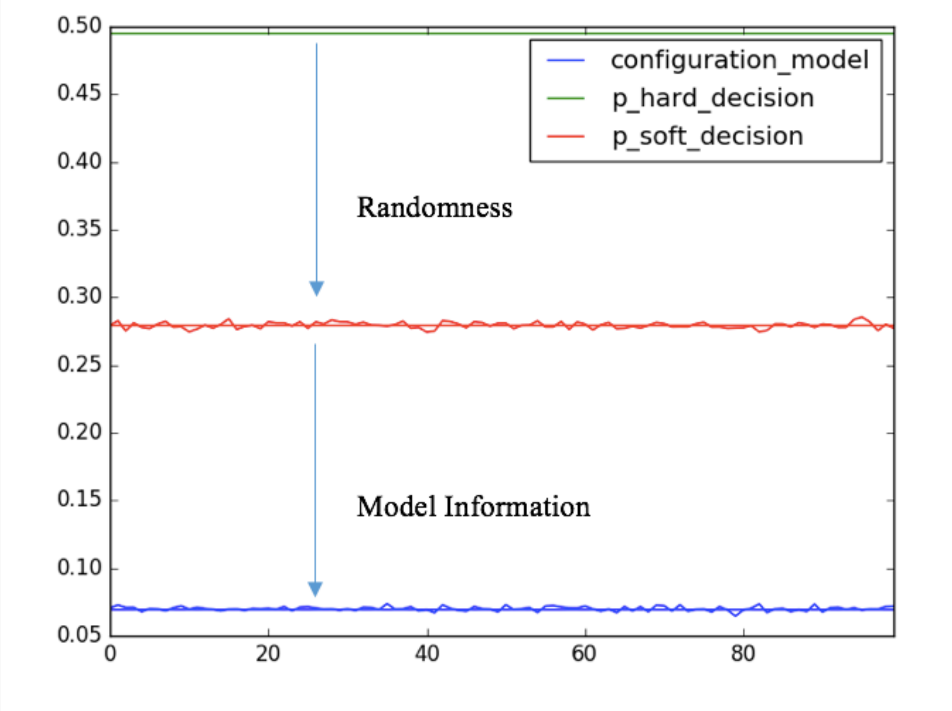
P-weighted random without replacement for a given origin (Introduce randomness given P(A, B))

Hard Decision

Select top probability airports as destination

AIRLINE MODEL TEST

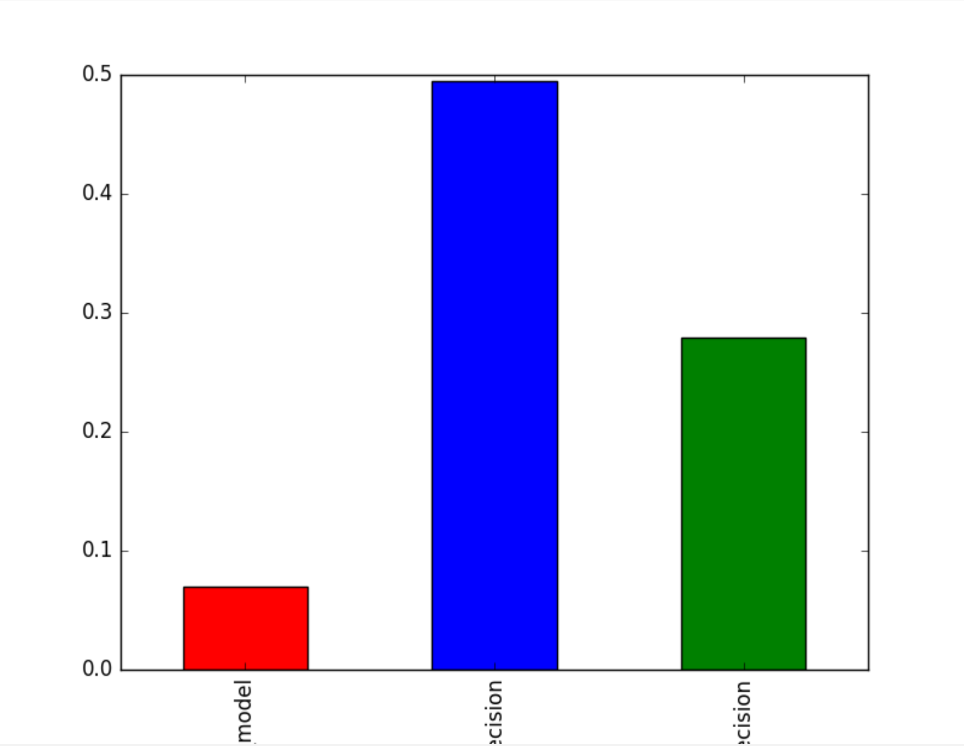
Random Simulation Process Evaluation



- Define proportion of the same edge between generated graph and true graph as absolute similarity.
- 100 simulations

AIRLINE MODEL TEST

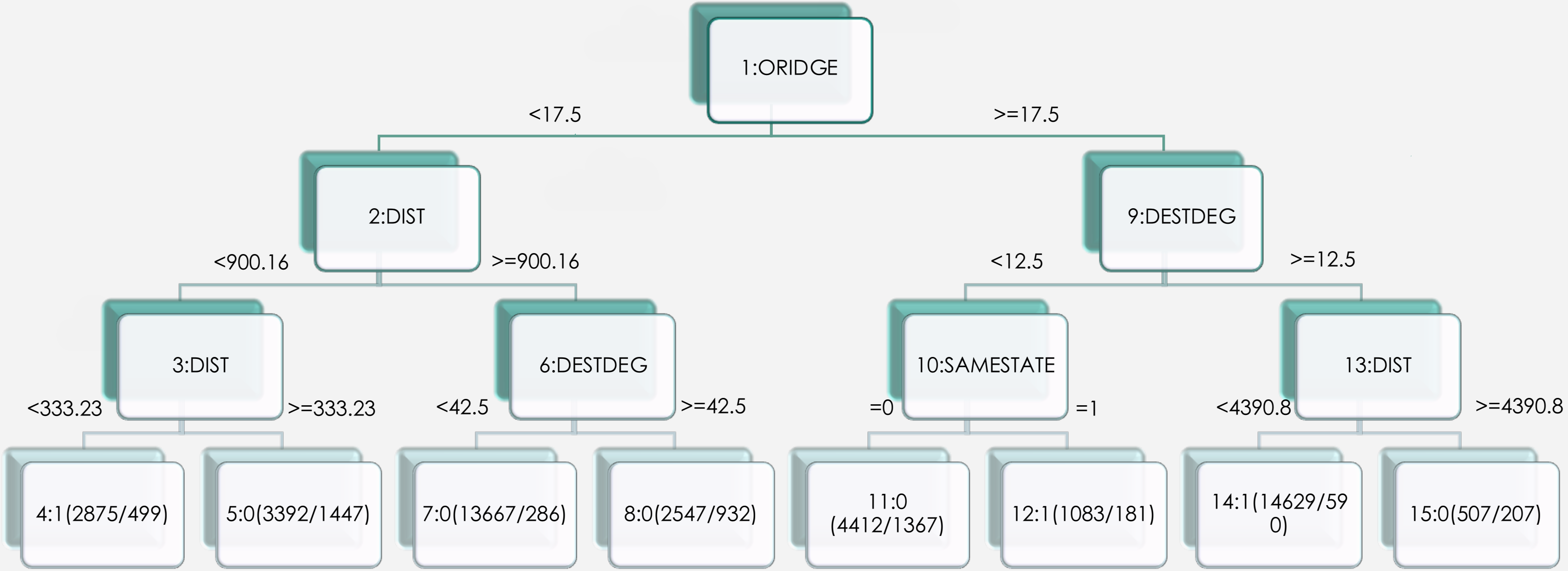
Random Simulation Process Evaluation



- Hard decision: 49% accuracy
- Soft random: 27% accuracy
- Configuration Model: 7% accuracy

MODEL INTERPRETATION

Decision Tree



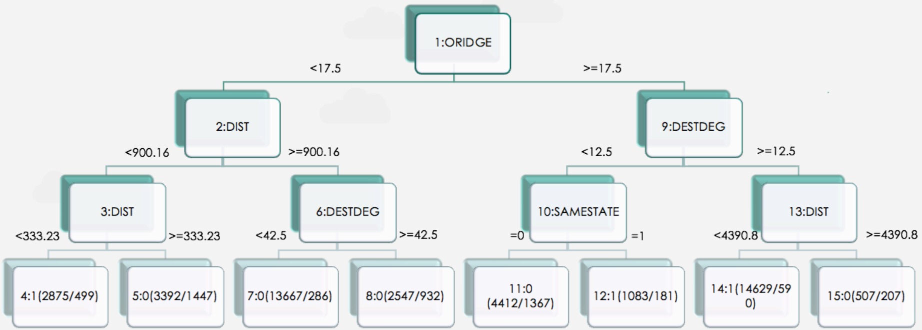
MODEL INTERPRETATION

Three Rules: That allow a light exist

Origin Degree
<17.5
↓
Distance(A,B)
<333km
↓
From small county airport to very close other airports

Origin Degree
>17.5O
↓
DestDegree
>12.5
↓
Distance(A, B)
< 4390km
↓
From big airports to other airports that within 4390km/ 6 hours fly

Origin Degree
>17.5
↓
DestDegree
<12.5
↓
A and B
in the same state
↓
From big airports to small county airports that are in the same state



CASE ANALYSIS

REFERENCES

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

QUESTIONS?

Go Ahead, Don't Hesitate!

THANK YOU!

Have a Nice Day!