CS-E5740 Complex Networks, Course Project

Thao Vy Le, Student number: 508379

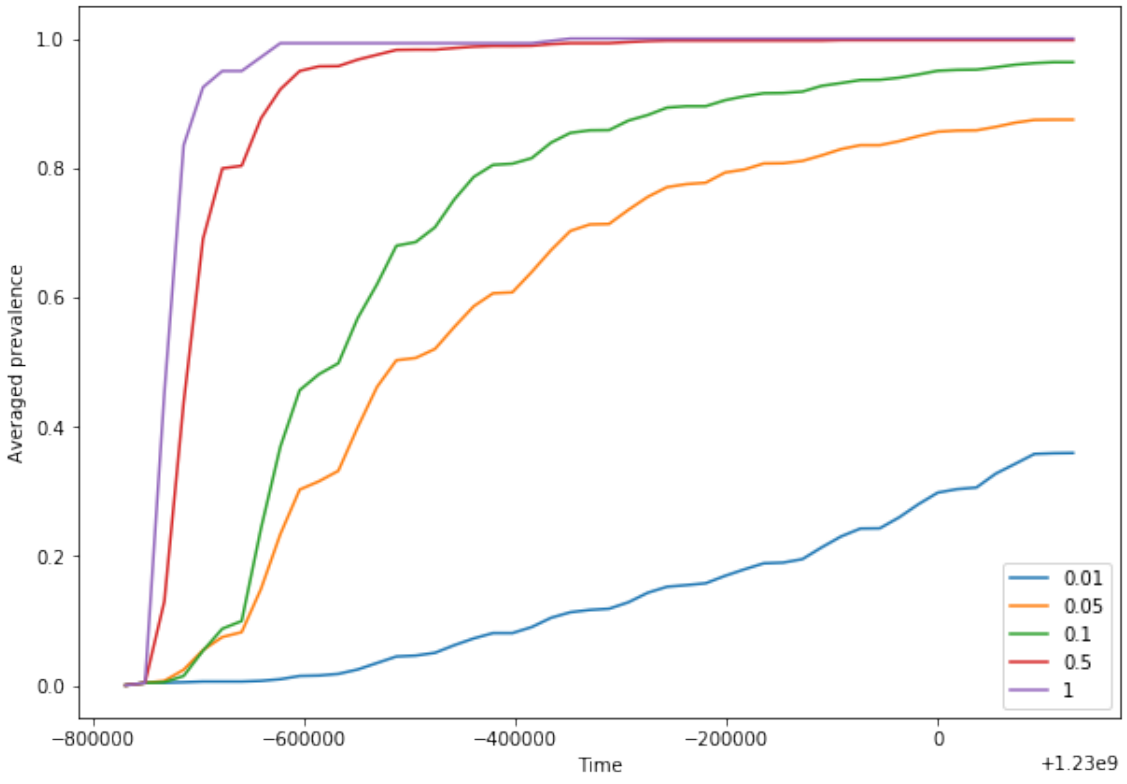
December 21, 2018

Task 1

a) If Allentown (node 0) is infected at the beginning of the data set, Anchorage becomes infected at time 1229290800.

Task 2

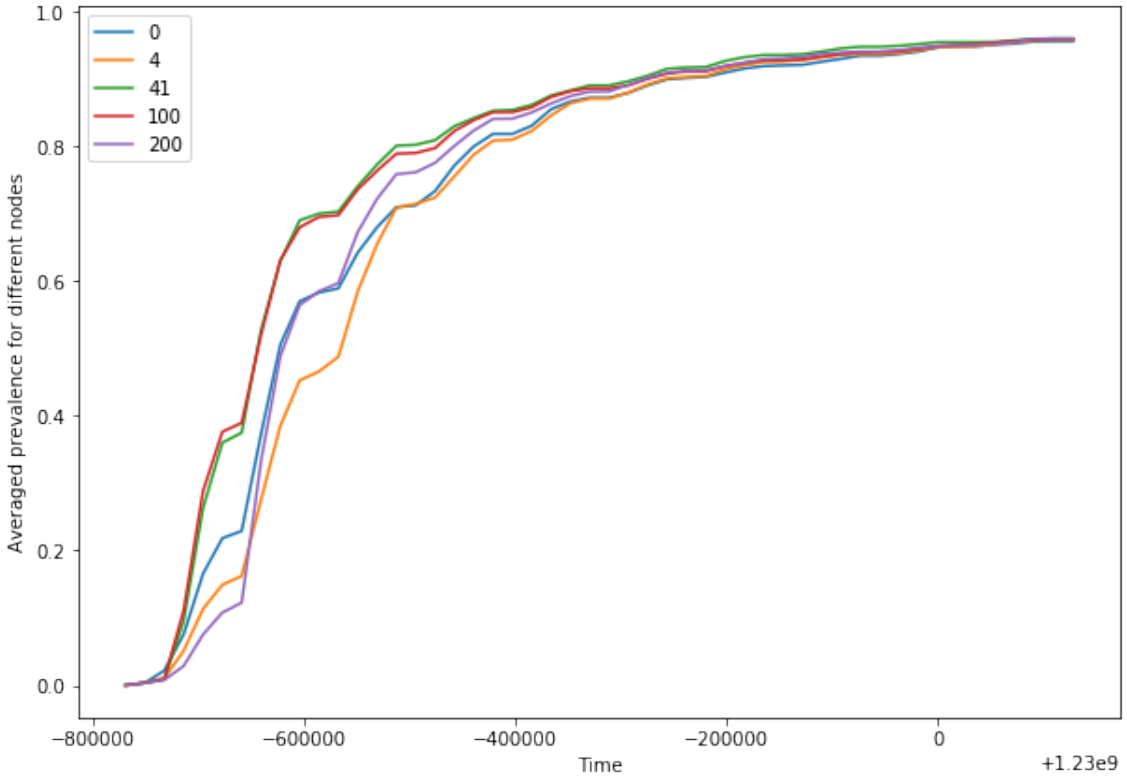
a) Plot the averaged prevalence p(t) of the disease as a function of time for each of the infection probabilities (0.01, 0.05, 0.1, 0.5, 1):



b) For probabilities between 0.5-1, the network becomes fully infected.
It's likely that the network has many small hub and once the disease reaches a hub, it spread quickly to the other nodes connecting to that hub. This explains the periodic steps in the graph.

Task 3

a) Averaged prevalence of the disease for each different node with p = 0.1 with different starting seed nodes 0, 4, 41, 100, 200 (averaged over 10 iterations):



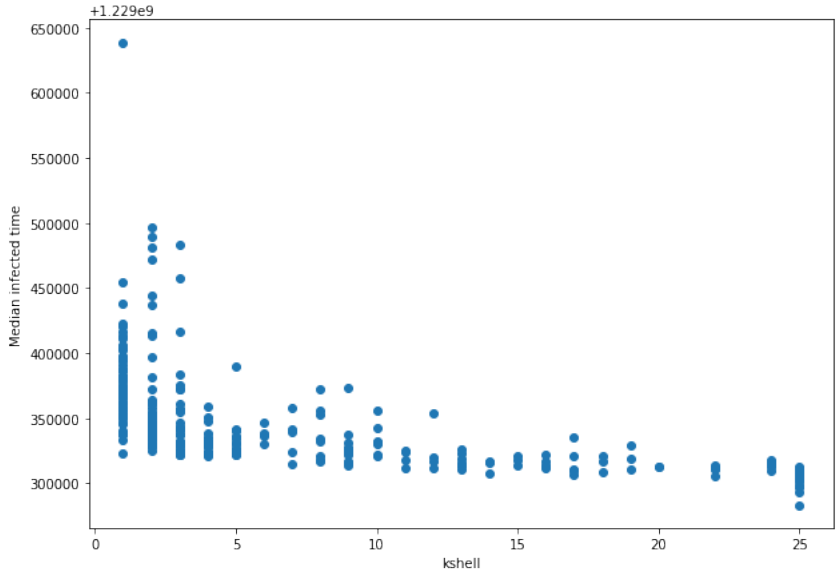
b) The spreading speed difference is visible in the beginning but not in the end. In the beginning, from different nodes, the disease can be either more difficult or easier to spread out to the hubs but once all the hubs are reached, there is no difference in spreading speed of the disease.

c) It is important to average the results over different seed nodes since spreading disease from different seed nodes can lead to significant differences in infection time of each node in the network. Therefore, to objectively assess the overall vulnerability of a node, it is important to average the results over simulating through different seed nodes.

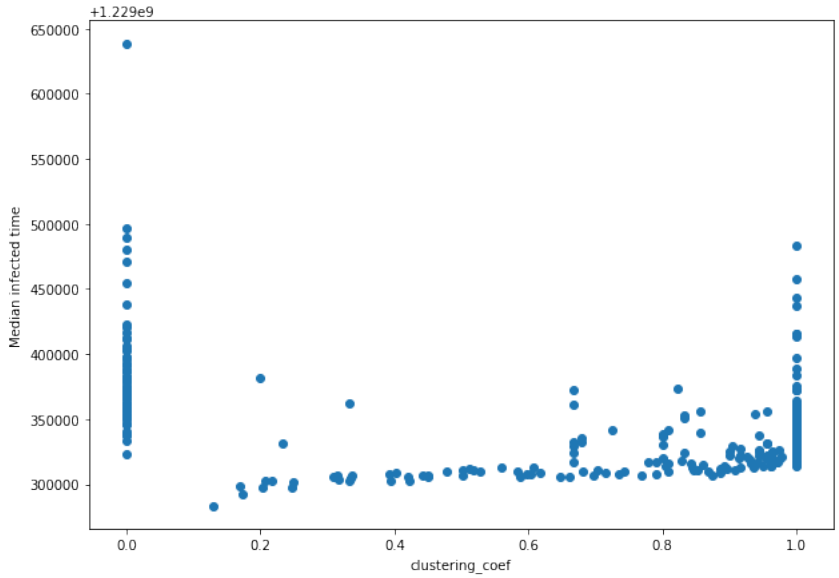
Task 4

a) Scatter plots showing the median infection time of each node as a function of nodal network measures:

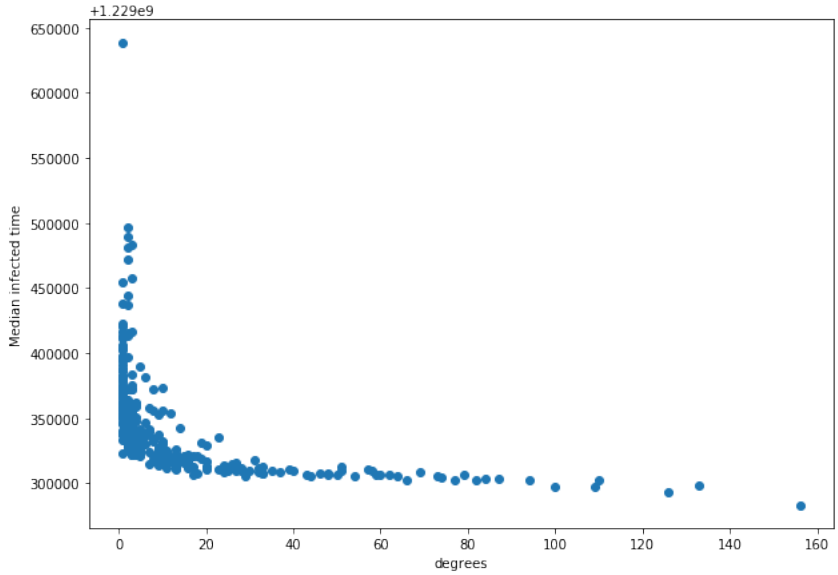
Median infection time vs. k-shell:



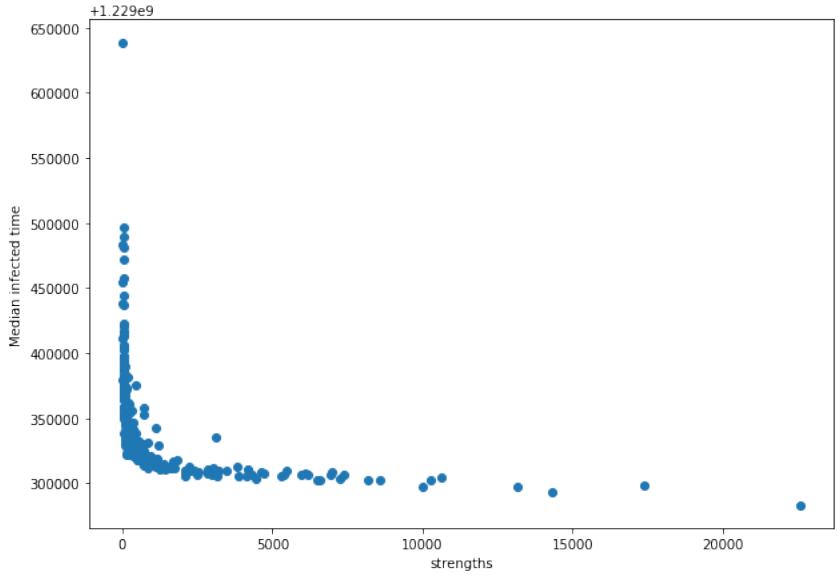
Median infection time vs. unweighted clustering coefficient:



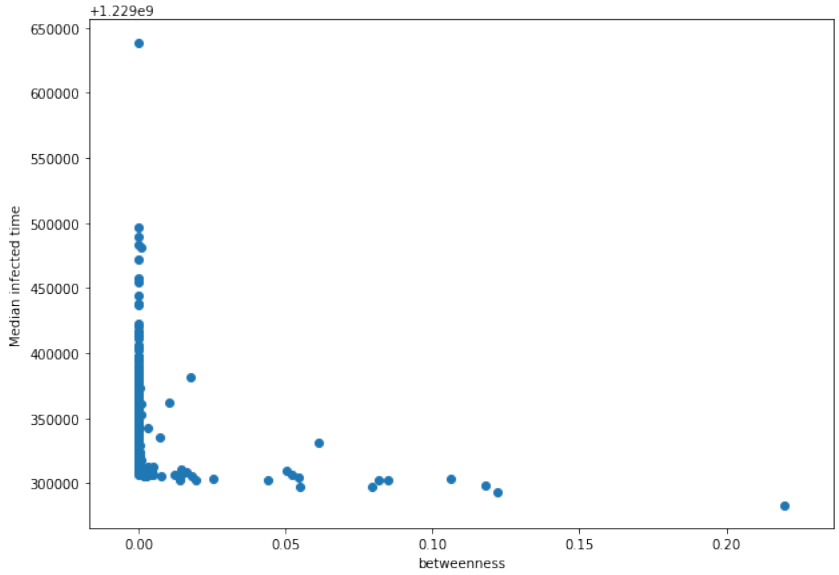
Median infection time vs. degree:



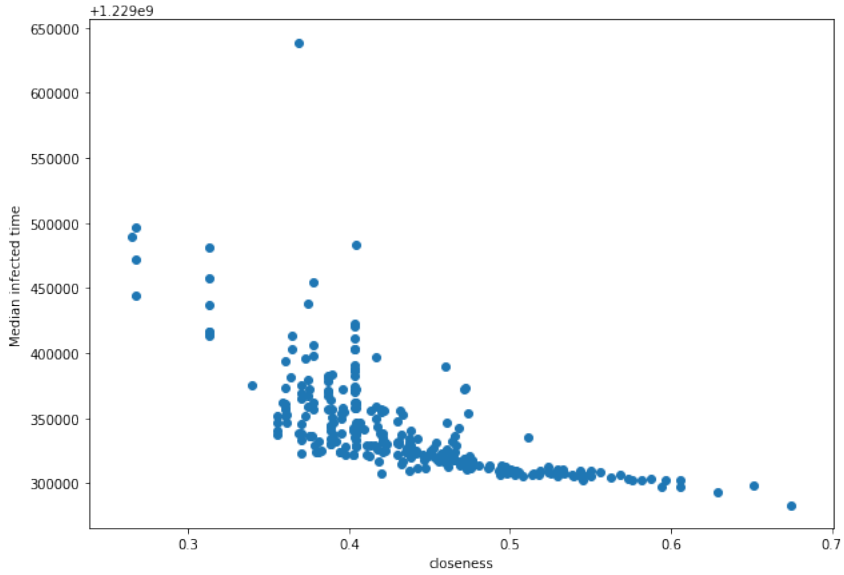
Median infection time vs. strength:



Median infection time vs. unweighted betweenness centrality:



Median infection time vs. closeness centrality:



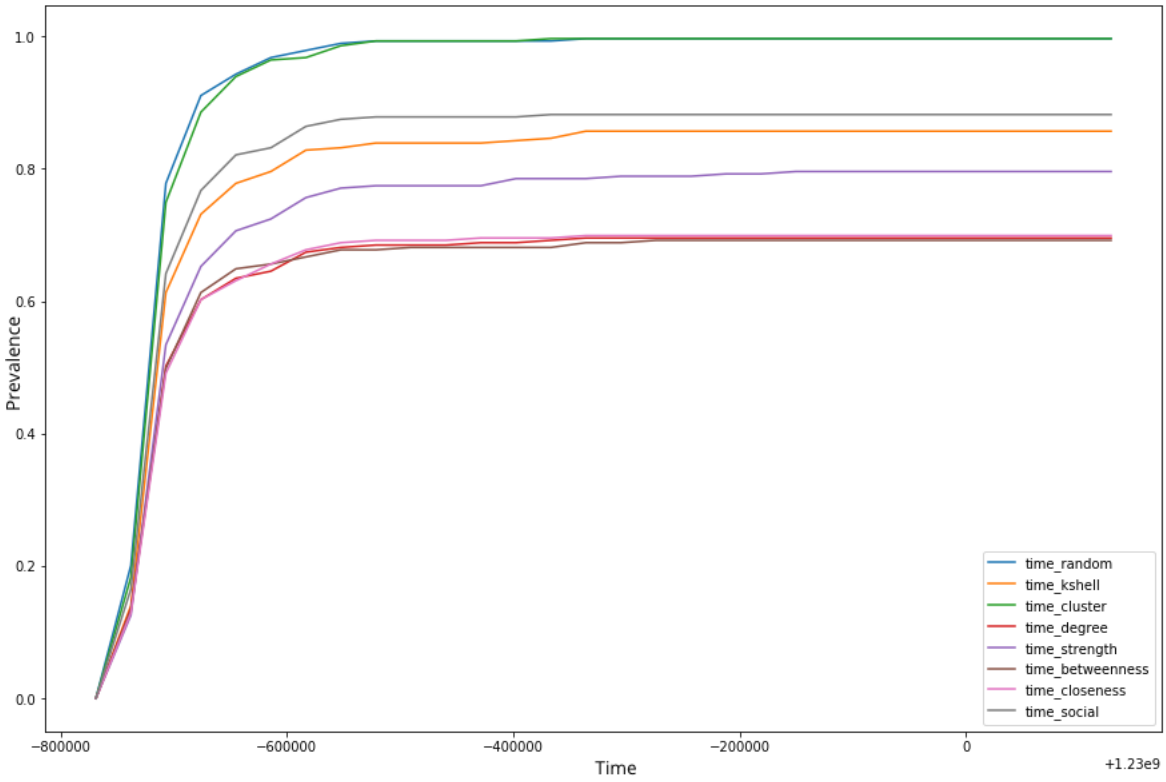
b) Spearman rank-correlation coefficient for each nodal network measures:
i) k-shell: correlation=-0.86, pvalue=6.37e-83
ii) unweighted clustering coefficient: correlation=-0.11, pvalue=0.06
iii) degree k: correlation=-0.922, pvalue=2.06e-116
iv) strength s: correlation=-0.857, pvalue=1.44e-81
v) unweighted betweenness centrality: correlation=-0.697, pvalue=5.65e-42
vi) closeness centrality: correlation=-0.823 pvalue=7.15e-70

Degree k has the highest absolute correlation so it is the best predictor of infection time.

c) Unweighted clustering coefficient has the lowest absolute correlation with the median infection time. This is likely the case since unweighted clustering coefficient is a measuring the connectivity of neighbor nodes with each other, while infection time is affected by links to/from hubs - something that unweighted clustering coefficient does not indicate clearly.
Unweighted betweenness centrality performs better than unweighted clustering coefficient but still rank low in the list. In my opinion, the main reason is that link weights are not taken into account in this measurement while it does matters in the models: places that have more flights back and forth tend to have more probability to be infected and spread out the disease. Besides, betweenness centrality places higher value in nodes that serve as bridge in shortest paths while disease spreading model is more important for hubs that have high connectivity with other airports (not the same definition of centrality that we are looking for here).
Closeness centrality, k-shell and strength measurements are also a good indicator of infection time with high absolute correlation (< -0.8) since they take into account the connectivity of the node with other nodes in the network. However, degree k is the strongest indicator since it indicates exactly how many connections each node have with other nodes.

Task 5

a) Prevalence of the disease as a function of time of 8 different immunized strategies:



b) Random node and unweighted clustering coefficient performed the worst since these strategies do not immunize nodes that are highly connected with other nodes in the network. Thus, the diseases can spread out to the whole network.

On the other hand, measurements like node degree, closeness centrality, strength and kshell seems to align with the result in Task 4 - which is a good indicator of node connectivity and also perform well in preventing diseases from spreading out to the network.

Betweenness centrality seems to be the odd one here since it was not a good indicator of infected time, but the strategy turns out to be effective. This might be because when protecting the "bridge" node, disease cannot spread out to the other side of the node.

c) Let p_k be the probability of nodes having degree k in any network => p_k follows Binomial distribution:

$$p_k = \binom{N}{k} p^k (1-p)^{N-k}$$

The average degree of random immunization strategy:

$< k >= \sum_k k * p_k$

The average degree of social net immunization strategy is the expected value of 2nd moment:

$< k_{nn} >=< k^2 > / < k >$

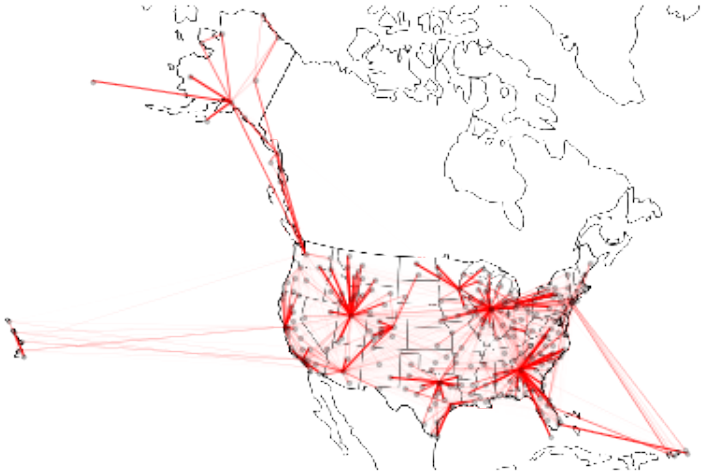
We have:
$< k^2 >= \sum_k k^2 * p_k$
$< k >^2= \sum_k k * p_k$

As $k >= 1$ and $p_k <= 1$ with all k, $< k^2 >>< k >^2$, we can conclude $< k_{nn} >>< k >$, social net immunization strategy has higher probabilities to pick high-degree node.

d) Although social network immunization strategy is not effective as other strategies, this strategy allows for faster runtime. A social network can become very large and complex, thus running through the whole network to calculate k-degree, k-shell or other nodal measurements would take up significant amount of time or even become an impossible task. Social network strategy, however, does not require to compute through the whole network but only a random group of nodes is needed for selecting their neighbors. Therefore, this strategy proves to be more effective in realtime and for big and complex network.

Task 6

a) Visualizing disease transmit links:



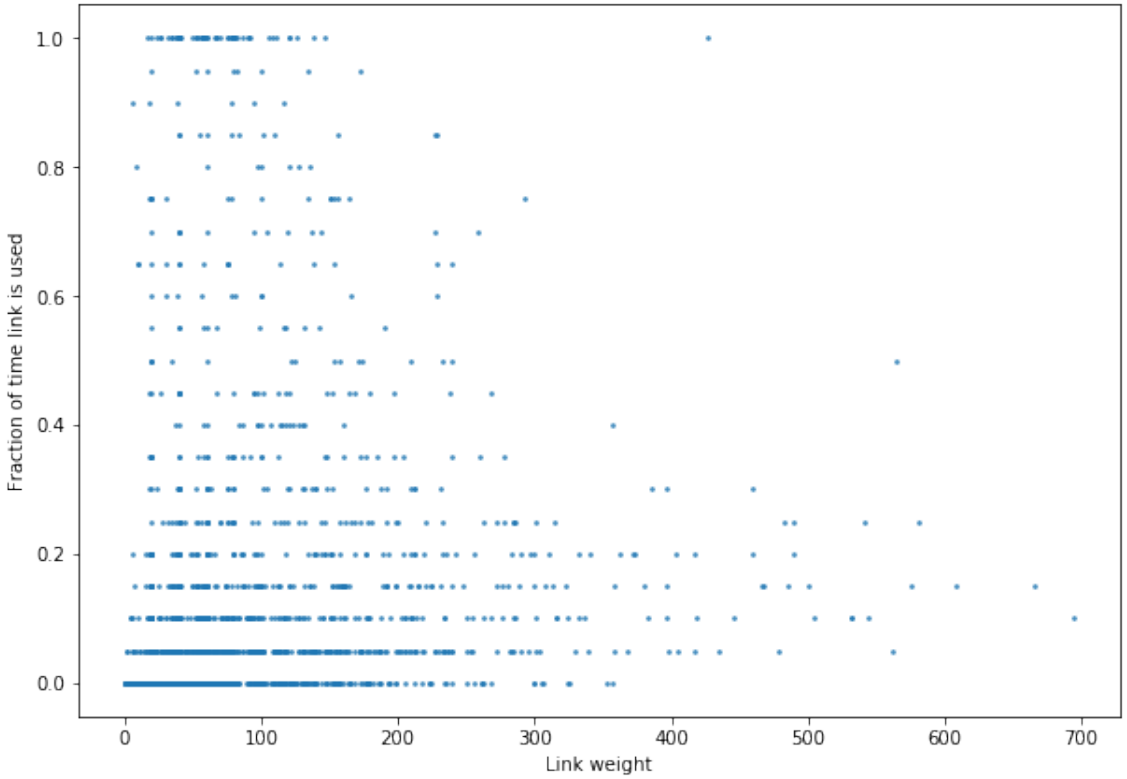
Maximal spanning trees:

The maximal spanning tree network has many similarities to the disease transmitting links visualization. This might be explained by the fact that a higher link weight means more flights between the nodes and thus correlates to higher probabilities of disease transmitting.

b) Also, looking into the visualization of disease transmitting links, the boldest links seems to cross each other at only a few points in the map. These are the central hubs with many flights connect to them. Perhaps trying to immunize those points is the most effective methods to prevent the diseases from spreading.

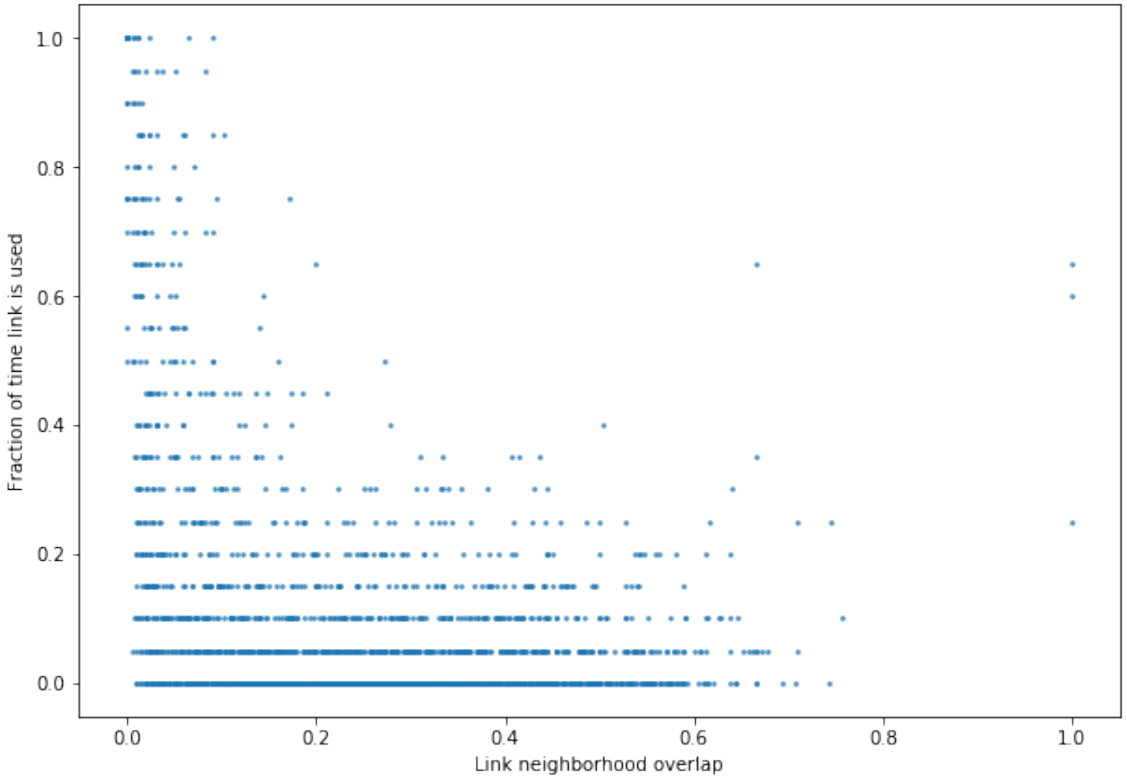
c) Scatter plots:

f_{ij} and link weight:



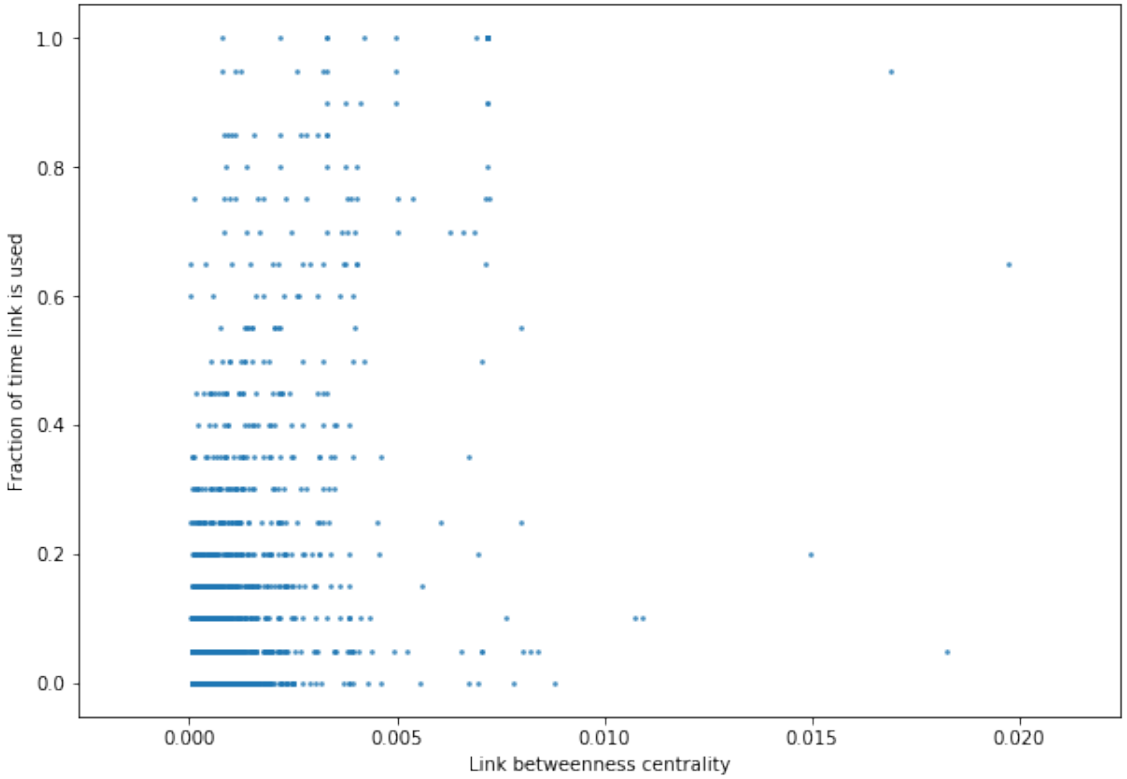
Spearman coefficients between f_{ij} and link weight : correlation=0.336, pvalue=2.44e-56)

f_{ij} and link neighborhood overlap:



Spearman coefficients between f_{ij} and link neighborhood overlap : correlation=-0.373, pvalue=6.85e-70)

f_{ij} and link betweenness centrality:



Spearman coefficients between f_{ij} and link betweenness centrality : correlation=0.44, pvalue=2.87e-100)

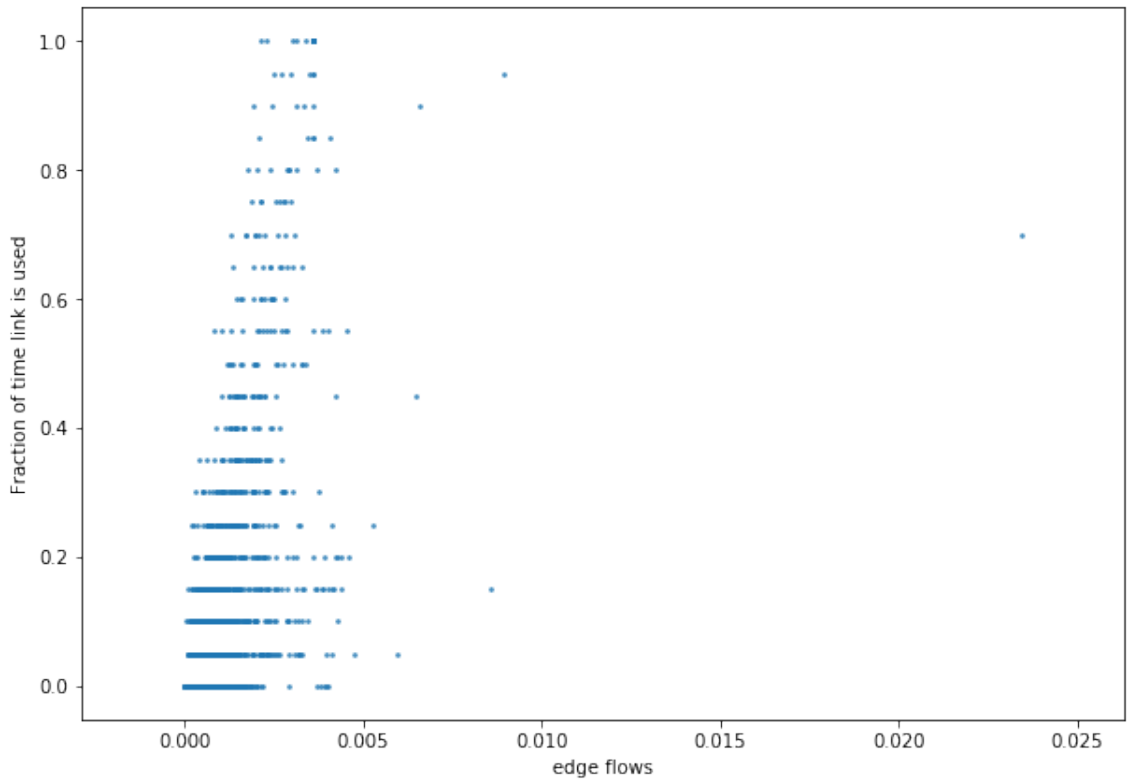
d) As discussed in earlier tasks, nodes that have the most connectivity to other nodes tend to be the "hubs" that spread the diseases. Therefore, links that connect to those hubs tend to be used more to transmit diseases. All of the link measurements above, however, does not indicate the degree of connectivity to hubs, thus all of those measurements have low correlation with f_{ij}.

Bonus Task 1

Looking at literature, there is a link measurement called edge current flow betweenness centrality, or random walk betweenness. This definition of betweenness not only taking into account shortest paths in the network but also other longer paths which might play an important role in connecting the information flow (in this case, spreading diseases) within the network. This is a relaxed version of betweenness centrality while shorter paths are given more weights than longer paths.

In this exercise, the weighted version of edge current flow betweenness centrality is

used.



Spearman coef of weighted edge current flow betweenness centrality : (correlation=0.675, pvalue=1.156e-277)

Task 7

Four ways how the model could be improved to be more realistic:
- The model is not currently allowed for recovery: an infected node remains infected forever or an immunized node remains immunized forever, which normally is not realistic. The model can be modified to accommodate a fixed time frame for a node being infected and then recovering after that.
- The model assumes the probability of being infected is the same in all case, which in reality is rarely true. This can be changed so that the probability of infecting diseases follows a distribution with certain parameters (e.g. normal distribution, poisson distribution, etc. which is more suitable), thus allowing for some degree of variations in probability of disease transmission in different places.
- Another thing this model is not taking into account is assuming airports are the same. However, in reality, airports with higher density of infected individuals tend to spread out disease faster. In order to account for this complexity, the model needs to be able to model disease spreading between individuals and simulate the differences in airport populations.
- Last but not least, the network and flight schedule is assumed to be static. In reality, it's

unlikely that there will be no changes made to adapt to the situation when an epidemic is spreading. Human behaviors shall change as a reaction/response to events happening, for example, less people wants to travel through places where diseases are spreading, and subsequently infected nodes might become less connected than earlier, reducing the chances of transmitting diseases. This dynamic could be taken into account on this model in different ways, for example, by modeling probability of flight cancellation for flights scheduled from/to infected sources.