CS-E5740 Complex Networks, Final Project

Yangzhe Kong, Student number: 765756

December 17, 2019

Contents

Task 1: Basic implementation **2**

Task 2: Effect of infection probability p on spreading speed **2**

Task 3: Effect of seed node selection on spreading speed **3**

Task 4: Effect of seed node selection on spreading speed **4**

Task 5: Shutting down airports **7**

Task 6: Disease transmitting links **9**

Task 7: Discussion **11**

Task 1: Basic implementation

a) If Allentown (node-id=0) is infected at the beginning of the data set, at which time does Anchorage (ANC, node-id=41) become infected?

Answer: Node 41 is infected at time:1229290800

Task 2: Effect of infection probability p on spreading speed

a) Plot the averaged prevalence $\rho(t)$ of the disease (fraction of infected nodes) as a function of time for each of the infection probabilities. Plot the 5 curves in one graph.

Answer: The plot is shown in Figure 1

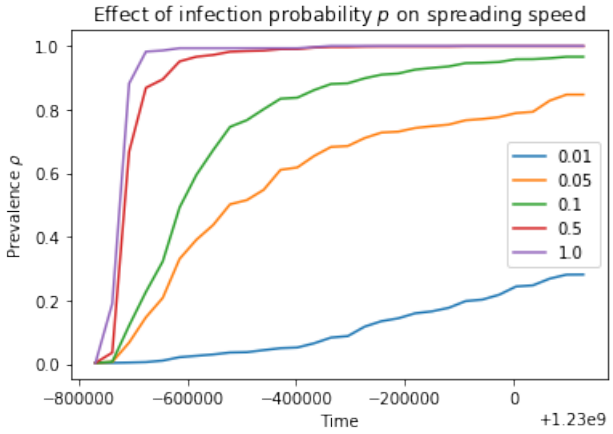

Figure 1: The averaged prevalence $\rho(t)$ of the disease (fraction of infected nodes) as a function of time for each of the infection probabilities

b) For which infection probabilities does the whole network become fully infected? What are the stepwise, nearly periodic "steps" in the curves due to?

Answer: Interestingly, only when p=1 will the whole network become fully infected. These nearly periodic steps may be due to the infection of a hub. If a hub is being infected, then the spreading speed would be higher, because typically a hub has more neighbors.

Task 3: Effect of seed node selection on spreading speed

a) Use nodes with node-ids [0, 4, 41, 100, 200] (ABE, ATL, ACN, HSV, DBQ) as seeds and p = 0.1, and run the simulation 10 times for each seed node. Then, plot the average prevalence of the disease separately for each seed node as a function of time.

Answer: The plot is shown in Figure 2

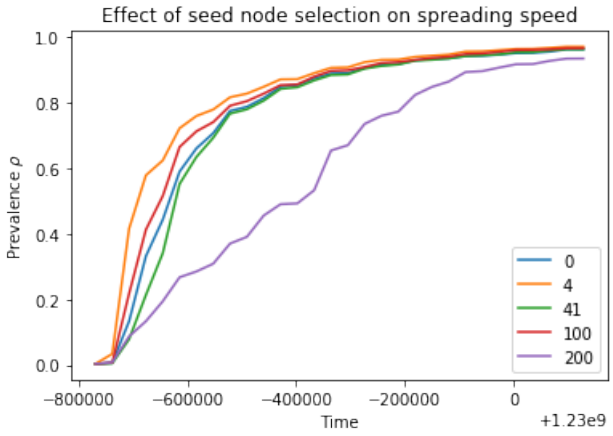

Figure 2: The average prevalence of the disease separately for each seed node as a function of time

b) You should be able to see differences in these spreading speed. Are these differences visible in the beginning of the epidemic or only later on? Why?

Answer: These differences are both visible in the beginning of the epidemic and later on. And actually they are more clear in the beginning since later on in all cases except for the case of using node 200 as seed node the prevalence all converge to somewhere around 0.95. A potential explanation is that seed nodes are located in component that have different connectivity measures, so in the beginning the spreading speeds will be quite different according to local vulnerability, but later on as the infection spreads to well-connected components, the spreading speeds would be similar.

c) In the next tasks, we will, amongst other things, inspect the vulnerability of a node for becoming infected with respect to various network centrality measures. Why is it important to average the results over different seed nodes?

Answer: As we see in b), selection of different seed nodes can affect a lot the spreading speed,it's reasonable to average their effects and reduce the variance. Also in the sense to control the variates that we are not interested in, it's important to average the results over different seed node.

Task 4: Effect of seed node selection on spreading speed

a) Run the 50 simulations, and create scatter plots showing the median infection time of each node as a function of the following nodal network measures:
i) unweighted clustering coefficient c
ii) degree k
iii) strength s
iv) unweighted betweenness centrality

Answer: The plot is shown in Figure 3, 4, 5, 6

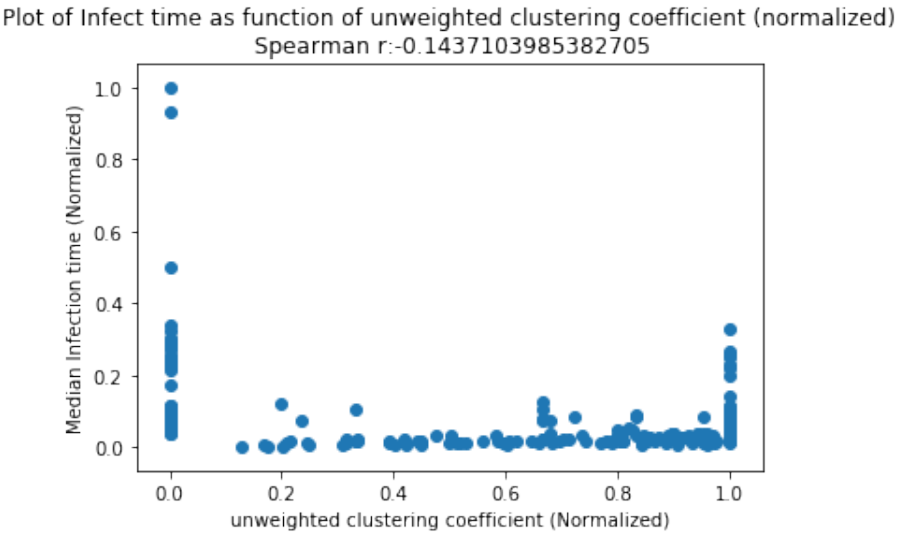

Figure 3: The median infection time of each node as a function of unweighted clustering coefficient

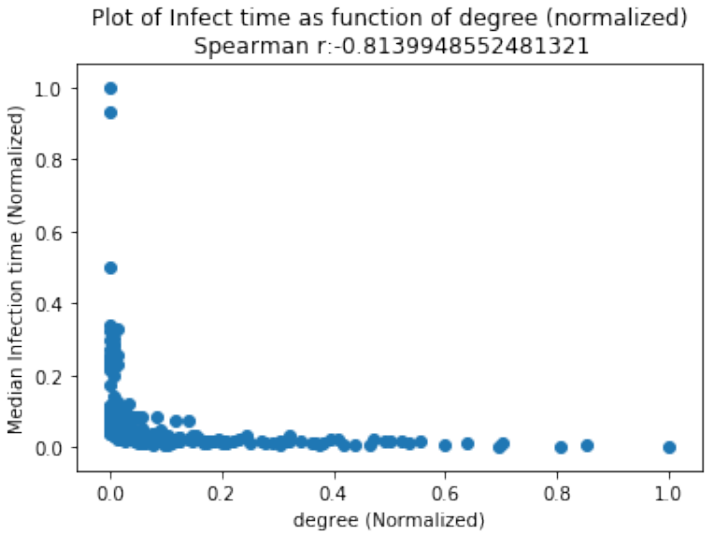

Figure 4: The median infection time of each node as a function of degree

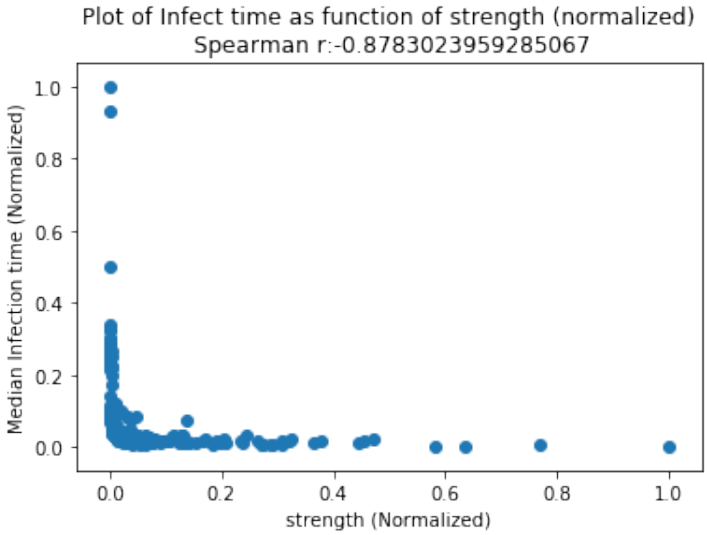

Figure 5: The median infection time of each node as a function of strength

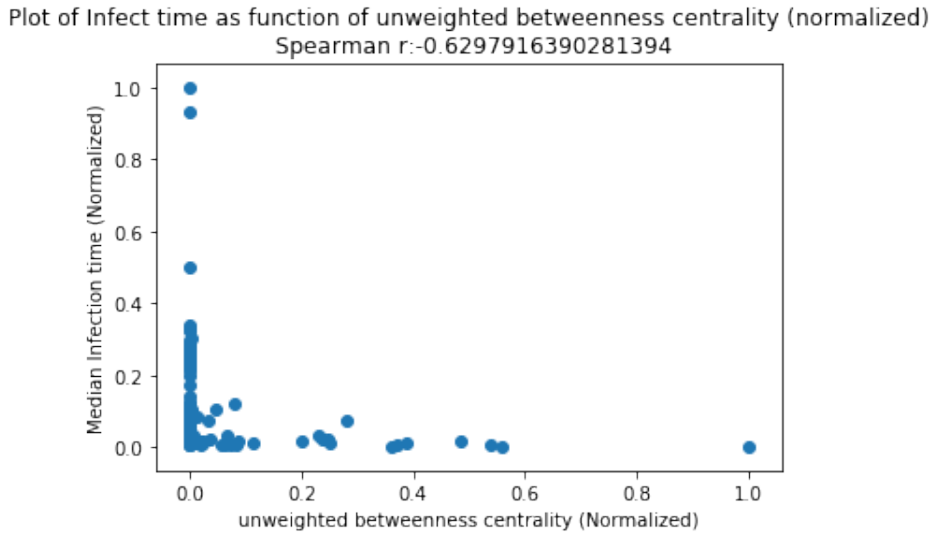

Figure 6: The median infection time of each node as a function of unweighted betweenness centrality

b) Use the Spearman rank-correlation coefficient for finding out, which of the measures is the best predictor for the infection times.

Answer: The results are shown below:
Spearman rank-correlation coefficient for
unweighted clustering coefficient: -0.1437103985382705
degree: -0.8139948552481321
strength: -0.8783023959285067
unweighted betweenness centrality: -0.6297916390281394

We can see that strength (rho: -0.878) and degree (rho: -0.814) seems to be better predictors for infection times

c) Discuss your results for each network centrality metric. Especially, explain the ranking of the network measures as measured by the median infection time.

Answer: **strength**($\rho = -0.878$) is a sum of weights of edges adjacent to the edge. Generally when an edge has a high strength value, it means that this edge is either in a sub-component with high connectivity or the entry path to a sub-component. Thus, in this case nodes connected with edges of high strength would certainly be vulnerable to infection.

degree($\rho = -0.814$) is a very similar measure but it is not weighted. It does not contain so much information as strength and therefore slightly worse than strength.

unweighted cluster coefficient ($\rho = -0.144$) is the worst because it doesn't provide too much information about the node and its neighbors as the previous ones

do.

unweighted betweenness centrality($\rho = -0.630$) is a measure of centrality in a graph based on shortest paths. We can see that this measure is not very good at predicting infection time. The potential explanation is that the best representation of the path from one airport to another is likely to be a random walk, it has not necessarily to follow the shortest path to a target airport. Thus, unweighted betweenness centrality cannot be the best one to predict infection time.

Task 5: Shutting down airports

a) Adapt your code to enable immunization of nodes, and plot the prevalence of the disease as a function of time for the 8 different immunization strategies (social net., random node, and 6 nodal network measures) when 10 nodes are always immunized

Answer: The plot is shown in Figure 7

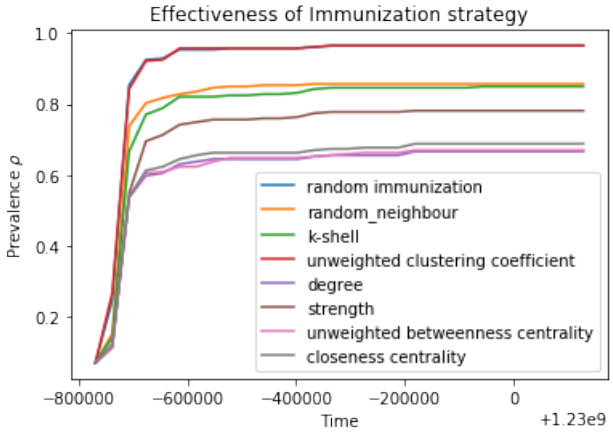

Figure 7: The prevalence of the disease as a function of time for the 6 different immunization strategies

b) Discuss the ranking of the immunization strategies. In particular, compare your immunization results with the results you obtained in the previous task (Task 4). Are there some measures that are bad at predicting the infection time but important with regards to immunization? Or vice versa? Why?

Answer: The results are shown in the following table and Figure 8

	Infection ratio	Spearman rank-correlation coefficient
unweighted betweenness centrality	0.67	-0.63
closeness centrality	0.688	-0.852
strength	0.781	-0.878
k-shell	0.849	-0.825
random_neighbour	0.857	N/A
random immunization	0.964	N/A
unweighted clustering coefficient	0.964	-0.144
degree	0.667	-0.814

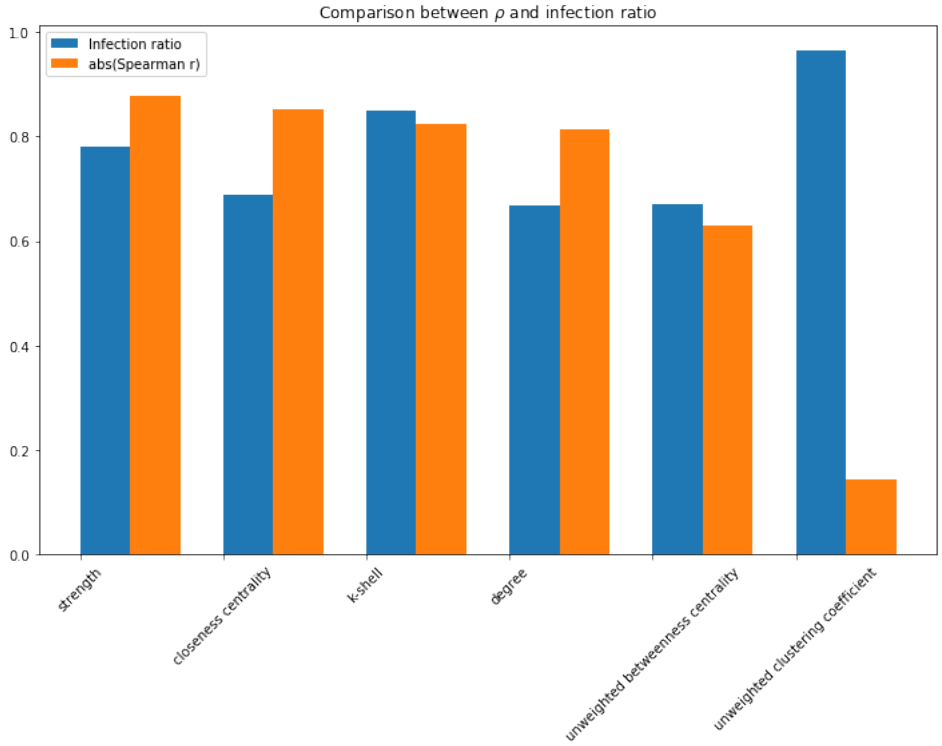

Figure 8: Comparison between ρ and infection ratio

strength, closeness centrality and degree are the ones that both do good in predicting infection time and immunization. We've already discussed strength and degree in part c) of task 4.

unweighted betweenness centrality is the one that do bad in predicting infection time but do well in immunization. Although we've said that the airline traffic flows do not necessarily follow the shortest paths, but measure based on shortest paths like closeness centrality and unweighted betweenness centrality can still be effective in immunization because the node with large closeness centrality and unweighted

betweenness centrality can be viewed as the one that has more control over the network, because more information will pass through that node. In this sense that node may be the gateway to a large sub-component and thus closing that node can help immunization.

Interestingly, k-shell is good at predicting the infection time but not so important with regards to immunization. The reason may be that k-shell is a very good measure of centrality, but infection will not be affected too much if we close down the central node. As we just discussed above the key to immunization is to close down gateway nodes.

c) The pick-a-neighbour immunization strategy probably worked better than the random node immunization. Let us try to understand why.
- First, if the degree distribution of the network is P(k), what is the probability of picking a random node of degree k?
- What is the expected outcome if you then pick a random neighbour of the random node (hint: see lectures 3 and 4)?
- Consequently, which of the strategies is expected to be more effective and why?

Answer: First, the probability of picking a random node with degree k is exactly $P(k)$. The probability of pick a random neighbour with degree k of the random node is $P(k) * \sum_k kP(k)$.

Since $\sum_k P(k) = 1$ and $\sum_k kP(k) > \sum_k P(k)$, we have $P(k) * \sum_k kP(k) > P(k)$. Thus, social network immunization is better since we increase the chances of choosing a node with a high degree.

d) Although the social network immunization strategy outperforms the random immunization, it is not necessarily as effective as some other immunization strategies (and there is random variation). Nevertheless, explain shortly, why it still makes sense to use this strategy in the context of social networks?

Answer: The main reason is its easy implementation because it requires no further information about the node.

Task 6: Disease transmitting links

a) Run the simulations, and compute the fraction of times that each link is used for infecting the disease (f_{ij}). Then use the provided function plot_network_USA which can be found in si_animator.py to visualize the network on top of the US map (see the example code given in the function). Adjust the width of the links according to the fractions f_{ij} to better see the overall structure. Compare your visualization with the maximal spanning tree of the network.

Answer: The results is shown in Figure 9 and Figure 10.

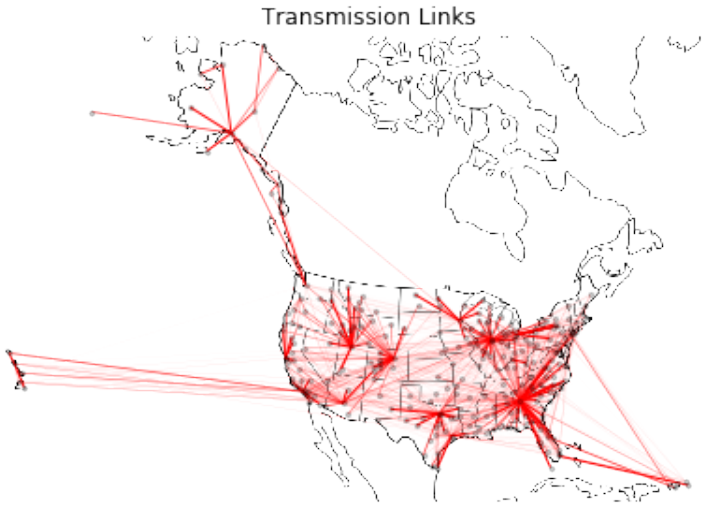

Figure 9: The network with line width in accordance wit f_{ij}

Figure 10: Maximal spanning tree of the network

b) What do you notice? How would you explain your finding?

Answer: We can notice that our plot and the maximal spanning tree are quite similar. A reasonable explanation is that in maximal spanning tree edges with large weights are selected, and because edges with larger weights are used more than the others, which leads to high probability to spread the infection for more times. And this is exactly what we are trying to calculate and plot using our simulation.

c) Create scatter plots showing f_{ij} a function of the following link properties:
i) link weight w_{ij}
iii) unweighted link betweenness centrality eb_{ij}
Compute also the Spearman correlation coefficients between fij and the two link-wise measures.

Answer: The Spearman correlation coefficients between fij and weight: 0.4054752528698251
The Spearman correlation coefficients between fij and link betweenness centrality: 0.5563146037032359

d) Explain the performance of the two link properties for predicting f_{ij}

Answer: Both 2 measures are not good enough in predicting f_{ij}. Link betweenness has relatively better performance because it is the sum of the fraction of shortest paths that pass through the node considering all pairs of nodes.

Task 7: Discussion

Even though extremely simplistic, our SI model could readily give some insights on the spreading of epidemics. Nevertheless, the model is far from an accurate real-world estimate for epidemic spreading.

Discuss the deficiencies of the current epidemic model by listing at least four (4) ways how it could be improved to be more realistic.

Answer: We can add a new state for nodes: removed (R) - immune from the pathogen (via vaccine or post-exposure) or dead, to change SI model into a SIR Epidemic Model to make it more realistic.

Similarly, we can make the infected individuals return to the susceptible state after getting infection with some probability. This model is called SIS model and it's appropriate for diseases such as common cold (rhinoviruses) that have chronic infections.

To make the model even more realistic, we can construct the SIRS model, which means that an airport can be susceptible again after getting immune. For example, an individual's immunity can wane over time even after recovering from seasonal influenza. The SIRS model can be used in this case to allow immune individuals to become susceptible again.

For some diseases, they have a latent phase in which the individual is infected but not yet infectious. We can add a latent/exposed (E) state in the SIR model and let infected (but not yet infectious) individuals instead of directly from S to I move from S to E and from E to I in order to incorporate this delay between getting infected (E) and really being

infectious (I).

Acknowledge

The ideas to improve the SI model are mainly from [1] and [2]. In addition, [3] is taken as a reference for the code of data visualization.

References

[1] "Welcome to EMOD modeling for general disease — Generic Model documentation", Institutefordiseasemodeling.github.io, 2019. [Online]. Available: https://institutefordiseasemodeling.github.io/Documentation/general/index.html. [Accessed: 17- Dec- 2019].

[2] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, and A. Vespignani, "Epidemic processes in complex networks," arXiv preprint arXiv:1408.2701, 2014

[3] adamilyas,"adamilyas/complex-networks", GitHub, 2019. [Online]. Available: https://github.com/adamilyas/complex-networks/blob/master/project/code/project.ipynb. [Accessed: 17- Dec- 2019].