CS-E5740 Complex Networks, Answers to the project exercise set

Benjamin Stumpf, Student number: 918561

December 10, 2020

Problem 1

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?
The timestamp, at which Anchorage becomes infected is 1229290800.

Problem 2

a) Plot the averaged prevalence $p(t)$ of the disease (fraction of infected nodes) as a function of time for each of the infection probabilities:

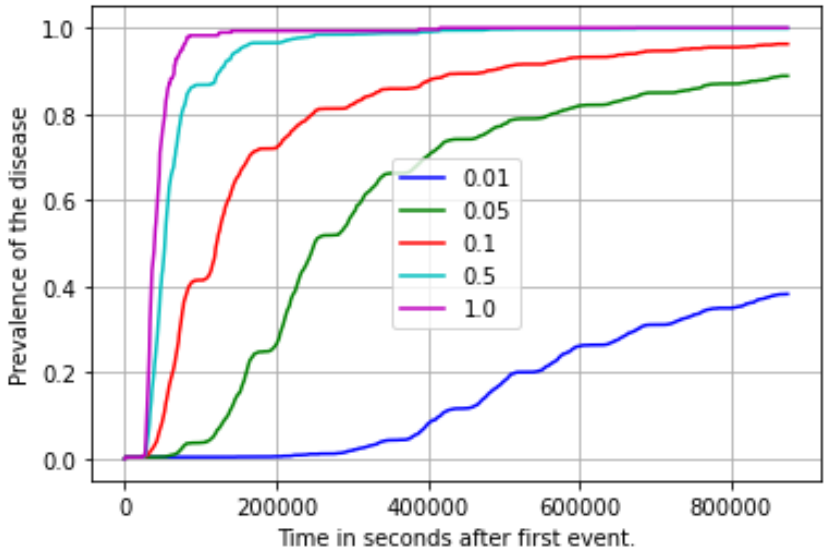

Figure 1: Averaged prevalence over 10 iterations as a function of time for the five infection probabilities

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

Only for the infection probabilities 1 and 0.5 does the whole network become infected. The periodic steps in the curves are probably due to airports with a lot of connections becoming infected, meaning that they can infect a lot of other airports, as they act as hubs.

Problem 3

a) Plot the average prevalence of the disease separately for each seed node as a function of time:

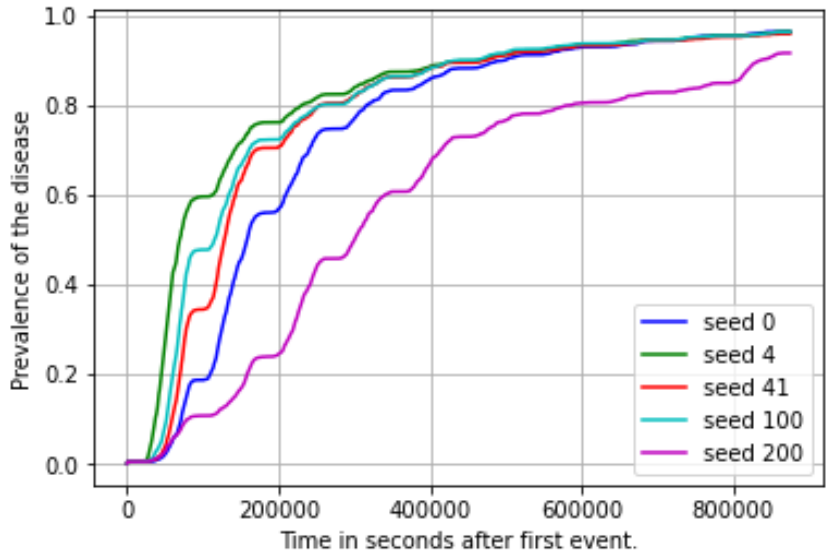

Figure 2: Averaged prevalence over 10 iterations as a function of time for the five seeds

b) The differences in spreading speed between seeds should be mostly visible in the beginning of the epidemic. Explain, why.

Indeed, the biggest difference in the spreading speed is visible in the beginning of the epidemic. This is due to the fact that the virus can spread quicker if our seed is a well connected node. Towards the end all graphs seem to converge towards 0.95, apart from seed 200.

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 will it be important to average the results over different seed nodes?

Because, for example, two nodes a and b of degree 1 each could both be different in terms of vulnerability: Let's imagine a scenario where node a is connected to a hub, while node b is connected to a node with low degree. Node a will be way more vulnerable as its only connected node is very likely to be infected early on as it is a hub while node b is less vulnerable as its only connected node is not that well connected itself and is therefore less likely to be infected early on.

Problem 4

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

i) unweighted clustering coefficient c (top left)
ii) degree k (top right)
iii) strength s (bottom left)
iv) unweighted betweenness centrality (bottom right)

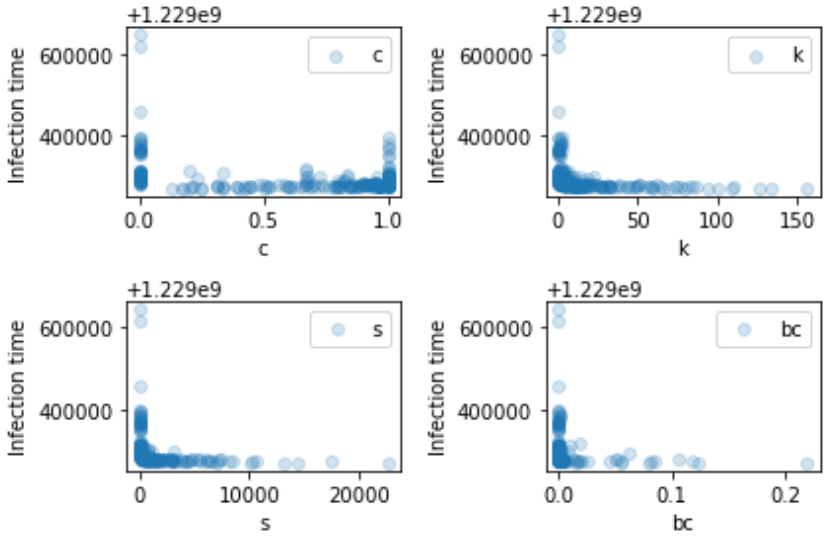

Figure 3: Median infection time as a function of four different network measures

b) Which of the measures is the best predictor for the infection times based on the Spearman rank-correlation coefficient?

The resulting spearman rank-correlation coefficients are:
Spearman r between median infection time and unweighted clustering coefficient: SpearmanrResult(correlation=-0.13463220408184623, pvalue=0.024515449702548084)

Spearman r between median infection time and degree: SpearmanrResult(correlation=-0.8270927388125542, pvalue=2.8380014170423936e-71)
Spearman r between median infection time and strength: SpearmanrResult(correlation=-0.896900598659845, pvalue=3.735386956075509e-100)
Spearman r between median infection time and unweighted betweenness centrality: SpearmanrResult(correlation=-0.6465007704210717, pvalue=2.056977118059764e-34)

The best measure to predict node infection time seems to be strength, as it has the highest absolute Spearmanr result (correlation=-0.896900598659845). This can be explained by the fact that a high strength indicates a high amount of neighbors and a high amount of flights between each neighbor, which obviously makes the node very vulnerable if one of the many neighbors is infected.

c) Which measure(s) would you use to pick the place to hide, i.e. which measure best predicts node infection time? Why?

The measures strength and degree both are suitable to pick a place to hide, with strength obviously being the better choice as laid out in b). The reason I would choose one of these measures is that they both have a high negative correlation with infection time according to the Spearman rank-correlation coefficient, with correlation=-0.896900598659845 and correlation=-0.8270927388125542 respectively. This high negative correlation can be explained that nodes with high strength and high degrees have many neighbors, meaning that they have a lot of nodes that they can be infected from. Strength has a higher negative correlation because here the neighbors additionally are connected to via links with high weight, which in our case means flight routes that are frequently used, making the infection risk even higher. As a result I would choose a node with a low degree and strength to hide.

Why does betweenness centrality behave differently than degree and strength?

We can see that betweenness centrality correlates worse with median infection time than strength and degree. We observe this by looking at the Spearman rank-correlation coefficient, which is correlation=-0.896900598659845 and correlation=-0.8270927388125542 for strength and degree respectively, and correlation=-0.6465007704210717 for betweenness centrality. The reason for this is that while nodes in the center might have high betweenness centrality and get infected early, the measure can't really make any useful predictions for the infection times of nodes with low betweenness centrality.

Why is clustering coefficient a poor predictor?

We can see that the clustering coefficient correlates very badly with median infection

time as the Spearman rank-correlation coefficient gives us correlation=-0.13463220408184623. The reason for this is that no matter whether the clustering coefficient is high or low, it won't tell us whether a node will get infected early or later, as the clustering coefficient measures how well connected a node's neighbors are, but not the node itself. As an example, a node of high degree will have a low clustering coefficient if its neighbors are not well interconnected, but it might still get infected early, while a node with low degree whose neighbors are well connected is less likely to be infected early on.

Problem 5

a) Plot of the prevalence of the disease as a function of time for the 6 different immunization strategies (social net., random node, and 4 nodal network measures)

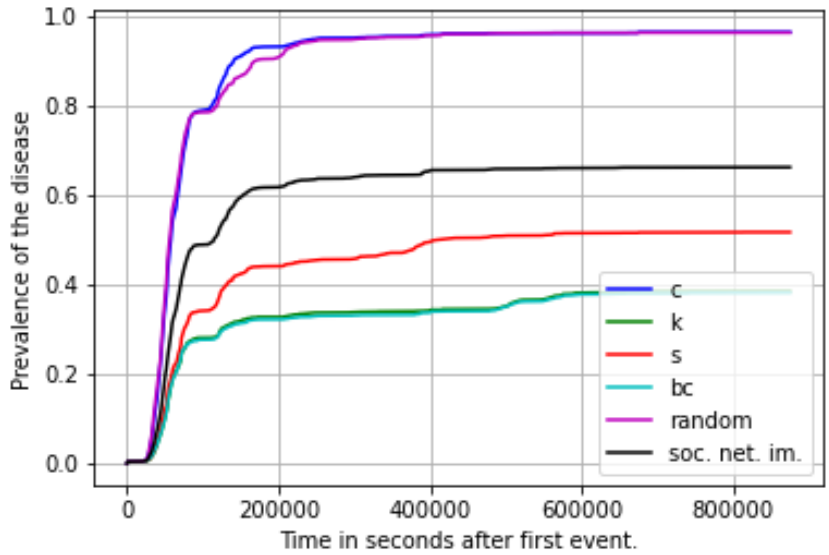

Figure 4: Prevalence of the disease as a function of time for 6 different immunization strategies

b) Which of the immunization strategies performs the best, and why?

Degree and Betweenness centrality perform best when it comes to immunization strategies. Degree immunization performs well because by making the 10 nodes with the highest degree immune we not only ensure that these nodes can't get infected, but also that they can't infect other nodes, which is crucial for these nodes with high degrees as they are hubs with the ability to spread the virus quickly if not immunized. The reason for Betweenness centrality immunization being a good strategy is that

these nodes with high betweeeness centrality being immune means that a lot of the shortest paths can't be use by the epidemic anymore, as an immune node can't be infected but also can't infect others. This leads to the epidemic spreading way slower than if these nodes weren't immune.

Why does betweenness centrality perform better as an immunization strategy than as a predictor for a safe hiding place?

Betweenness centrality performs better as an immunization strategy because we chose the 10 nodes with the highest betweenness centrality. The nodes with high betweenness centrality are actually also a good predictor for a safe hiding place, however the betweenness centrality of all nodes can not be used as a good predictor for a safe hiding place as nodes with low betweenness centrality can either be infected very early or later on and it is near impossible to make a prediction for them.

c) First, if the degree distribution of the network is $P(k)$, what is the probability of picking a random node of degree k?

The probablity is $P(k)$

What is the expected outcome if you then pick a random neighbour of the random node?

It is expected that the random neighbour picked has a higher degree than our initial random node. This is also known as the friendship paradox, which describes that on average, a person has less friends than their average friend. When we follow a random link, each node has k opportunities to be picked, so as many opportunities as its degree, as denoted by $p(k_{nn} == k) \propto kp(k)$. Applying this to the whole network leads to $\langle k_{nn} \rangle = \frac{\langle k^2 \rangle}{\langle k \rangle}$, which means that for a node with degree k the average neighbor has a degree of $\frac{\langle k^2 \rangle}{\langle k \rangle}$. This paradox can also be applied outside social networks as for example in our use case, meaning that we are more likely to pick a neighbour that has a higher degree than one with a lower degree.

Consequently, which of the strategies is expected to be more effective and why?

As a consequence, picking a random neighbor is expected to be a better immunization strategy as we expect to pick a node with higher degree than when just picking a random node. As we already observed earlier, higher degree nodes being immunized is a very effective strategy to delay the pandemic. Our conclusion seems to be correct as we can observe in the graph from a) that random neighbor is a better immunization strategy than random node.

d) Explain shortly, why it still makes sense to use this strategy in the context of social

networks?

It makes sense because it is very easy to apply. One doesn't have to look for highest degree nodes as this information is sometimes maybe not available but can just pick a random neighbor of a random node. Furthermore, social networks are of dynamic nature, meaning that relying on static strategies like highest degree or betweenness strategy might not be very effective as the numbers are constantly changing.

Problem 6

a) Links of the network on top of a map of the USA with width of the links according to the fractions f_{ij} to better see the overall structure and comparing it to the maximum spanning tree of the same network.

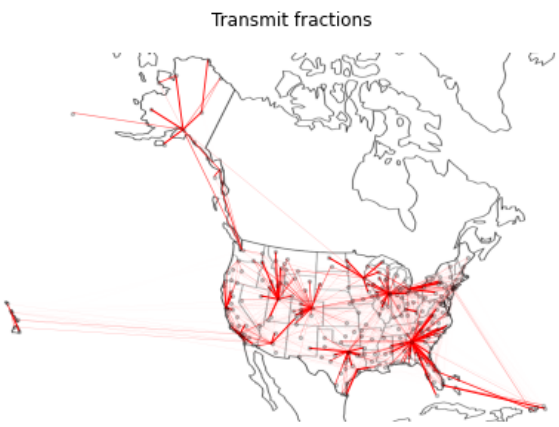

Figure 5: Map of USA with network links. Width adjusted to the fractions f_{ij}

Figure 6: Maximum spanning tree of the network on top of the map of the USA

b) Explain why your visualization is similar to the maximal spanning tree.

We see that the links with the biggest width from figure 5 are the same as the ones included in the MST. We further observe that the hubs are visible in both figures. So overall both figures are similar. This is because in a maximum spanning tree, the edges with the biggest weight are returned. In our example weight corresponds to how often a flight takes place, making it more likely that such a link is often infecting another node, which is why the two visualizations looks very similar.

c) Create scatter plots showing f_{ij} as a function of the following link properties:
i) link weight w_{ij}, shown in figure 7
ii) unweighted link betweenness centrality eb_{ij}, shown in figure 8

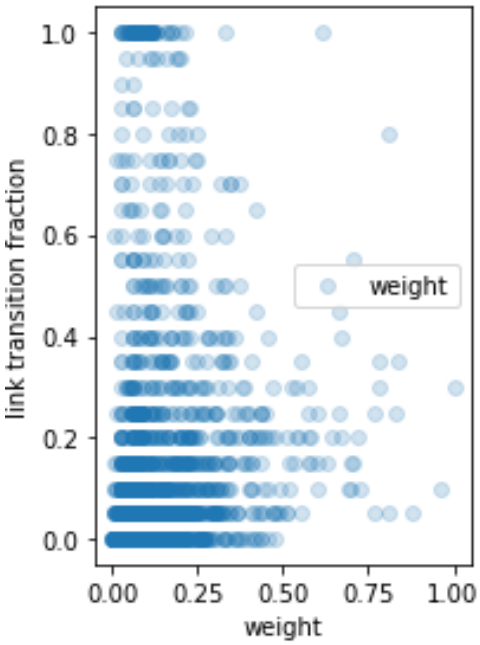

Figure 7: f_{ij} as a function of link weight w_{ij}

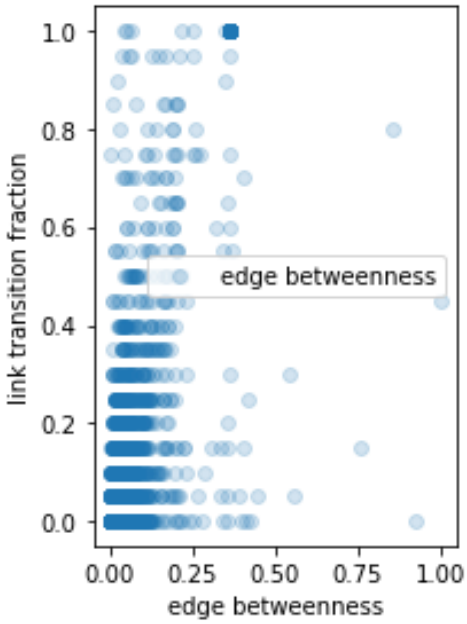

Figure 8: f_{ij} as a function of unweighted link betweenness centrality eb_{ij}

Compute also the Spearman correlation coefficients between f_{ij} and the two link-wise measures.

The spearman correlation coefficient between f_{ij} and link weight w_{ij} is: 0.36814387134428583.

The spearman correlation coefficient between f_{ij} and unweighted link betweenness centrality eb_{ij} is: 0.5252580776819579

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

We can observe that both link properties score rather badly in terms of correlation with f_{ij}. Link weight scores badly because while it might be true that links with high weight are more likely to transfer the disease a low link weight is not necessarily unlikely to transmit the disease, especially if a link is one of the few links in total leading to a node. Therefore link weight correlates badly with f_{ij}. Betweenness centrality correlates a bit better with f_{ij} because links that are on the shortest path between nodes tend to be centrally located and thus have more chances to transmit the disease. However, this doesn't always have to be the case which is why the correlation is not that high.

Problem 7

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

One way to improve the current epidemic model would be to use a more realistic model than the SI model. One example of such a model would be the SIR model, which allows infected nodes to recover, which is a more realistic scenario for most pandemics than nodes staying infected forever as it is the case in our present model. The SIRS model goes even a step further than the SIR model by including an immunity period, in which the node can not be infected again. Overall, it is important to note that the choice of the model is heavily dependent on the disease that has to be modelled, as different diseases can have different features that heavily influence their spreading capabilities.
The idea of using a SIR or SIRS model was inspired by the paper given in the exercise instructions [1]

Another way to improve the current model is to not only observe air travel, but also travels by train, car, ship etc.. Some diseases might even spread through animals, in which case the modelling would have to be even more elaborate. Other ways of spreading that should also be considered are for example through food, water or air. In order to take all of these ways of spreading into account there should probably be one node per human being instead of one node per airport, as this is a pretty significant simplification. That way nodes could also be deleted, if the pandemic starts killing people which in turn would have an impact on infection risk as less nodes are around.

Building on top of the previous argument, another feature to improve the model would be to allow for governmental restrictions and recommendations to be modelled as well. This again can be implemented way easier if the nodes are distributed as one per human. That way a curfew for example could be modelled by minimizing infection rate for a certain area where some of our nodes are between certain hours. Also a mask mandate could be modelled by reducing infection risk in areas where masks have to be worn.

Finally, the model should be able to show the effects of a vaccination being deployed. The same way a pandemic starts it also should be able to end in the model. This can easily be implemented by adding a variable to each node stating whether that node is vaccinated or not, which makes the node immune after a certain period of time. That way interesting predictions could be made about how quickly the pandemic disappears if for example 50, 60 or 70 percent of the nodes are vaccinated.

Sources

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