CS-E5740 Complex Networks, Final Project

Tommaso Canova, Student number: 101613341

December 26, 2023

Preface

The topic of this project work that we have been running for a long time has, unfortunately, become all too real in the recent years. This shows that pandemics are not entirely unexpected, and the way how they spread through the globe has also been foreseen for a long time. In fact, if one would use global air transport data, a more complex disease model (SEIR), and the real geographic origin of spread, one would get a timeline that eerily matches early 2020.

Introduction

In the project work, your task is to implement a Susceptible-Infected (SI) disease spreading model and run it on top of a temporal network from air transport data, containing information on the departure and arrival times of flights. You will study the dynamics of spreading and how it depends on where the process starts as well as the infectivity of the disease, and use static network centrality measures to understand the roles that specific nodes play.

Model specifications

In the SI model, each node is either Susceptible (S) or Infected (I). When an Infected node is in contact with a Susceptible node, the Susceptible node may become infected with some probability $p \in [0,1]$, reflecting the infectivity of the disease. Infected nodes remain Infected forever. In our model that mimics the spreading of disease through the air transport network, nodes are airports and time-stamped connections are flights between them. Initially, only one airport node (called the seed node) is set to the Infected state, while all other airports are Susceptible. Now, following the SI process, a flight from an Infected source airport infects its Susceptible destination airport with probability $p \in [0,1]$.

Note that a flight can carry the infection only if its source airport is infected at the time of the flight's departure! Infected airports remain infected for the rest of the simulation.

Task 1: Basic implementation

a) *If Salt Lake City (SLC, node-id=27) is infected at the beginning of the data set, at which time does Anchorage (ANC, node-id=41) become infected?*
Assuming $p = 1$ of infection the implemented SI allows us to observe that if *Salt Lake City* (SLC, node_id=27) airport is infected at the beginning of the simulation *Anchorage* (ANC, node_id=41) airport become infected at time: **1229283600**

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. You should be able to spot stepwise, nearly periodic plateaus in the curves.*
The result of the simulation is reported in Figure 1.

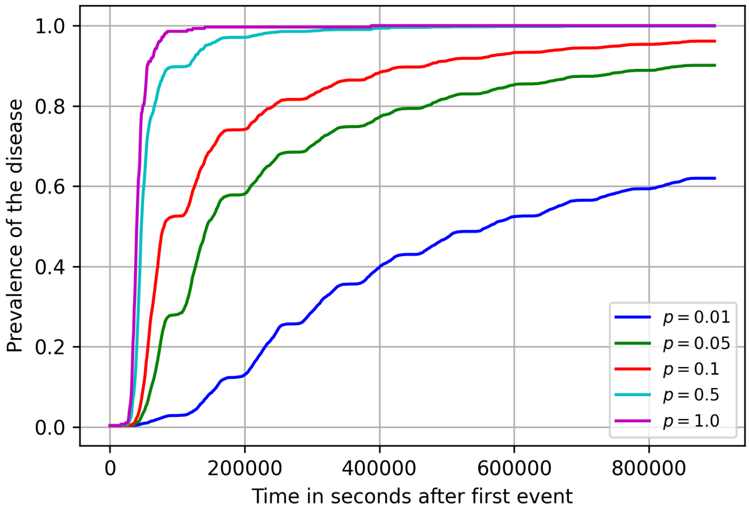

Figure 1: Infection spreading speed comparison between different infection probabilities

b) *For which infection probabilities does the whole network become fully infected? How can we explain the periodic "steps" in the curves?*
For each of the following infection probabilities [0.01, 0.05, 0.1, 0.5, 1.0] the SI model has been run 30 times and then the result for each probability has been averaged.

The SLC airport has been considered as the first infected node for every simulation. As we can be observe from Figure 1 the network become **fully infected** when the model is run with probabilities $p = 0.5$ and $p = 1.0$. Moreover, the curves manifest some periodic "steps" over time, each of these steps represent the short amount of time when most of the planes in charge of transmitting the infection are effectively flying. In fact, the infection occurs only when the planes effectively land to destination, for this reason the virus cannot spread during those steps and the prevalence of the disease remains constant.

Task 3: Effect of seed node selection on spreading speed

a) *Use nodes with node-ids [5, 38, 118, 134, 139] (DTW, SMF, CRW, DHN, GGG) as seeds and $p = 0.1$, and run the simulation 30 times for each seed node. Then, plot the average prevalence of the disease separately for each seed node as a function of time (recycling your code for Task 2).*
The result of the simulation is reported in Figure 2.

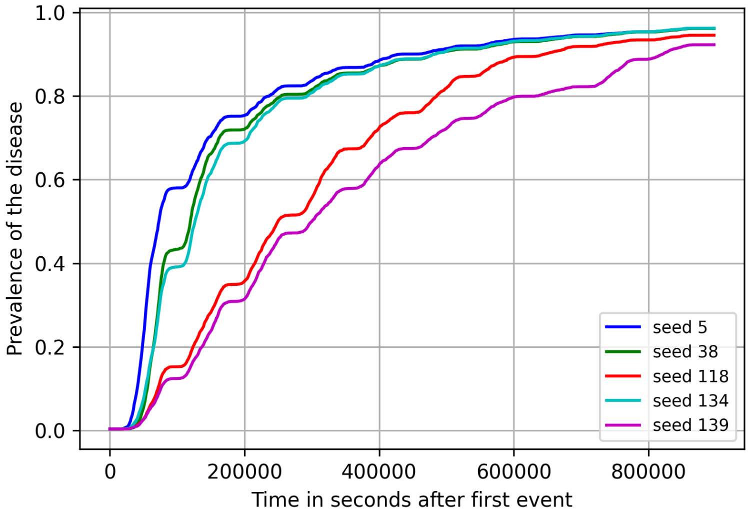

Figure 2: Average prevalence of disease for different random nodes over time

b) *The differences in spreading speed between seeds should be mostly visible in the beginning of the epidemic. Explain, why.*
As we can observe from Figure 2, the disease is spread faster during the beginning of the simulation when nodes 5, 38 and 134 are considered as initially infected nodes. This high prevalence of disease is due to the fact that these three nodes have a high degree and for this reason many flights will departure from their airport and then

more infections will probably be carried out to different destinations. Afterwards, this degree dependency in the infection will decrease over time since more nodes will be infected. Considering the network as a directed graph, Table 1 reports the out-degree of each initially infected used used for each simulation.

Node	Degree
5	109
38	24
118	2
134	13
139	1

Table 1: Out degrees of initial infected nodes used in Task 3 simulation considered the graph as directed

c) *Why is it important to average the results over different seeds? What kind of problems could follow from using only a single seed, for example in the next task where we'll inspect the vulnerability of a node for becoming infected with respect to various network centrality measures?*
By averaging the results over different seed we are able to obtain more robust results, thus reducing the randomness due to the initial selected node. On the other hand, if we would rely just only a single seed node the outcome would not be representative of the whole network, and therefore, the network dynamics would not be captured as well. As we have noticed from the previous simulation, changing the initial infected seed node leads to different results in the fraction of infected nodes at the end of the simulation.

Task 4: Where to hide?

a) *Run 100 simulations, and create scatter plots showing the median infection time of each node as a function of the following nodal network measures: unweighted clustering coefficient c, degree k, strength s, unweighted betweenness centrality*
The obtained plots are reported in Figure 3.

b) *Use the Spearman rank-correlation coefficient for finding out which of the measures is the best predictor for the infection times*
For each node attribute the Spearman rank-correlation coefficient has been computed and the result is reported in Table 2.

c) *Based on your results, answer the following questions:*

- *Which measure(s) would you use to pick the place to hide, i.e. which measure best predicts node infection time? Why are these measures good in predicting*

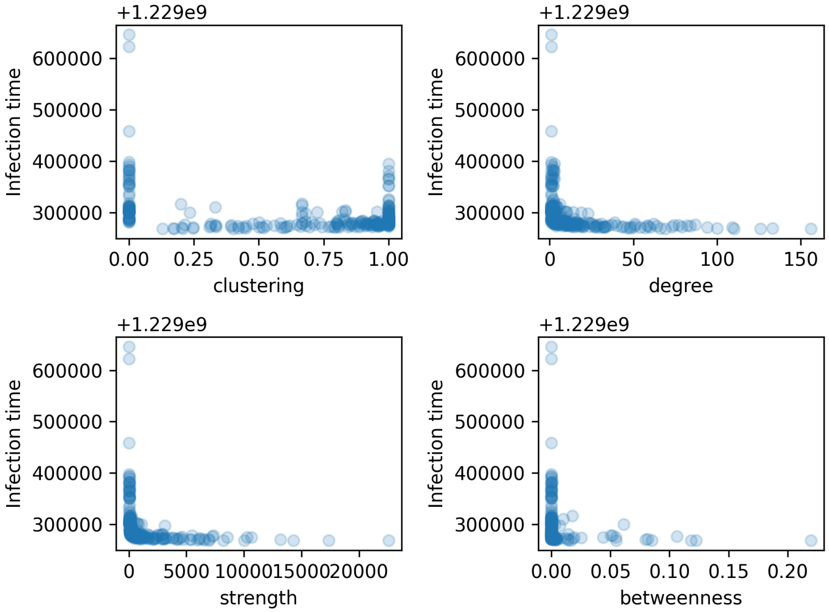

Figure 3: Scatter plot of infection time based on different node attributes

Node Attribute	**Clustering**	**Degree**	**Strength**	**Betweenness**
Spearman Coefficient	-0.142	-0.819	-0.887	-0.626

Table 2: Spearman rank-correlation coefficient for different node attributes

the infection time?

Observing the results reported in Table 2 it is clear that the two measures that best predict the infection times of the nodes are strength and degree. Indeed we can observe that we have a higher negative correlation between the infection times and these two measures, this means that nodes with high values on these two attributes will be infected earlier during the spread of the disease. These results suggests that the best place to hide in order to be as safe as possible from the epidemic are the **nodes with low degree and low strength**. As we can recall from the definition the degree of a node represents the amount of links between that node and other nodes of the network, on the other hand the strength of node in a network is computed as the sum of the link weights of its neighbours. For instance, in this case the weight of each link has been considered as the amount of flights between two airports. It is therefore evident that the fewer connections and flights there are between one or more airports, the less likely it is that they will be infected.

– *Why does betweenness centrality behave differently than degree and strength?*
The Betweenness centrality allows us to find hubs in the networks, thus nodes

that connects communities to each other. In this case this measure behave differently with respect to degree or strenght measures since its value can only determine the speed the infection but it does not guarantee the effective spread of it. For instance, hubs with low degree could be infected later than nodes with high degree.

- *Why is the clustering coefficient a poor predictor?*
  The clustering coefficient gives us an information on how much the neighbours of a node are connected to each other. However, this information is not useful in predicting the infection time since the disease can be transmitted only trough direct links, therefore the connectivity of the neighbours can be ignored.

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 6 different immunization strategies (social net., random node, and 4 nodal network measures), always immunizing 10 nodes.*
The result of the six different immunization strategies is reported in Figure 4

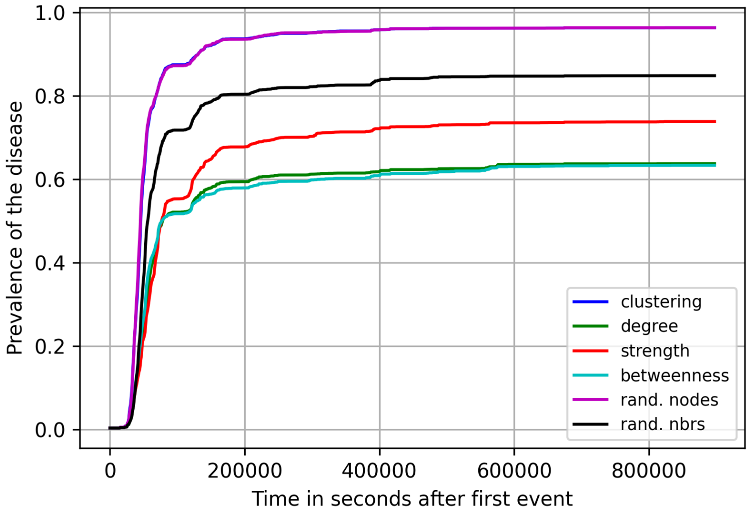

Figure 4: Prevalence of disease over time according to different immunization strategies

b) *Based on your results, answer the following questions:*

- *Which immunization strategy performs the best, and why?*
  The best immunization strategy is obtained by exploiting the betweenness centrality as a criteria to find the best 10 airports that need to be immunized. This

strategy is the best compared to the others because nodes with high betweenness centrality act like hubs, consequently, by immunizing them we are able to keep separate the communities, avoiding to let the disease spread through the hubs.

- *Why does betweenness centrality perform better as an immunization strategy than as a predictor for a safe hiding place?*
  As explained above, this measures works well as immunization strategy, however it should not be considered as a good predictor to find a safe place to hide during an epidemic. In fact, the virus is spread through direct links, therefore it does not matter if a node has high or low betweenness centrality, since this measure gives us only an information on shortest path and not on direct links.

c) *The random 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?*
  The probability of picking a random node of degree k is equal to $\frac{k}{2e}$, where we multiply by two the number of edges e since we are dealing with an undirected graph. While if we have to pick n nodes with degree k the total probability will be $\frac{nk}{2e}$.
- *What is the expected outcome if you then pick a random neighbour of the random node?*
  Following a random link of a random node our expected outcome will be equal to $\frac{\langle k^2 \rangle}{\langle k \rangle}$.
- *Consequently, which of the strategies is expected to be more effective and why?*
  According to Figure 4, selecting a random neighbour of a random node can be considered the best strategy compared to random node selection since we know that, thanks to the Friendship paradox, it will have an higher degree than the random node itself. For this reason, if we immunize the random neighbour we will be able to avoid the infections over other links and therefore to slow down the epidemic, since the neighbour will have an higher degree than the node itself.

d) *Although the random neighbour immunization strategy outperforms the random immunization, it may be less effective than some other immunization strategies. Nevertheless, explain shortly, why it still makes sense to use this strategy in the context of social networks.*
First of all, we have to recall that the Friendship paradox holds for social networks, therefore using this immunization strategy can be still considered a good choice. Secondly, this immunization strategy has a lower computational cost compared to other measures because we do not need any prior knowledge of the network. For instance, using other metrics like clustering coefficient, we should compute it for the whole

network and then sort the result in order to immunize the best top k nodes, which is computationally more demanding with respect to a random neighbour immunization.

Task 6

a) *Run the simulations, and compute the fraction of simulations where each link transmitted the disease (f_{ij}). For example, if the disease spread from node a to node b in 10 simulations out of 50, fab = 0.2. Compare your visualization with the maximal spanning tree of the network.*

Transmit fractions
Maximum spanning tree

(a) Fraction of links spreading the disease during the simulations (opacity is a function of f_{ij})

(b) Maximal spanning tree of the network

Figure 5: Disease spreading comparing the simulations with the maximal spanning tree

b) *Explain why your visualization is similar to the maximal spanning tree.*
Recalling the theory, the maximal spanning tree represent the subset of the original links so that the tree contains all the nodes of the original network and the sum of the link weights is maximized. Observing the plot in Figure 5a, the higher the link transmission fraction the higher is the chance that a link is in charge of transmitting the disease, in fact, this concept is strongly related with the weight of a link since, the more flights between two airports and the higher chances that a link is infectious. For this reason, both plots shows similar results. In addition, the plot on the left can be seen as a smoothed version of the MST (Figure 5b) because the opacity of the links is reduced based on the fraction of times the link is present during the simulation instead of just considering it as a whole.

c) *Create scatter plots showing f_{ij} as a function of the following link properties: link weight, unweighted link betweenness centrality. Compute also the Spearman correlation coefficients between f_{ij} and the two link-wise measures.*
Spearman correlation coefficients for the two properties and plots of f_{ij} according to the two measures are reported in Table 3 and Figure 6 respectively.

Link Attribute	**Weight**	**Unweighted link betweenness centrality**
Spearman Coefficient	0.457	0.551

Table 3: Spearman rank-correlation coefficient for different node attributes

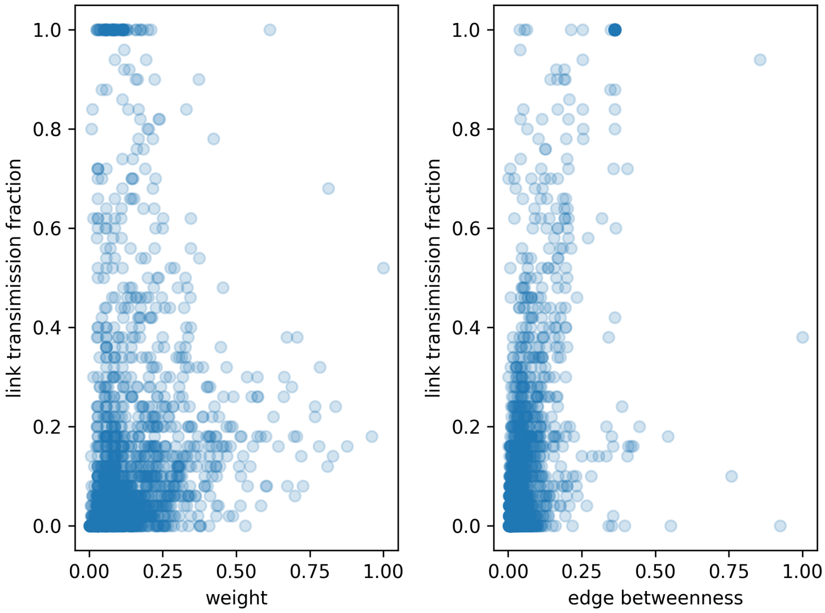

Figure 6: Link transmission fraction according to link weight and edge betweenness measures

d) *How well do the two link properties predict f_{ij} and why? Explain their performance based on how they are defined.*

As we can observe from Table 3 the coefficients for both measures are positive, so we can state that an higher value of these measures contributes with an higher disease transmission rate among the links.

As mentioned before, the information concerning the amount of flights between two airports is encapsulated in the weight attribute. As expected, the more flights between two nodes and the higher the probability of spreading the disease.

On the other hand, the edge betweenness represents the fraction of all shortest paths that traverse a specific edge relative to the total number of shortest paths in the network. A higher edge betweenness indicates that a particular edge is traversed by a greater number of shortest paths, implying that the associated link serves as a critical connector between one or more clusters. For this reason links with higher BC are the one more involved in transmitting the disease from one community to the other, as they act as bridges.

Task 7: Discussion

Even though extremely simplistic, our SI model can 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.

The model could be improved according to different strategies, here listed four:

1. The model could be made more complex, transitioning from a simple SI to a SEIRS, in which the following properties are introduced:

    - Exposed nodes (E), which, after being exposed to the infectious agent, are not yet infectious (representative of the incubation period).
    - Recovered nodes (R), which have experienced the disease and are considered immune to it.
    - Susceptible again nodes (S): nodes that, after being of type R, have lost immunity and are susceptible to infections again.

2. In this case we have just considered the airports as they were all the same entities, however they differ from city to city, in terms of size, workers involved, amount of customers and so on. For this reason the model could be improved by considering in the dataset both the amount of people living in each airport city and at the same time the average amount of people that travel on each flight. With these data we would be able to better understand how the disease can spread outside the airport context.

3. The dataset could be enriched by considering the real departure and arrival times in order to separate morning and night flights. Having this information could be useful to detect hidden pattern in the disease spreading. For instance, people may travel more after landing to an airport in the morning with respect to night flights, and therefore spreading more the disease.

4. To better understand how to prevent the spread of the disease, another enhancement that can be introduced into the model is to use an infection threshold, upon reaching which the airport is temporarily closed for a specified time period (also variable). This solution could be implemented in addition to the SEIRS model. By conducting various simulations with variable thresholds, a better understanding of the contribution of each airport to the virus spread could be gained. At the same time, it is possible to identify ”safety thresholds” for which the proper functioning of airports can be ensured, minimizing the spread of the disease.