CS-E5740 Complex Networks, Project

Sara Cabodi, Student number: 784287

December 20, 2019

Compile with `pdflatex ex_template.tex`

Task 1

a) If Allentown (node-id = 0) is infected at the beginning of the data set, Anchorage (ANC, node-id = 41) becomes infected at time **1229290800** (seconds after the epoch).

Task 2

After running the SI model 10 times with each of the infection probabilities [0.01, 0.05, 0.1, 0.5, 1.0] and having Allentown (node-id=0) as seed node, I recorded all infection times of the nodes.

a) The following graph shows averaged prevalence $\rho(t)$ of the disease as a function of time. The latter is represented starting from 0, which corresponds to the first flight's departure time (1229231100), and goes until 892800, that is the difference between the last arrival's time (1230123900) and the first one reported before. The choice was made just for plotting purposes and it's consistent throughout the project.
As expected, the fraction of infected nodes increases with the probability. With the lowest p (red curve), the maximum averaged prevalence is between 0.2 and 0.3 and the grade of growth of the function is quite low. Concerning the two next probabilities, $p = 0.05$ and $p = 0.1$, the trend is quite similar. At the beginning (until $t = 200000$) the grade of growth is sharper that the second part of the spreading, where the tendency is smoother. Lastly, for $p = 0.5$ and $p = 1.0$, the fraction of infected nodes almost reaches the maximum in a short amount of time and then they have a stable trend around it.

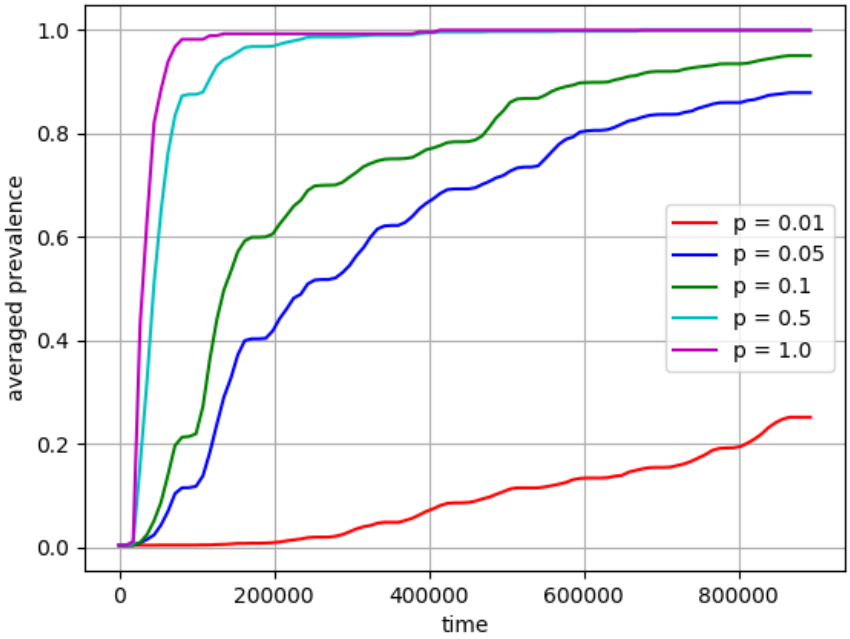


Figure 1: Plot of averaged prevalence $\rho(t)$ of the disease as a function of time

b) The whole network become fully infected for $p = 0.5$ and $p = 1.0$.
The stepwise are most probably due to the fact that there aren't any flights during night hours, and so the disease doesn't spread.

Task 3

a) Using as seed nodes those with node-ids [0, 4, 41, 100, 200] (ABE, ATL, ANC, HSV, DBQ) and p = 0.1, I ran the simulation 10 times for each seed node. Then, I plotted the average prevalence of the disease separately for each seed node as a function of time.
The result shows 5 curves, each one corresponding to a different seed node. The blu curve (seed-id=4, ATL) is the one that spreads the disease faster than the others. This is probably because Atlanta is the biggest hub among those considered and, with the same probability of infecting other nodes, has a higher chance of spreading it given the higher number of flights departing. The opposite can be said about the purple curve (seed-id=200, DBQ) where the seed is a small town in Iowa, Dubuque, and so has a smaller rate of spreading. The other curves are in between, but a bit closer to the Altanta one, and have similar trends.

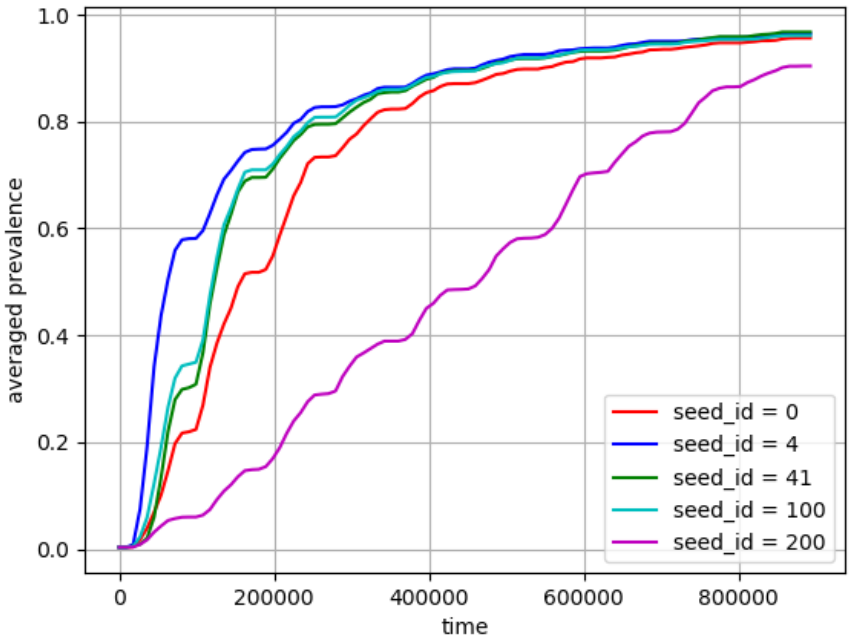


Figure 2: Plot of averaged prevalence $\rho(t)$ of the disease as a function of time

b) The differences in spreading speed are visible in the beginning of the epidemic. This happens because the speed is directly proportional to the dimension of the seed airport, which means that there is a high number of flights and therefore many departing ones that spread the disease.

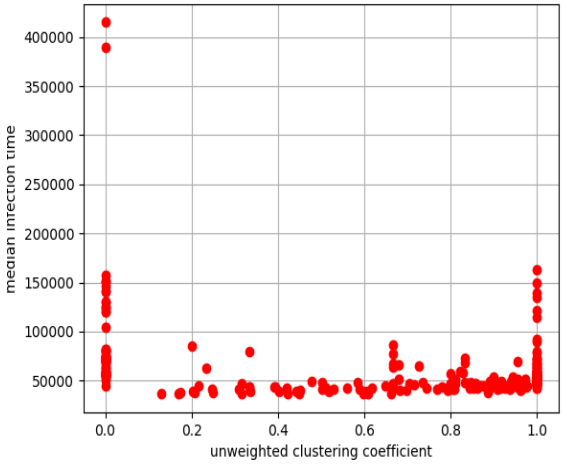
c) In statistics it's usually important to identify important parameters that can bias the outcome of the result of an inspection over certain data. In this task, we have discovered that the choice of the seed-id can chance the result of average prevalence in time. Therefore, in the next task it will be important to average over trials with difference seed-ids in order to get the best randomized result as possible.

Task 4

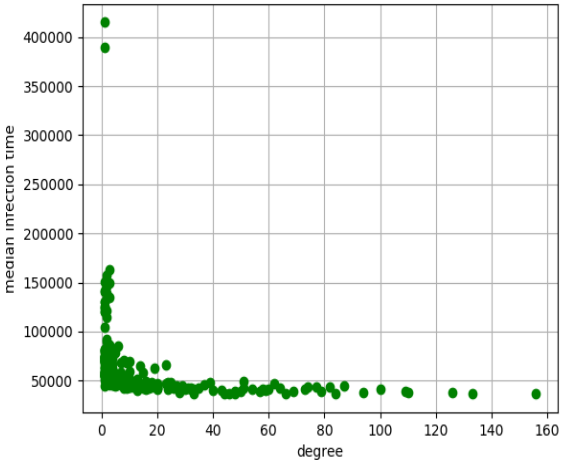
This task aims at analysing the network in order to select a refuge. To do this, I ran the SI model 50 times with $p = 0.5$ using different random nodes as seeds and record the median infection times for each node.

a) The following scatter plots show the **median infection time** of each node as a function of 4 nodal network measures:

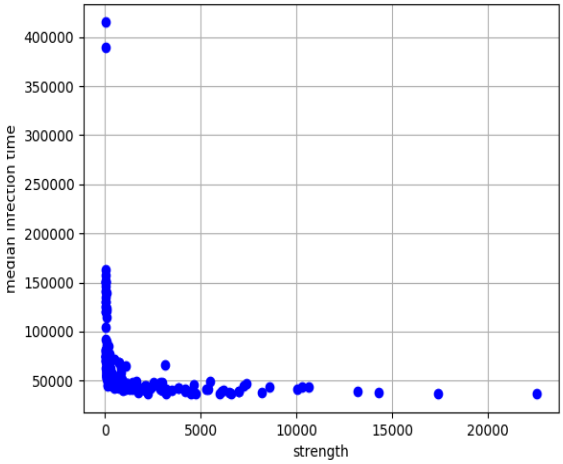
– unweighted clustering coefficient c
– degree k
– strength s
– unweighted betweenness centrality



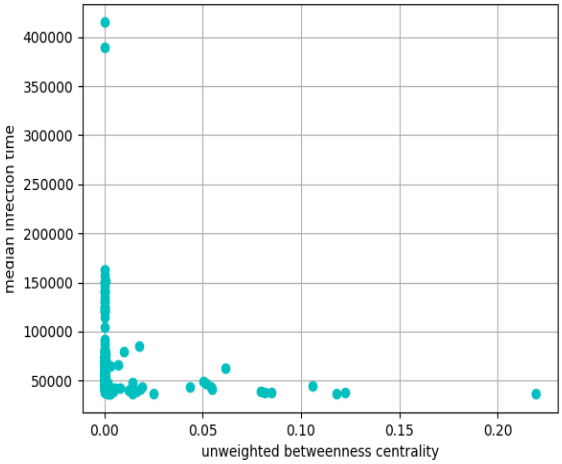
(a) Clustering coefficient



(b) Degree



(c) Strength



(d) Betweenness centrality

b) The best predictor for the infection times is is strength with a Spearman rank-correlation coefficient of -0.876.

c) Now I will discuss each plot separately.
The median infection times are normalized on the first flight's departure time, as before.

a) ***Clustering coefficient***
Starting from low clustering coefficient, we notice that the spreading through these nodes happens especially in the first half of the total time. There very few nodes with low clustering coefficient that have a high median time of infection. With the increase of the clustering coefficient, the median infection time concentrate in the initial part of the spreading, meaning that the higher c, the faster the node gets infected. It seems that there is not a defined trend and that the clustering coefficient cannot say much about nodes capacity of lasting before being infected. This hypothesis is confirmed by the Spermann rank-correlation coefficient that is equal to -0.124, which indicates a poor correlation of this measure (c) with the median infection time.

b) ***Degree***
With a probability $p = 0.5$, there seems to be very few nodes that resist the infection until a median infection time of almost 400000 seconds after the first infection time. These are probably isolated points with very low degree, or simply outliers in the statistical inquiry (?? Too much?). In general, we can see that nodes with low degree (up to $k = 10$ approximately) tend to get infected later, while those with higher degree get infected quite early in the spreading. From the plot, it seems that the higher the degree, the faster the node gets infected. This hypothesis is party confirmed by the Spermann rank-correlation coefficient of -0.811, which highlits a quite high correlation with the median infection time.

c) ***Strength***
The trend in the strength's scatter plot resemples a bit the one of the degree. It seems that for low s values the behaviour is quite the same as the previous one. For higher values though, we can denote a sparser distribution. This could be explained with the fact that is not as important to have a very high strength as it is to have a reasonable amount of it (devo rifrasare ma non so bene come!! TODO - conta piu che ci sia e non che sia fortissimo). This measure is sligthly better than the degree in terms of correlation, since it has a Spermann coefficient of -0.876, which is also the highest between the four.

d) ***Betweenness centrality***
This measure is in line with the previous two. For low values of betweenness centrality, it seems that the nodes are more likely to resist longer the infection. The more the x-value increases, the faster the node tends to be infected. This analysis is in accord with the definition of the measure of centrality, which is higher in nodes that work more as bridges. The distribution resembles the ones of degree and strength, but with lower correlation. As a matter of fact, it has a Spermann coefficient of -0.620, which highlights a partial correlation with the median infection time, but not as strong as the two before.

Task 5

a) In this task I compare different immunization strategies:

- *Social network immunization*, also called "pick-a-neighbour" strategy
- *Immunization of random nodes*
- Immunization of nodes that possess the largest values of
  - *unweighted clustering coefficient* c
  - *degree* k
  - *strength* s
  - *unweighted betweenness centrality*

In the following graph I plot the prevalence of the disease as a function of time for the 6 different immunization strategies, when 10 nodes are always immunized.

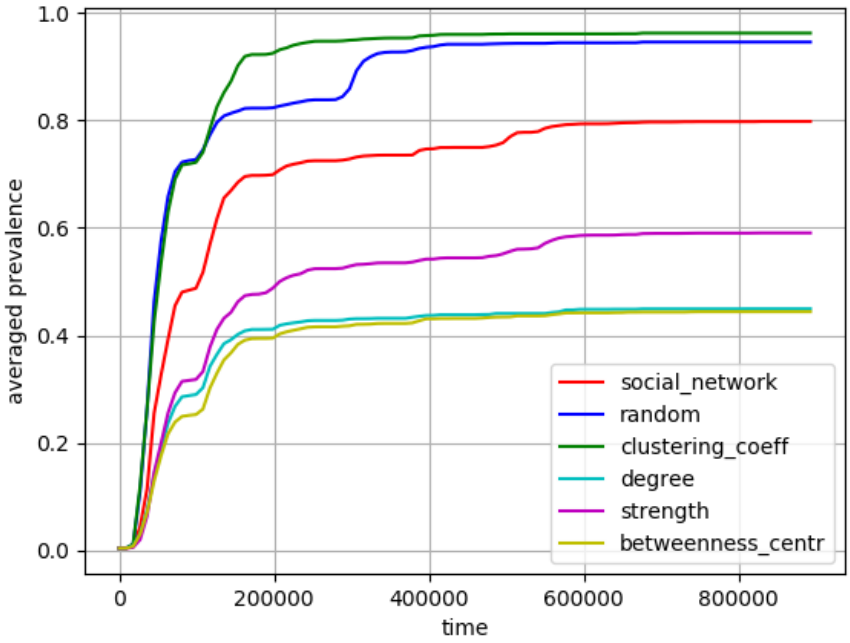


Figure 4: Plot of averaged prevalence $\rho(t)$ of the disease as a function of different immunization strategies

b) The betweenness centrality is the winning strategy, as it shuts down the biggest hubs in the network, those which have highest betweenness centrality and that work as bridges. Nodes with high values of this centrality measure turn out to be the most

effective in spreading the infection: intuitively, once they are infected, they can propagate the disease to many other adjacent nodes. Moreover, this measure is not as important for the prediction of infection time as it is with regards to immunization. This is probably because in the first case the choice of the seed-node is randomized and not based on this measure, while the choice of immunization is done on purpose, shutting down those with the highest values (i.e. those which could spread it the fastest). The degree strategy is quite as effective as the best one. As a matter of fact it is just slightly above the betweeeness centrality curve. This is because nodes with high degree can spread the disease to adjacent nodes almost as fast as those with high betweeness centrality. Therefore, immunizing nodes based on these two measures is quite the same and in both cases effective in terms of containing the spreading. Link weight (number of flights) is not so important, or at least less important that the degree: once the infection reaches a given node, an outgoing link with k flights has a probability to transmit the infection $(1-(1-p)^k)$, with $p = 0.5$, the probability is $1-0.5^k$, which means $0,875$ at $k = 3$, and almost $0,95$ at $k = 4$. This partly explains why strength is less effective than pure degree at determining shutting nodes.
Given the above considerations, it is easy to understand that a random node choice is at the exact opposite, as it doesn't target degree, but simply a random node, with average degree.
On the other hand, the social (random neighbor of a random node) is a better strategy, as the probability of piking a given node is proportional to its degree. Therefore this strategy is an intermediate one between maximum degree (or, though worse, maximum strength) and random.
The clustering coefficient strategy has a poor performance too, as (intuitively) removing a given node in a cluster will not affect overall infection times, as the infection will easily find alternative paths through other nodes in the cluster.

c) The probability of picking a node of degree k is

$$P(k) = \frac{n(k)}{N}$$

where

- N is the total number of nodes
- $n(k)$ is the number of nodes of degree k

The expected outcome if I then pick a random neighbour of the random node can be characterized by the nearest-neighbour's expected degree formula as follows

$$\langle k_{nn} \rangle = \frac{\langle k^2 \rangle}{\langle k \rangle}$$

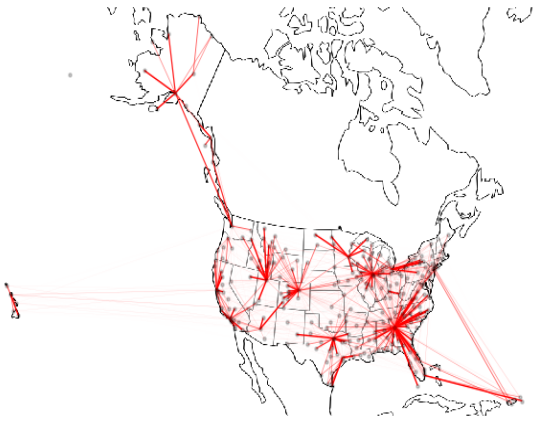
Given these results, in accordance with the *Friendship paradox*, a social network approach is the more effective, since the degree of a *friend of randomly selected individual is likely to have higher than average centrality.*

d) Although the social network immunization strategy outperforms the random immunization, it is not necessarily as effective as some other immunization strategies. However, it still makes sense to use this strategy in the context of social networks because of it's statistical nature. Usually, social networks have a huge number of nodes and edges and, therefore, it is very difficult and expensive to calculate degrees and strengths measures for each node and rank them. So, it is helpful to calculate the social network strategy since is a statistical measure that is always feasible to compute, even though it might not be the best one.

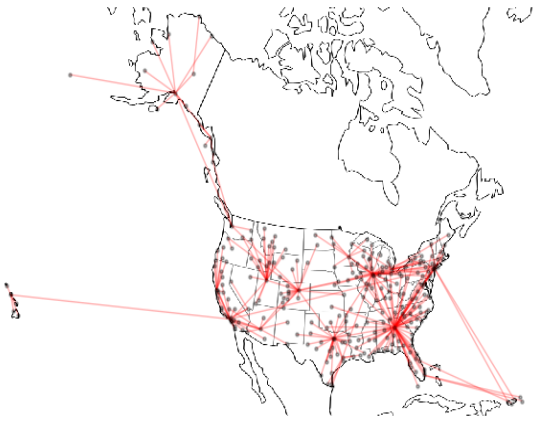
Task 6

In this task we will discuss the role of links. We will do this by recording the number of times that each link transmits the disease to another node. I ran 20 simulations using random nodes as seeds and $p = 0.5$. For each simulation, I recorded which links are used to infect yet uninfected airports (either susceptible airports or the infecting flight arrives before the already recorded infection time).

a) After running the simulations, I computed the fraction of times that each link is used for infecting the disease (f_{ij}). Then, I visualized the network (a) on top of the US map, adjusting the width of the links according to the f_{ij} fractions to better see the overall structure. Here is my result compared with the maximal spanning tree of the network.



(a) US network with fractions f_{ij}



(b) Maximal spanning tree of the US network

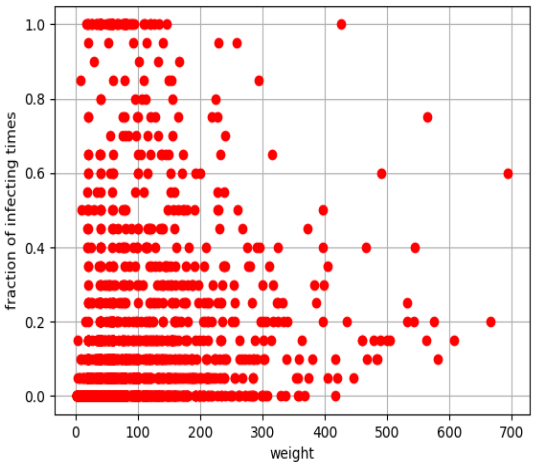
b) The use of the f_{ij} fractions definitely highlights the biggest hubs in the network together with the most used links for disease spreading. A similar information is also present in the MST plot, which aims at maximizing the overall sum of edge weights (in this case the number of flights between two airports). So, we can argue that they both tend to highlight the importance of big hubs and popular communications, that are also those which tend to spread most of the disease.

   More in detail, I can explain the strong similarity between the maximum spanning tree and the sub-graph of infected links as follows:

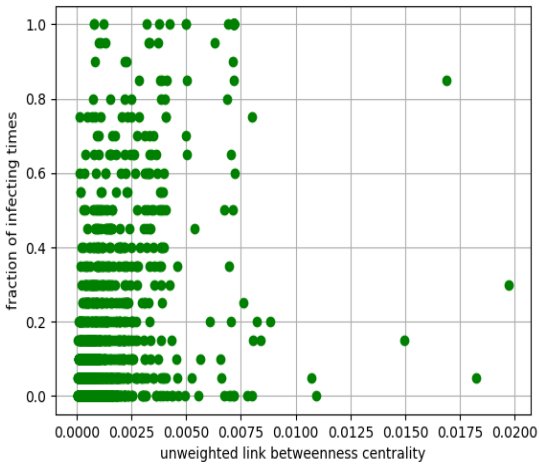
   - For any single infecting propagation experiment (from a selected source), the set of infecting links is a tree, that becomes a spanning tree whenever the number of considered flights is enough, for the chosen infection probability p, to spread the disease to all nodes
   - Given a set of experiments, the sub-graph of infecting links is the union of infecting trees of all experiments. The more the infecting trees are similar each other, the more they will match with a given spanning tree
   - Link weights play a role, as well as hubs, in both the maximum spanning tree and infection trees, which intuitively explains why they show strong similarities

c) Then, I created scatter plots showing f_{ij} as a function of the following link properties:

   - link weight w_{ij}
   - unweighted link betweenness centrality eb_{ij}



(a) Link weight



(b) Unweighted link betweenness centrality

Moreover, I computed the Spearmann correlation coefficient between f_{ij} and the two link-wise measures and here are my results:

- For link weight $\longrightarrow$ 0.36
- For betweenness centrality $\longrightarrow$ 0.49

d) The best predictor between the two considered is the link betweenness centrality. The edge betweenness centrality is defined as the number of the shortest paths that go through an edge in a graph or network. Weight-based edge scoring, as well as unweighted betweenness centrality, both capture some piece of information that is relevant to infection transmission:

- links with high weight have a higher probability to transmit infection, as they represent many flights, and they are likely to connect hubs (or important nodes), but weight is a local measure, that misses any information related to the global graph topology
- Betweenness centrality captures the importance of a given node/edge in shortest paths, which are good candidates to transmit infection

Of the two measures, the second one, though not perfect, is a better predictor of a link infection score, as the global (topological) nature of betweennes is more important than the purely local nature of edge weight.

Bonus Task

I had a look at a few papers from literature, coming up with some ideas, but no conclusive solution.

First of all, papers generally address the importance of nodes, rather than links, in order to transmit the infection. Though edges and nodes are clearly related each other, they can show some differences: see for instance the notion of a bridge and an articulation point in a graph (a bridge always implies the existence of two articulation points, whereas an articulation point can exist without any bridge).

An interesting idea that I found (*"Spreading dynamics in complex networks"*), is the importance of peripheral nodes(edges) in transmitting the infection. To this respect I found a comparison between PageRank and k-shell in order to evaluate the infection transmission importance.

When trying to combine the considerations on spanning trees (done in ex.6) with the above ideas, I came up with two strategies:

- Trying to combine link weights with unweighted betweenness centrality: I did this by computing weighted betwenness centrality. But I had to manipulate weights in order to lead to shortest paths: I thus first took the inverse $w' = \frac{1}{w}$, then I normalized it by taking the square/cube root, in order to take into account the non linear aspect of weights on spreading probability.

- Exploit the contribution of k-shells: k-shells are important to capture (almost) unique paths (with no/few alternatives), as low k values represent nodes with low degree. For instance, peripheral nodes (or bridges to peripheral nodes are expected to have low k-shell values. I first scored each edge with the lower k-shell value between the two adjacent nodes (worse case). Then, by just considering k-shell edge scores (alone), I was unable to beat the betweenness centrality results. So I tried to combine the two measures in one, taking k-shell scores as correction weights for betweenees centrality: I divided the betweenness centrality value by k-shell edge score in order to normalize it.

- Once I consolidated the approach trying with $w' = \frac{1}{w}$, $w' = \frac{1}{\sqrt{w}}$ and then $w' = \frac{1}{\sqrt[3]{w}}$, I applied the same procedure as the k-shell one to other measures of centrality of nodes. For each edge I took the lower measure of the two ending nodes and divided the betweenness centrality value by that value. I tried the following nodes centrality measures:

  - k-shell
  - PageRank
  - Closeness
  - Eigenvector

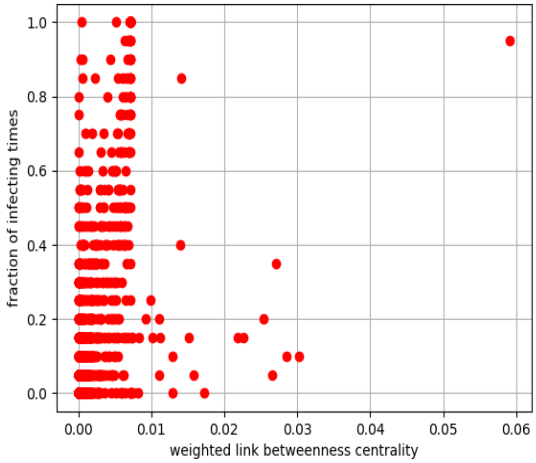
- Finally, I applied the procedure for the strength of nodes to see the difference from previous results, given that it was the best measure for predicting the infection times.

In the end, I found out that most of the measures of nodes centrality give pretty similar correlation coefficients. In particular, the best measure (highest Spearmann correlation coefficient) is obtained with eigenvector centrality whith a correlation of 0.6766.
Here there are all the correlation coefficients (computed with the contributions of centralities) that I computed:

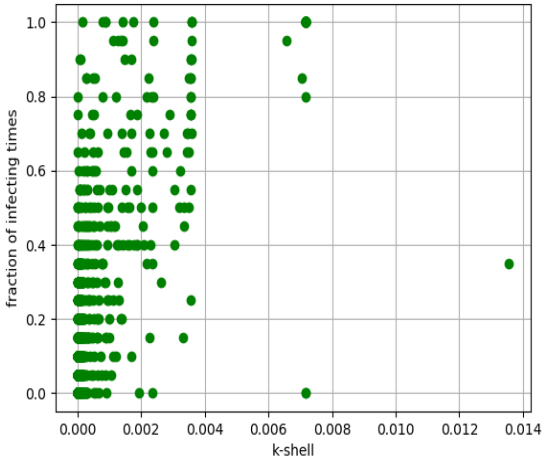
weighted	0.6366
w/k-shell	0.6741
w/PageRank	0.6710
w/Closeness	0.6457
w/Eigenvector	0.6766
w/Strength	0.6746

Note: each measure is a weighted betweenness centrality normalized as explained before.

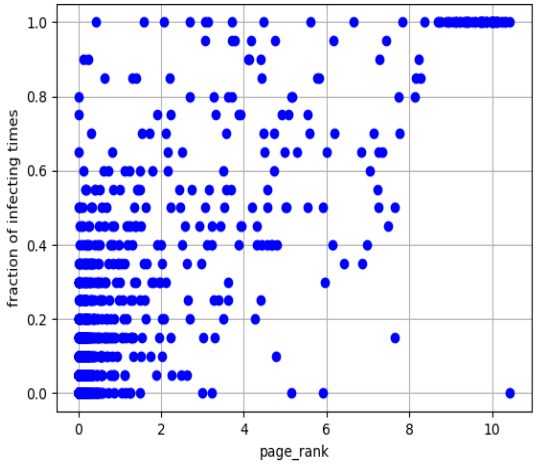
The following scatter plots show the results obtained from the procedure explained.



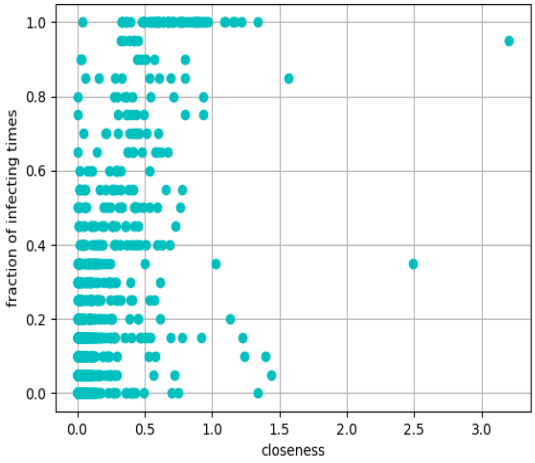
(a) Weighted betweenness centrality



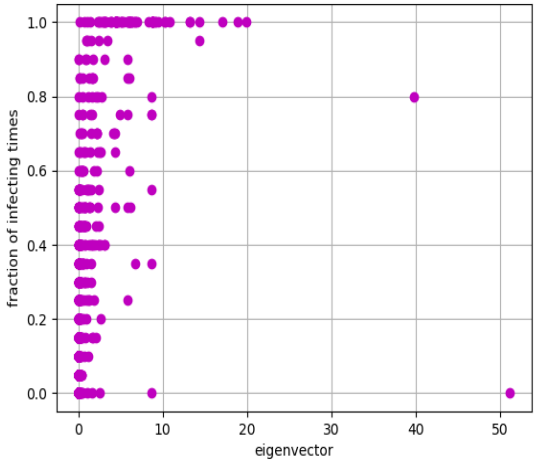
(b) k-shell weighted betweenness centrality



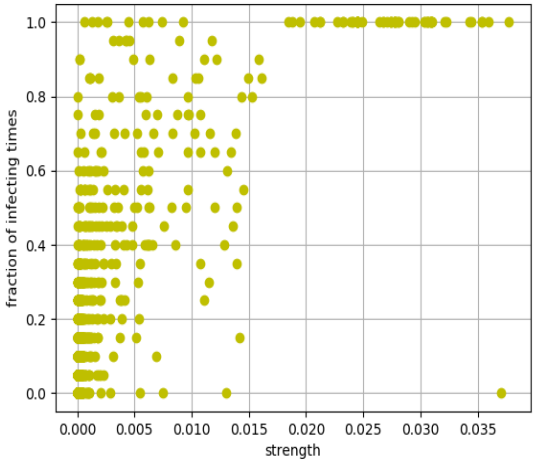
(c) PageRank weighted betweenness centrality



(d) Closeness weighted betweenness centrality



(e) Eigenvector weighted betweenness centrality



(f) Strength weighted betweenness centrality

Task 7

Even though extremely simplistic, this 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.
Here are some deficiencies of the current epidemic model and possible ways to improve them:

- Firstly, the model is a SI one, so the infected nodes stay infected forever. This aspect could be improved and made more world-like by turning it into a SIS or SIR model. The first one allows the possibility for a node to become susceptible again (so it could be infected again), while the second one considers the possibility for a node to become 'healthy' and immunized (so it cannot be infected again).

- The current model allows two possible states for each node: susceptible or infected (model 0/1). One idea could be adding a *"degree of infection"* value that scales from 0 to 1 and which represents the fraction of the infection in the considered airport. This could be a cumulative quantity that would insert a temporal window from the first landing of an infectious flight and the actual infection of the whole airport. The idea would be that each flight would carry a certain amount of infection that would contribute to the total amount of infection of the arrival airport (up to 1). Though it is true that an infection process is not linear/additive, the idea is to consider not just a 0/1 infection.

- Another deficiency is the probability of infection. Currently, the threshold is the same for each node. An idea could be to calculate for each target node a different value depending on how much is infected the departure one. In a sense it could be linked to the previous point exploiting the concept of *"degree of infection"* of each node and weighting the probability on that value, which is dynamic and not static, as it is right now.

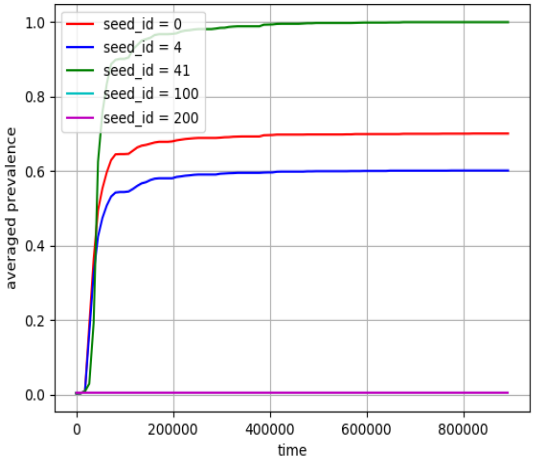
- Lastly, one could take into account other aspects of the population in order to calculate a more suitable infection probability. For instance, the size of the airport, the number of passengers for each flight and the population/size of the city (related the the considered airport) could be interesting measures to add that can characterize better the distribution. This is related to the idea that an infection may spread with higher probability in highly populated areas and crowded flights could have a higher chance to carry an infection.

Bonus Task

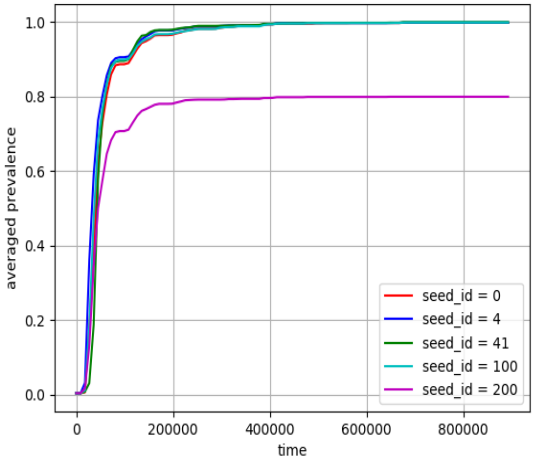
In this task I tried to simulate a variation of the SI model, implementing two different and well-know versions: SIS and SIR. I modified my code for the spreading simulation in order to include two additional parameters: *infection duration* and *model*. In the case of SIS model, once a node has been infected, after "infection-duration" seconds it becomes susceptible again. For the SIR model, after "infection-duration" seconds it becomes resistant, so it cannot be infected again. I chose to run the simulation with the same seed-ids of Task 3 (ids = 0, 4, 41, 100, 200) and a probability of infection $p = 0.5$. I ran the simulation for each one of the two models with different infection duration times. In particular, I decided to analyse the averaged prevalence $\rho(t)$ of the disease as a function of time with the following infection duration:

- 15k seconds, which are a bit more than 4 hours
- 50k seconds, which are around 13 hours (almost half a day)
- 100k seconds, which is a bot more than 27 hours (more than a day)
- 200k seconds, which is around 55 hours (more than 2 days)

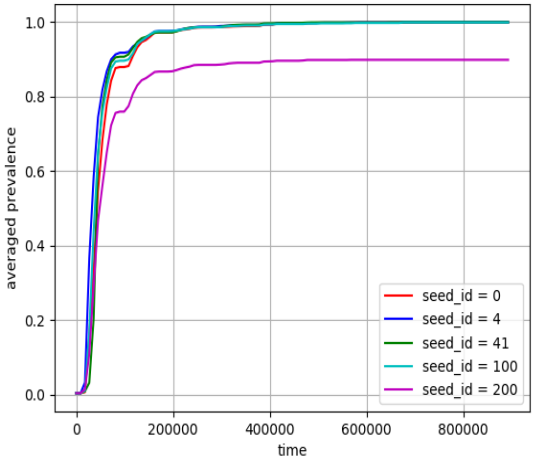
The results presented in the plots in the following pages show that the spreading speed and the quantity of nodes infected depend not only form the seed node, as already stated in the previous tasks, but also from the infection duration time. In particular, in both cases it is quite clear that the more the infection duration time increases, the more the dependency from the seed id increases and almost every curve follows the same trend. Furthermore, if we look at the difference between the two models, there are more sharp dissimilarities in the plots with lower infection duration times. This is probably because this parameter influences more the fraction of infected nodes and blocks before the spreading in the SIR model than in the SIS one. When we increase the window of time in which the nodes are infected, it is more likely that in both cases a bigger averaged prevalence is reached in both models and so there appears to be less difference between the two.



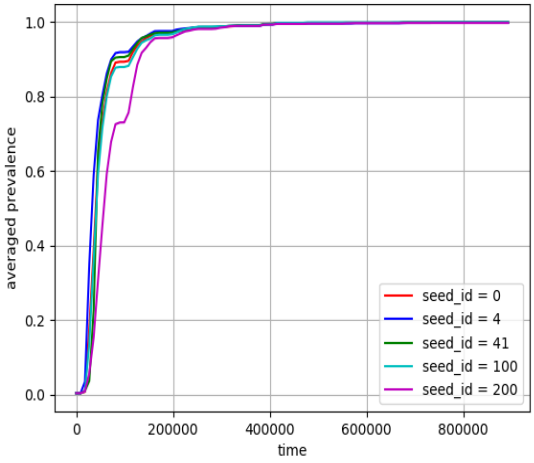
(a) $\rho(t)$ with infection duration = 15k



(b) $\rho(t)$ with infection duration = 50k

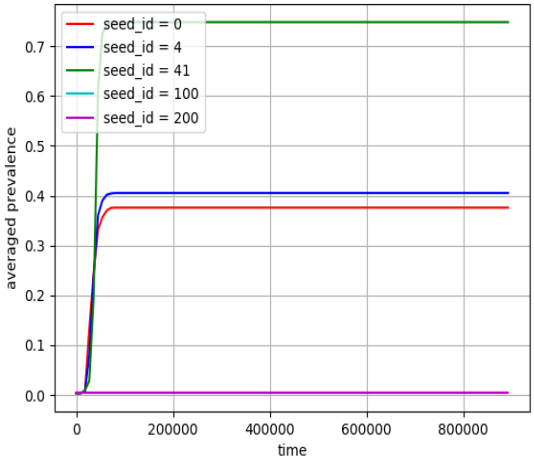


(c) $\rho(t)$ with infection duration = 100k

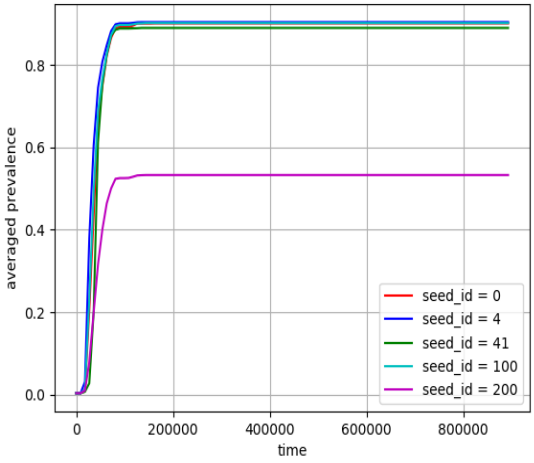


(d) $\rho(t)$ with infection duration = 200k

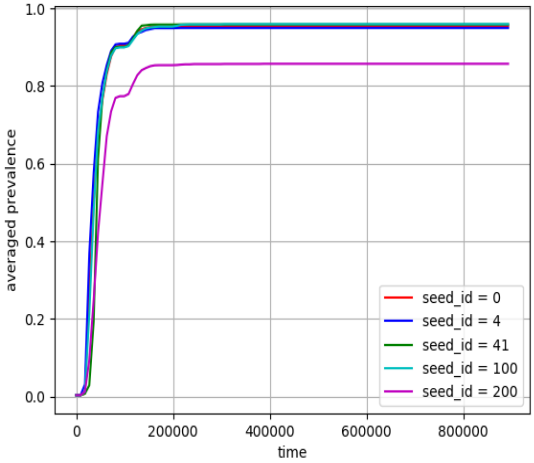
Figure 8: SIS model



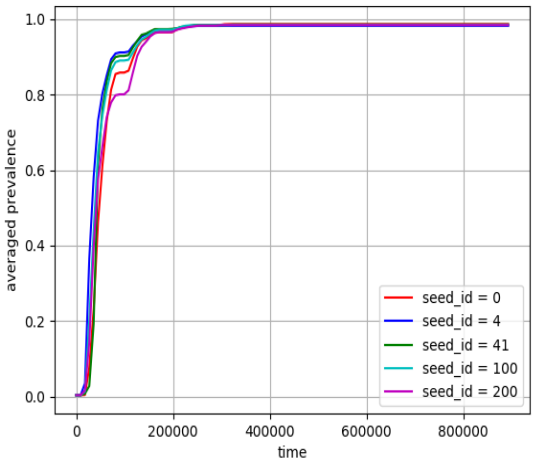
(a) $\rho(t)$ with infection duration = 15k



(b) $\rho(t)$ with infection duration = 50k



(c) $\rho(t)$ with infection duration = 100k



(d) $\rho(t)$ with infection duration = 200k

Figure 9: SIR model