CS-E5740 Complex Networks, Answers to the course project

Tommaso Brumani, Student number: 100481325

December 19, 2022

Task 1: Basic Implementation

a) If Salt Lake City (SLC, node-id= 27) is infected at the beginning of the data set, we have that Anchorage (ANC, node-id= 41) becomes infected at time $t =$ **1229283600**, which falls within the expected range of 1229283000–1229284000.

Task 2: Effect of infection probability p on spreading speed

The SI model was run 10 times for each of the infection probabilities $[0.01, 0.05, 0.1, 0.5, 1.0]$, maintaining Salt Lake City (SLC, node-id= 27) as the seed node.
Prevalence was computed at each iteration, and was later averaged for each infection probability.

a) The average prevalence for the disease was plotted as a function of time for each infection probability.
The results are reported in Figure 1.

As expected, some nearly periodic plateaus can be spotted in the curves.

b) The entire network becomes infected within the observed timeframe only for $p = 0.5$ and $p = 1.0$.
The periodic 'steps' in the curves signify that there regularly are periods of time in which the prevalence does not increase.
As this phenomenon occurs independently of infection probability or current prevalence value, we can explain it by assuming there are certain times during which fewer planes depart from airports, such as for example during nighttime or specific days of the week.

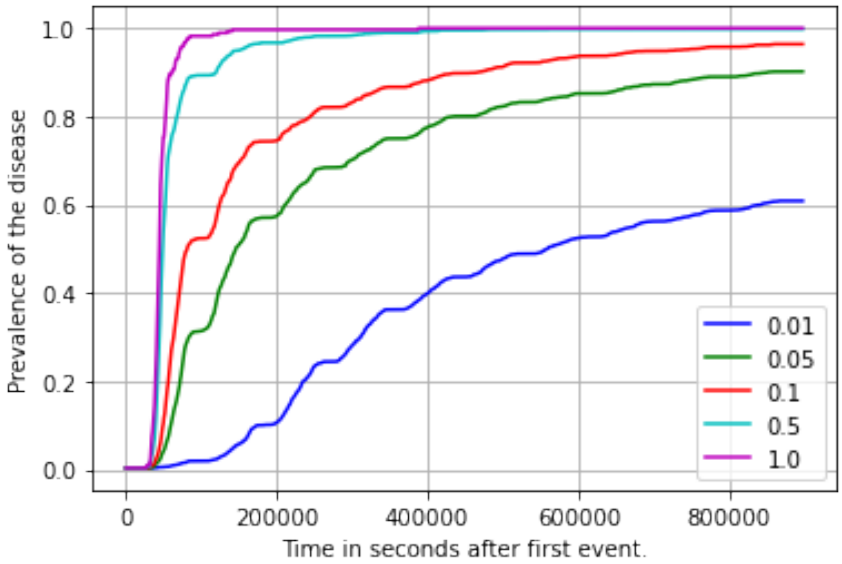

Figure 1: Averaged prevalence evolution for each infection probability.

Task 3: Effect of seed node selection on spreading speed

a) The spreading of the infection was evaluated when using different seed nodes, by setting $p = 0.1$ and running the simulation 10 times for each of the seed nodes in the list $[5, 38, 74, 134, 143]$.
The average prevalence of the disease as a function of time was then plotted for each seed, and the results are reported in Figure 2

b) The differences in spreading speeds between the different seeds are more apparent during the first 200000 seconds of the simulation.
This is likely because at the beginning the individual differences in positioning within the network of the different nodes are more important in determining the spreading of the infection (some nodes might be more central than others).
As the spreading continues, however, the overall network dynamics become dominant as all the hubs are quickly reached, thereby resulting in a similar development of the disease.

c) It is important to average results over multiple seeds because, depending on the specific task and network, 'lucky' or 'unlucky' choices of seed may result in extremely atypical conclusions.
As an example, choosing a node that is only connected with a few others might cause those nodes to be infected much sooner than those farther away, even though they might be less likely to be infected quickly on average.

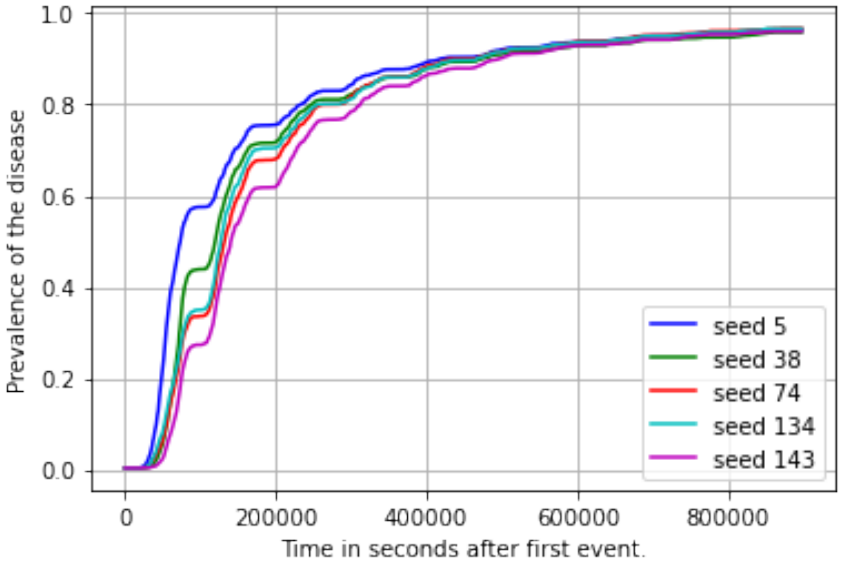

Figure 2: Averaged prevalence evolution for each seed node.

Task 4: Where to hide?

The infection was simulated 50 times with $p = 0.5$, using different random nodes as seeds each time and recording the median infection times for each node.

a) The median infection times for each node are reported in the scatterplots in Figure 3 as a function of different nodal network measures (unweighted clustering coefficient c, degree k, strength s, and unweighted betweenness centrality bc).
For those nodes that did not get infected in more than half of the iterations, the median infection time was manually set to the largest median time observed in the various simulations.

b) The Spearman rank-correlation coefficient was calculated between the computed median infection times and the node's network measures to estimate which measure is the best predictor for the infection times. The results are reported in Figure 4, from which we can surmise that degree and strength appear to be the most promising quantities to observe, as they are strongly negatively correlated with the median infection times of nodes.

c) - Based on the results obtained, a good choice for a place to hide from the infection would be a node with very low strength or degree, possibly even zero.
This is (intuitively) because an airport that is connected with few others, and that receives few flights from them, is less likely to be reached soon by the spreading disease (unless, of course, it is its source).

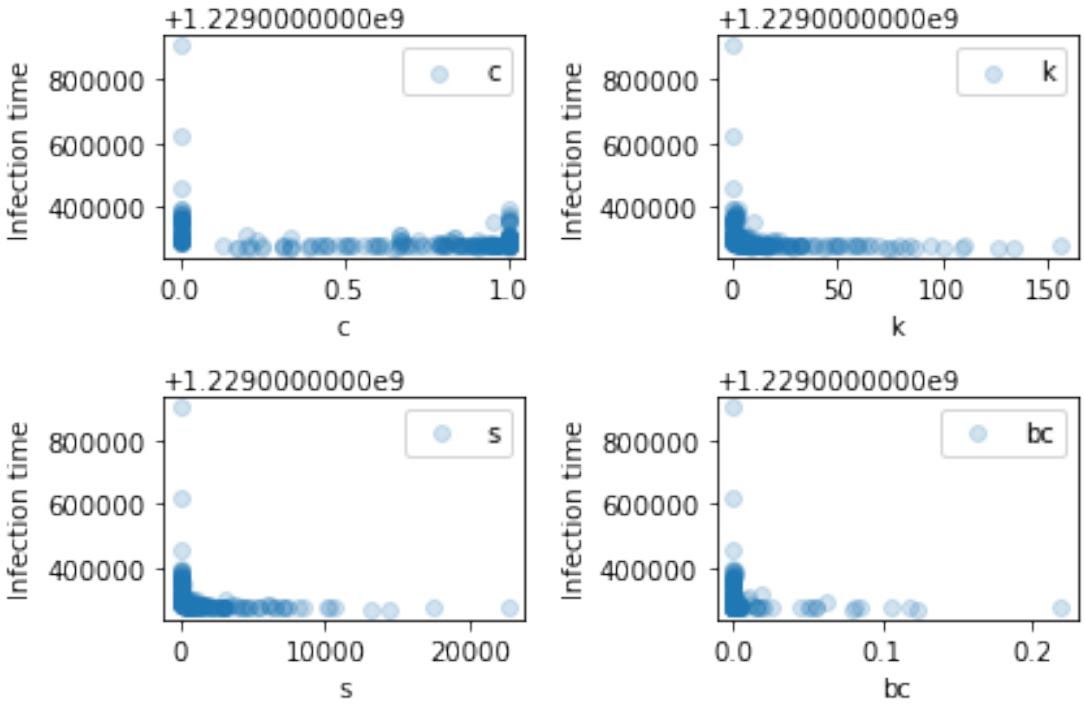

Figure 3: Median node infection time as a function of c, k, s, and bc.

```
Spearman r between median infection time and c: -0.13945645629907016
Spearman r between median infection time and k: -0.823412202366824
Spearman r between median infection time and s: -0.896892756970813
Spearman r between median infection time and bc: -0.6468527487894815
```

Figure 4: Spearman r coefficient for median infection times wrt. c, k, s, and bc.

In this regard, strength might be a better measure of the node's likelihood of infection because it contains more information regarding the node's exposure to the disease (not only the number of connected airports but also the frequency of flights between).

- Betweenness centrality behaves differently than degree or strength because, rather than being a good measure of how quickly a node is likely to be infected, it is a good measure of how important it is in the spreading dynamics of the disease.
  A node with low betweenness centrality could still be infected very quickly, on average, if it was simply neighboring other nodes with high betweenness centrality.

- Clustering coefficient is a poor predictor of median infection times because both nodes that are very well connected (and thus more exposed to the disease) and nodes that are mostly isolated may have comparable c values, so long as their neighbors are connected with each other.

Task 5: Shutting down airports

a) A number of immunization strategies were tested by running 20 simulations for each one, using $p = 0.5$ and randomized seed nodes (although the same node was kept for the same iteration across different strategies to reduce variance).
An immunized node is unable to be infected throughout the entirety of the simulation.
The 6 adopted immunization strategies were:

- Immunizing the 10 nodes with the highest c value.
- Immunizing the 10 nodes with the highest k value.
- Immunizing the 10 nodes with the highest s value.
- Immunizing the 10 nodes with the highest bc value.
- Immunizing 10 randomly selected nodes.
- Immunizing one randomly chosen neighbor for each of 10 randomly selected nodes.

The average evolution of prevalence in the network as a function of time was calculated for each of the immunization strategies.
The results are reported in Figure 5.

b)
- The best-performing immunization strategy is the one immunizing the 10 nodes with the highest betweenness centrality.
  This is likely because, as explained above, betweenness centrality is a good indicator that a node quickens the spreading of the disease between other nodes, as a high value corresponds to a node that is in the shortest path between many other pairs of nodes.
  If such a node is immunized, the disease will surely have to take other, longer paths to reach some of the susceptible nodes.
  In contrast, other strategies like the ones based on degree and strength, immunize nodes that are likely to be infected quickly (because of their high number of connections and/or flights) but are not necessarily going to slow down the infection.
  Nevertheless, a node with a high degree or strength is more likely than average to be the one spreading the disease between two nodes (and more likely to have a higher betweenness centrality), and thus results in improved performance

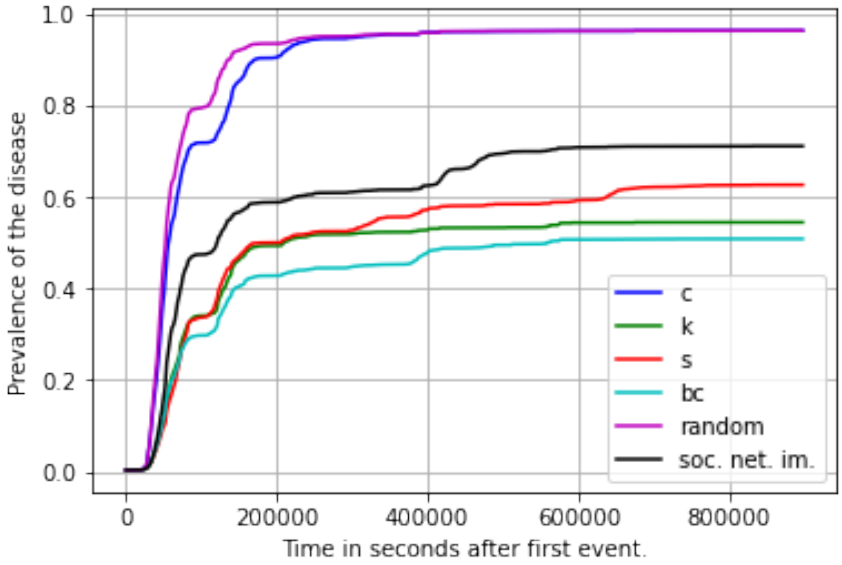

Figure 5: Averaged prevalence evolution for each immunization strategy.

compared to random immunization.
Moving on, the strategy of immunizing nodes with a high clustering coefficient expectedly proves very ineffective: a node with a high clustering coefficient will have many neighbors connected with one another, meaning it is quite likely that its immunization will have barely any effect on the spreading of the disease.
Finally, the pick-a-neighbor strategy performs discretely: while it does not specifically target nodes with high centrality or number of connections, the friendship paradox (a node's neighbor on average having a higher degree than the node itself) will result in this type of strategy having a tendency to immunize hub nodes, thus approximating the performance of the method based on node degree.

- Betweenness centrality performs better as an immunization strategy than a predictor for hiding places because it indicates whether a node is involved in many shortest paths for the spreading of the disease rather than whether it is more likely to be infected quickly or not.
  For example, as mentioned before, a node connected exclusively to another with high betweenness centrality would have an extremely low *bc* value, but would nevertheless find itself on average at a very short distance from the infection's seed (and would thus be infected quickly).

c) As expected, the pick-a-neighbor immunization strategy performed better than the random node immunization.

- The probability of picking a random node of degree k from the network is

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

where N is the number of nodes in the network, N_k is the number of nodes of degree k, and $P(k)$ is the network's degree distribution.

- Picking a random neighbor of a random node is then going to result in a probability of picking a node of degree k equal to:

$$p(k_{nn} = k) = \frac{kNP(k)}{N < k >} = \frac{kP(k)}{< k >}$$

- Consequently, a node's neighbor will, on average, have degree

$$< k_n n >= \frac{< k^2 >}{< k >}$$

which is higher than a random node's ($< k >$), meaning, choosing neighbors as immunized nodes will on average result in choosing nodes with a higher degree, thereby resulting in a slower spreading of the disease.

Task 6: Disease transmitting links

Over the course of 20 iterations with $p = 0.5$ and randomized seed nodes, the node responsible for transmitting the disease to each node in the network was recorded.

a) The fraction of iterations where each link in the network was the one responsible for transmitting the disease between its two nodes was computed.
The network was plotted over the map of the USA by weighing each link based on its fraction of viral iterations, and the results are reported in Figure 6.
For comparison, the maximal spanning tree of the network was also plotted over the same map, and the results are reported in Figure 7.

b) The visualization of the two networks is similar because, by definition, the maximal spanning tree contains the edges with the highest weights.
Since a link with a large weight corresponds to a connection between two airports that is traveled by a high number of planes, these links are also more likely to be the ones transmitting the disease, and thus will have a stronger coloration on the transmission frequencies map.

c) Scatter plots were computed showing the frequency of transmission for each edge as a function of link weight and unweighted link betweenness centrality, and are reported in Figure 9.

The Spearman correlation coefficients between the transmission frequency and these two quantities were also computed for each link, and are reported in Figure **??**.

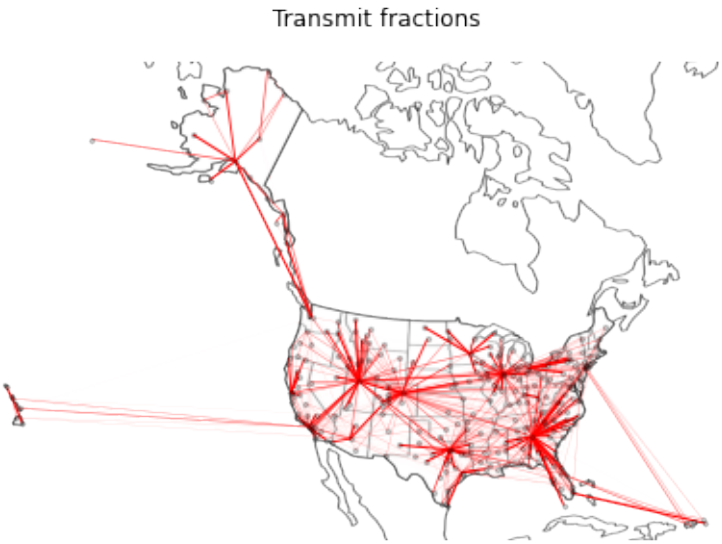

Figure 6: Network with weighted links plotted over a map of the USA.

Figure 7: Maximal spanning tree of the network plotted over a map of the USA.

d) Edge betweenness appears to be somewhat positively correlated with f, whereas for the weight this correlation seems to be weaker.
This is likely because while a link with a larger weight is more likely than one with a smaller one to be a vector of spreading for the disease, the betweenness centrality of the node implicitly determines how probable it is for that node to be on the shortest path between any two others, and thus, to be activated during the simulation before

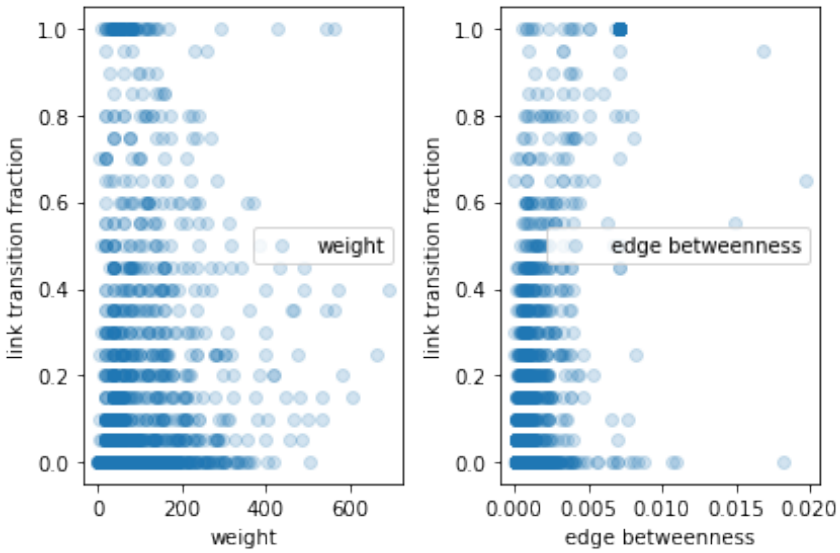

Figure 8: Scatterplots of link frequency of transmission as a function of w and bc, respectively.

```
Spearman r between link transmission fraction and weight: 0.31542302562171826
Spearman r between link transmission fraction and edge betweenness: 0.504116536
1335789
```

Figure 9: Spearman correlation coefficients between f, w, and bc.

other links infect its nodes.