Modeling and Data Analysis in Complex Networks

Ajaya Adhikari - Lorenzo Gasparini - Ioannis Papadopoulos
Group 20

Introduction

In this assignment we will analyse a dataset about the mitigation of infectious disease at school. We consider a temporal network, hence the links between the nodes can change at each timestep. Each row of the dataset denotes a link between two nodes at a certain time step. We have used the framework Networkx and the programming language Python for the network analysis. First, we will explore the topological features of this network, without considering the temporal aspect i.e. the network is aggregated over all timesteps. Second, the temporality of the network will be considered. In this case a lot of interesting aspects can derived, such as how influential a node is as a seed to propagate a disease throughout the network.

Topological features

Number of nodes: N	242
Number of links: L	8317
Link density: p	0.285
Average degree: E[D]	68.74
Degree variance: Var[D]	708
Degree correlation (assortativity)	0.118
Meaning: Positive value indications that there is a correlation between nodes of similar degree, while negative values indicate that there is a correlation between nodes of different degree.	
Clustering coefficient: C	0.526
Average shortest path between all node pairs: E[H]	1.732
Diameter: H_{Max}	3

Spectral radius	80.246
Algebraic connectivity	130.557

Degree distribution

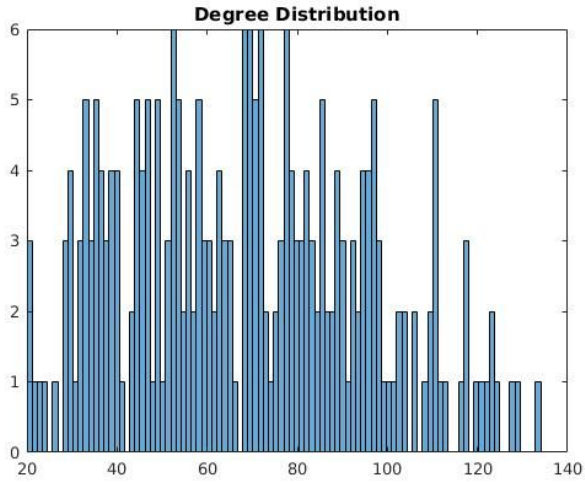

Figure 1: Degree distribution.

This network cannot be modelled by scale-free networks because scale-free networks degree distribution follows a power law. This is clearly not the case for our network (figure 1). Typically, the degree distributions of random graphs are similar to the normal distribution. The degree distribution graph visualized on figure 1 is more similar to a normal distribution than a power law, hence random networks can better model our network.

Small-world property

According the Watts–Strogatz model, small-world networks have short average path length while highly clustered. In our case the average path length is indeed short (1.7), and the clustering coefficient is not very high (0.52). We suspect that our network has the small-world property on a moderate level.

Temporal networks

Infection spread over time

The simulation of infection spread is done by infecting one node at the first time step. At each timestep the neighbors of the infected nodes also get infected. The number of infected nodes per timestep is saved. This whole process is done for each node as the initial infected

seed. The average over all these processes is visualized in figure 2 and the standard deviation of this infected nodes per time step is shown in figure 3. The execution time for the whole simulation with the resulting mean and standard deviation was 5 minutes.

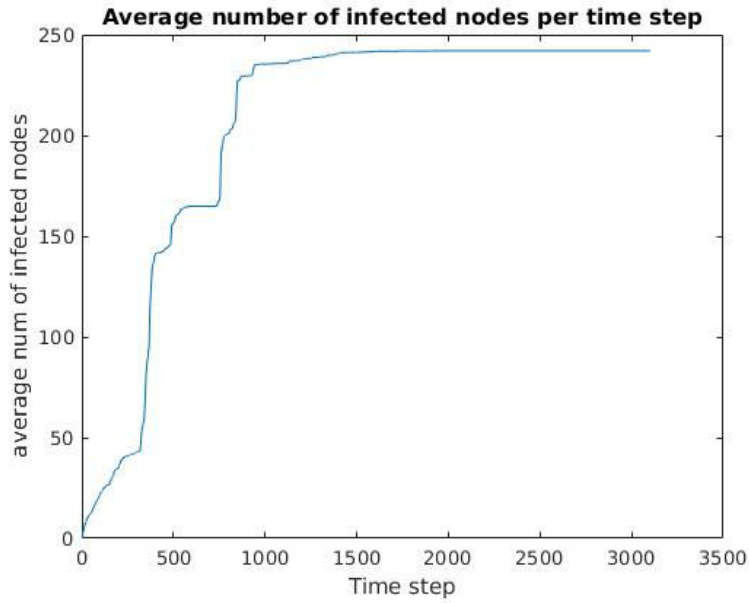

Figure 2: Average number of infected nodes per time step

Figure 2 and 3 shows that there are only about 3000 time steps, while the assignment states that the time steps go from 0 to 5846. The data does not show this. After going back and looking at the data carefully, we found out that there is a big jump from timestep 1555 to time step 4302. Hier, we considered this step as one time step.

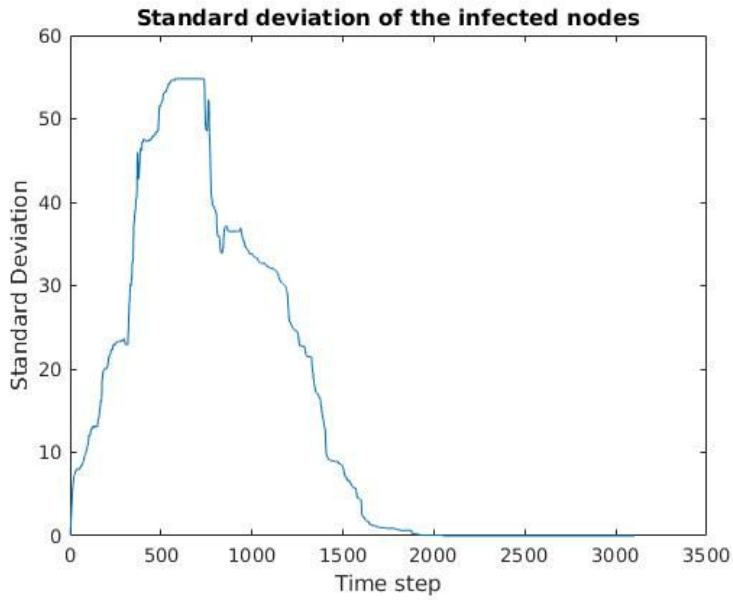

Figure 3: Standard deviation of the infected nodes per time step

Influence

We computed the influence of each node by calculating the time it takes to infect 80 % of the total nodes when this node is selected as the seed node. A node is influential if its time is short. We created a sorted list of all nodes according this time. Figure 4 shows the distribution of the time step when 80% of the nodes are infected with different nodes as seed. Node 182 is most influential, it need 49 time steps to infect 80% of the nodes. Most of the nodes need around 200 time steps, while three nodes (234, 151, 146) need more than 4600 time steps.

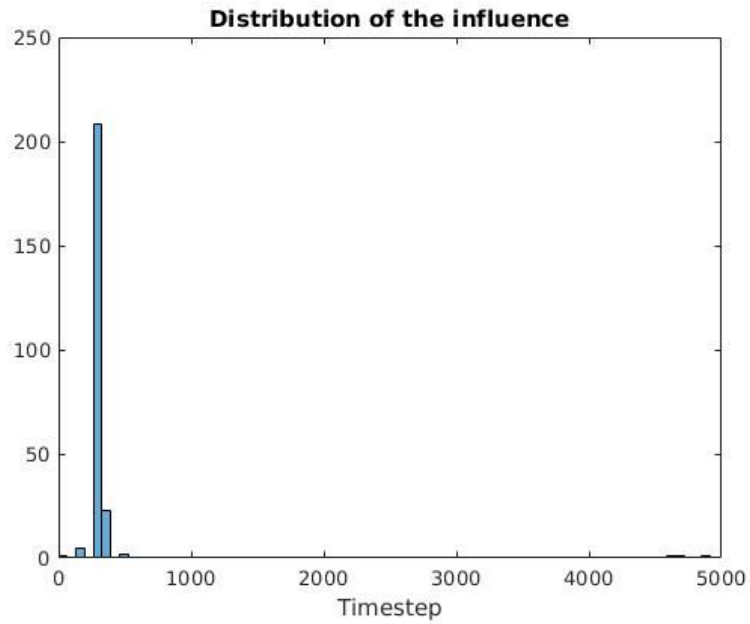

Figure 4: Distribution of influence

Degree and clustering coefficient

We are going to analyse whether the degree and clustering coefficient of each node in the aggregated network is correlated with the influence of the node. Two sorted list containing the nodes have been created and the nodes are sorted in descending order according to the corresponding coefficient. The similarity between the influence (R) and degree (D) coefficient is computed with the following formula. Only the highest f percent of the nodes are considered in the formula.

$$r_{RD}(f) = \frac{|R_f \cap D_f|}{|R_f|}$$

Figure 5 visualizes the similarity between the influence versus the degree and clustering coefficients. It is clear that the degree coefficient is more explanatory than the clustering coefficient.

When a node has lot of neighbors, they get infected at one step, while a node with high clustering coefficient might not have a lot of neighbors, hence it will take more time steps to infect the cluster. We suspect that this explains the difference on the graph.

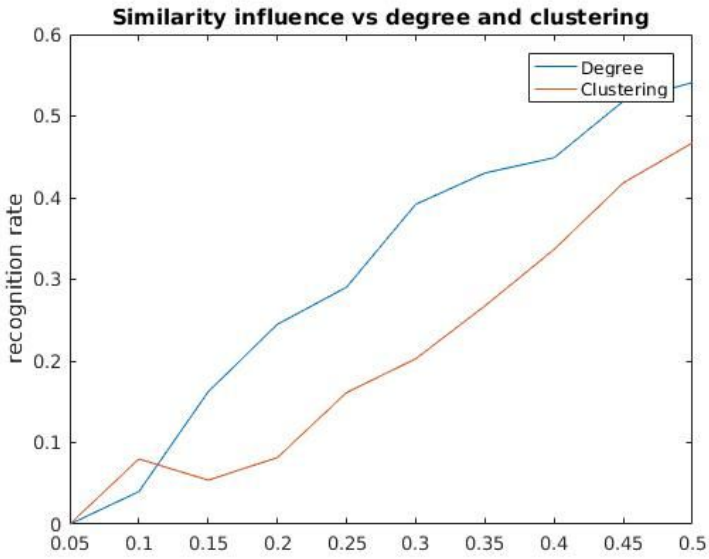

Figure 5: Similarity influence vs degree and clustering coefficient