GTDM Fourth Assignment

Group 2

December 15, 2021

Bertoletti, Luca — Brueckner, Xaver Matthias Heinrich
Ly Nguyen, Thi Phuong — Funicello, Alfredo
Degl'innocenti, Giorgio — Luraschi, Matias Santiago

Instructions

Consider the graph stored in the file *graph2.gml*, containing a sample of a population composed by 70 persons. For each person the age, the gender, and the name (anonymised, identified by a number from 1 to 70) have been registered. The persons are forming the nodes of the graph and there is an (unoriented) edge between two nodes if the two persons are used to spend more than 5 hours per week together, in person or on social media, videoconference, etc.

Identify if are there communities in the graph, and analyse if the members of each community have some common characteristics. Are there any hub nodes, that is any node with a particularly big number of connections to the others?

Imagine now that a fake news spreads in the population represented by your graph, starting from one single person, that we consider 'infected by the fake news' at time 0. At each time step, each non infected person vi becomes infected (that is receives the fake news) with probability

$$P(\textit{infection of vi at time } t+1) = \begin{cases} (0.2) \cdot n_i(t) & \text{if } n_i(t) \leq 5 \\ 1 & \text{otherwise} \end{cases} \tag{1}$$

Where $n_i(t)$ is the number of infected neighbours of vi at time t.

Are you able to simulate the spread of the fake news in the population? Is there any difference in the mean speed of the spread if the infection starts from each of the identified communities?

1 Identification of communities

Each node has 3 characteristics:

1. Age
2. ID
3. Gender

The Figure 1 is a visualization of the graph we're working with.

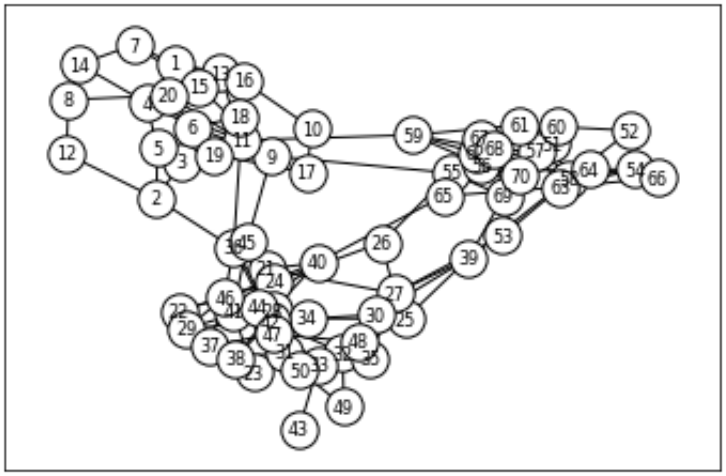


Figure 1: Visualization of the graph from the *graph2.gml* file

Using the *Girvan-Newman* algorithm, which is based on the concept of *betweenness of edges*, we find that the graph has 3 communities, shown in Figure 2.

Two communities have 20 members, while the other has 30.

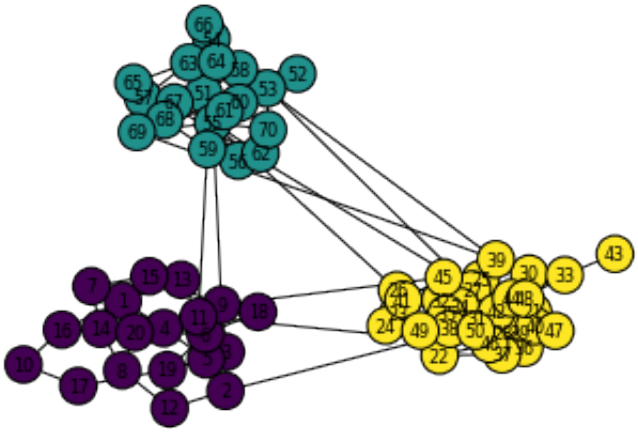


Figure 2: Graph showcasing the sub-communities

Through the SimRank matrix, plotted as a heatmap in Figure 3, we can discern the pattern of similarity in age between the 3 groups.

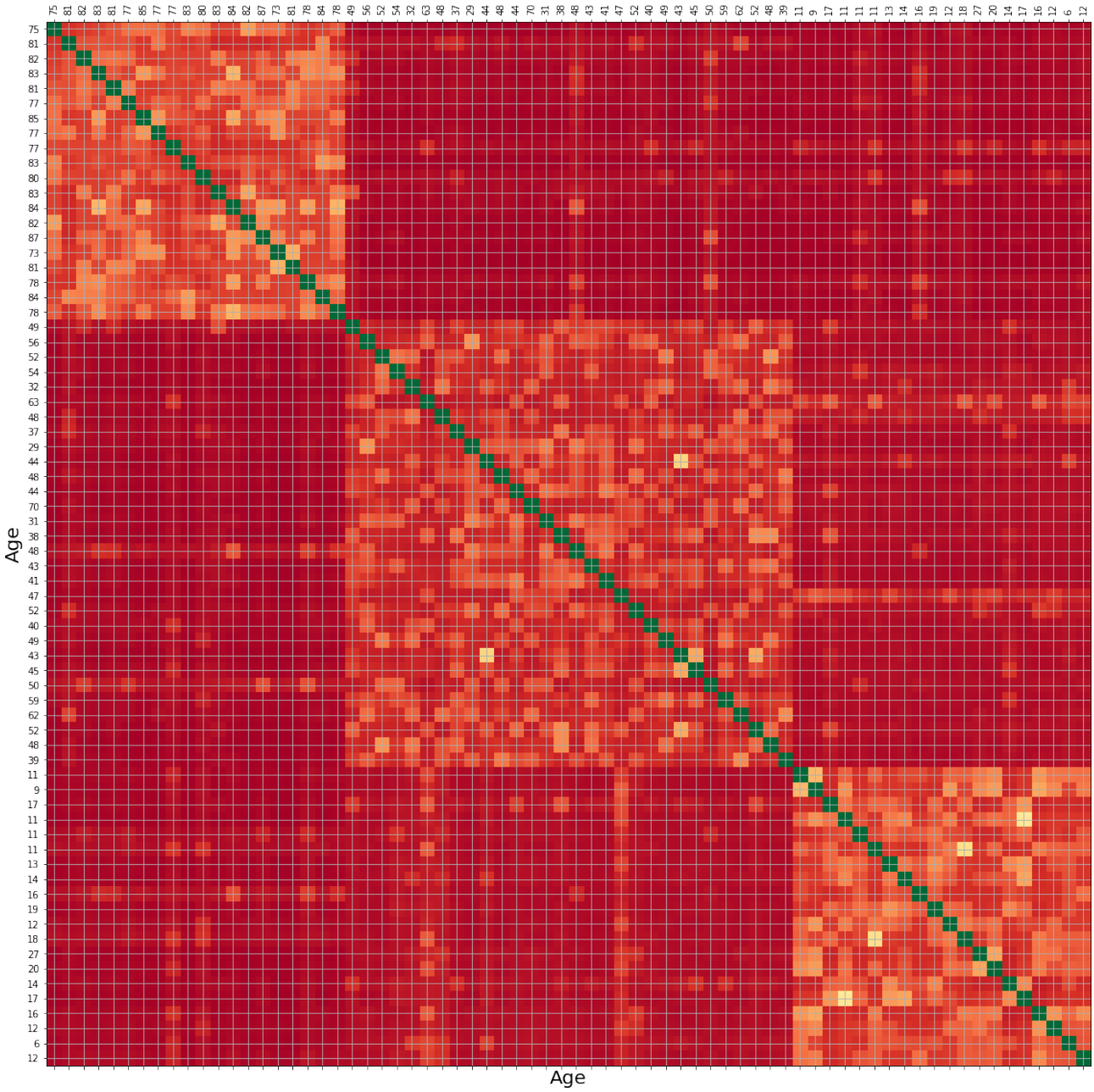


Figure 3: SimRank matrix plotted on Age feature of the nodes in the graph

We can see that one of the community is composed by people in the age bracket 71-90, the last one 30-70, the third one 0-20. The Figure 4 shows the age distribution between communities.

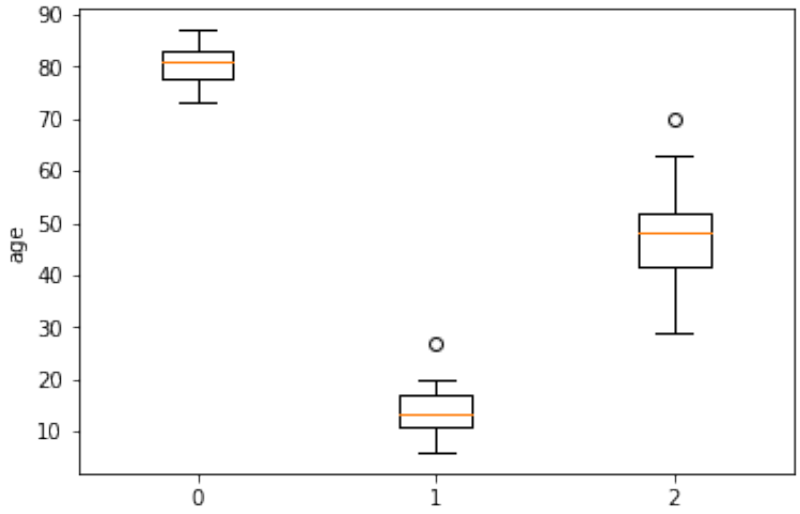


Figure 4: Age distribution between groups

For the other characteristics ID and Gender no pattern could be found. ID is just a anonymized proxy for the name and the gender is evenly distributed in each of the three identified communities.

Another interpretable characteristic could be the amount of edges per node [called average degree from here on] in each of the communities, since it is an indicator for how interconnected the individual communities are.

Average degree per community

0: 3,95

1: 4,25

2: 4,13

From this analysis we can deduct that a younger average age in a community results in a higher average degree. This makes sense intuitively since younger people generally spend more time amongst peers in a similar age e.g. in school, sport clubs, etc. compared to older people aged 70+.

2 Hubs

Through the degree classification of the nodes we found 4 *influencers* in the graph:

11: with 7 neighbors

27: with 7 neighbors

55: with 9 neighbors

58: with 7 neighbors

In Figure 5 is shown the degree distribution of the nodes

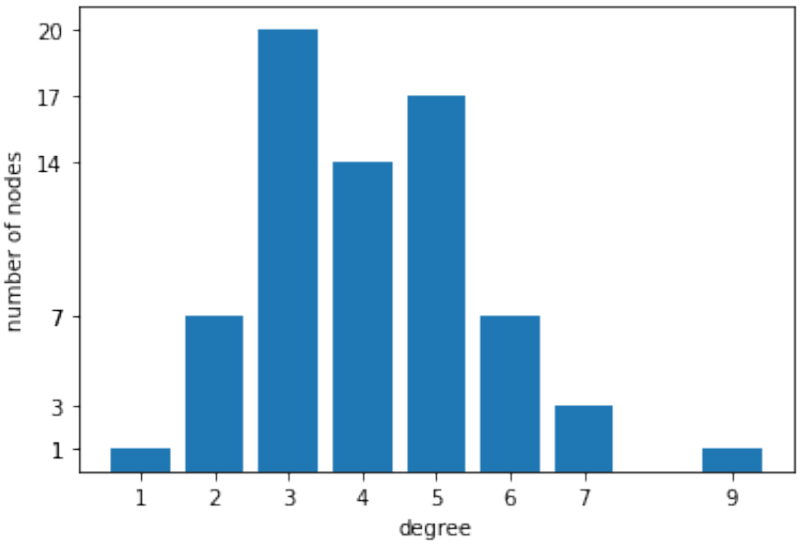


Figure 5: Degree distribution between nodes

We also computed the *closeness centrality* of the nodes, defined as:

$$C(u) = \frac{n-1}{\sum_{v=1}^{n-1} d(v,u)},$$

where $d(v,u)$ is the shortest-path distance between v and u, and n is the number of nodes in the graph. Higher values of closeness indicate higher centrality. The highest value nodes are:

21: 0.334%

27: 0.341%

11: 0.345%

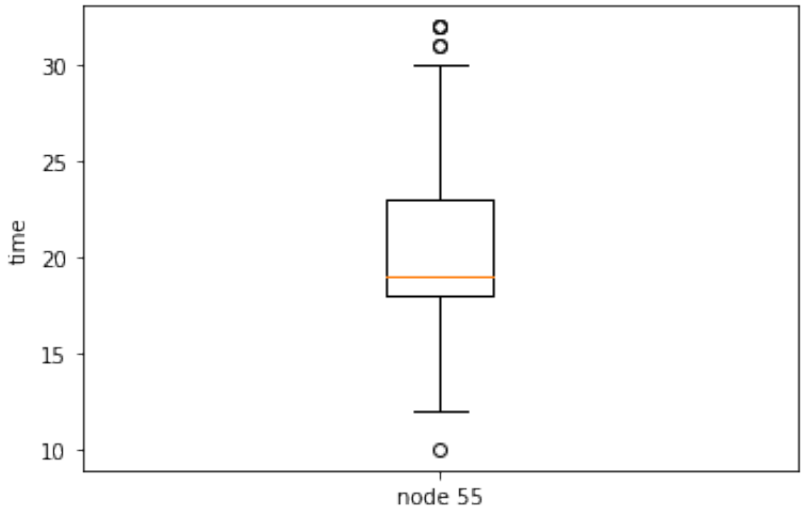
9: 0.350%

55: 0.352%

3 Spreading fake news

We simulated the spread of a fake news in the social graph.

We used as first patient0 the node 55 which is the node with highest degree and highest centrality, achieving an average of 20.37 steps needed to infect the whole graph with a standard deviation of 4.156.



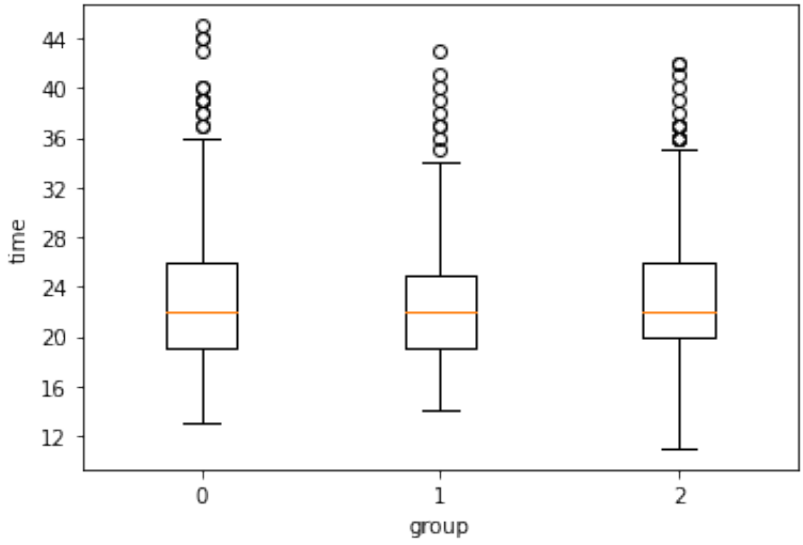
We then computed with 2000 iterations runs of spreading by selecting the patient0 randomly between the different communities. From this we received the average time steps needed on average per community.

The results are:

Community 0: 22.94 steps with a *std* of 5.21

Community 1: 22.40 steps with a *std* of 4.43

Community 2: 23.16 steps with a *std* of 4.92



From this we can see that even though the average time steps needed are similar for all three communities with a relatively high standard error, community 1 gives us the lowest average of the three which makes sense considering our earlier observation showing that community 1 is the most interconnected of the three based on average degree resulting in a faster average spread of the fake news.