GTDMO - 4th assignment

Group 6

October 23, 2022

1 Problem definition

Consider the graph stored in the file graph1.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. Imagine 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

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

where ni(t) is the number of infected neighbours of vi at time t.

1. Identify if are there communities in the graph, and analyse if the members of each community have some common characteristics.
2. Are there any hub nodes, that is any node with a particularly big number of connections to the others?
3. Are you able to simulate the spread of the fake news in the population?
4. Is there any difference in the mean speed of the spread if the infection starts from each of the identified communities?

2 Solution

Our graph $G = (V, E)$ consists of $|V| = 70$ nodes and $|E| = 140$ edges.

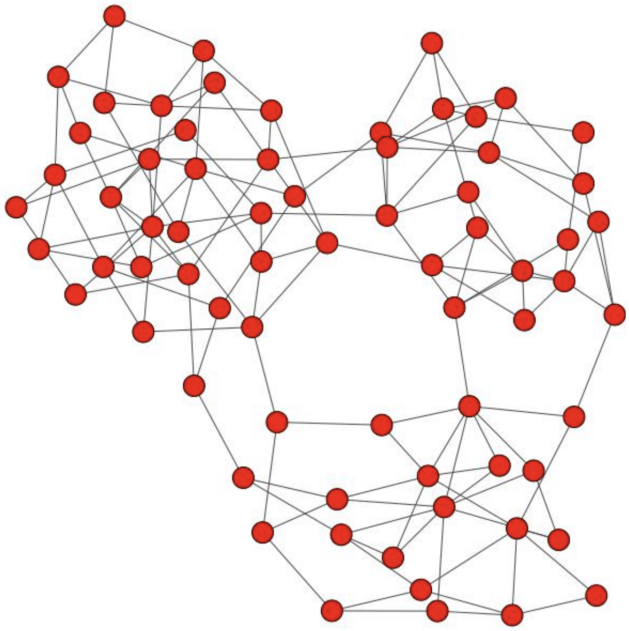

Figure 1: Net graph representation of people relation by gender

For each individual we have three main attributes which are: name, age and gender, observing this initial row representation we notice that people seem to be clustered into 3 different communities, however when we introduce the gender characteristic in the graph we notice that we have the same number of men and women, but it is not possible to assume any cluster by gender.

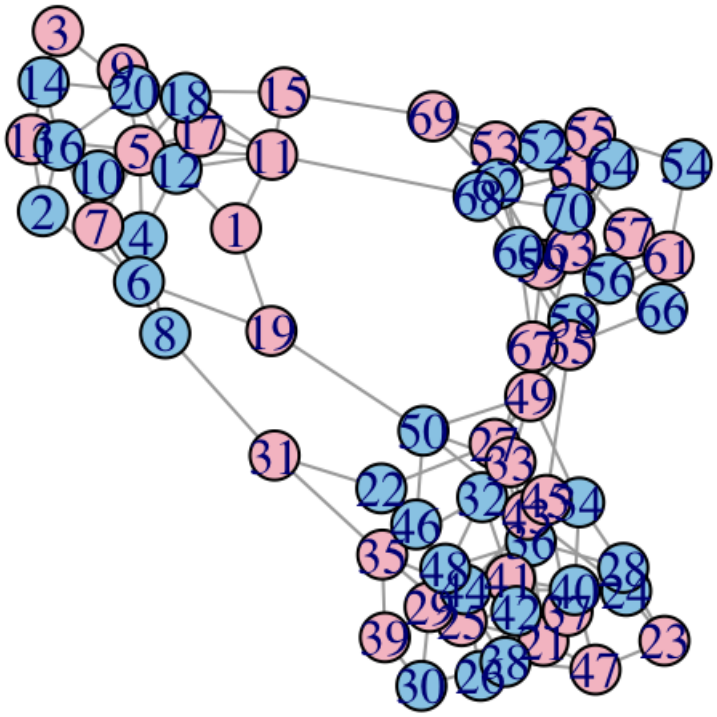

Figure 2: Net graph representation of people relation by gender

Instead, when we introduce the characteristic of age we notice that we have a well defined division in three different clusters.

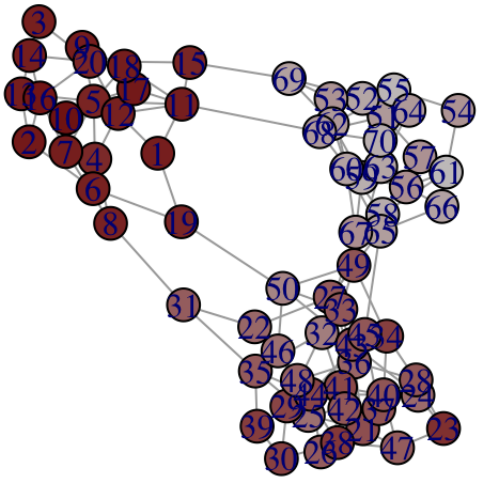

Figure 3: Net graph representation of people relation by age

looking to the cross section data divided into ten-year tracks, on Graph 4, it is possible to conclude that our sample is formed by younger men, identify a higher concentration of young men

and a bigger number of women with more than 41 years, increasing the women age to 36,2 years old versus 35,1 years old for men's average age.

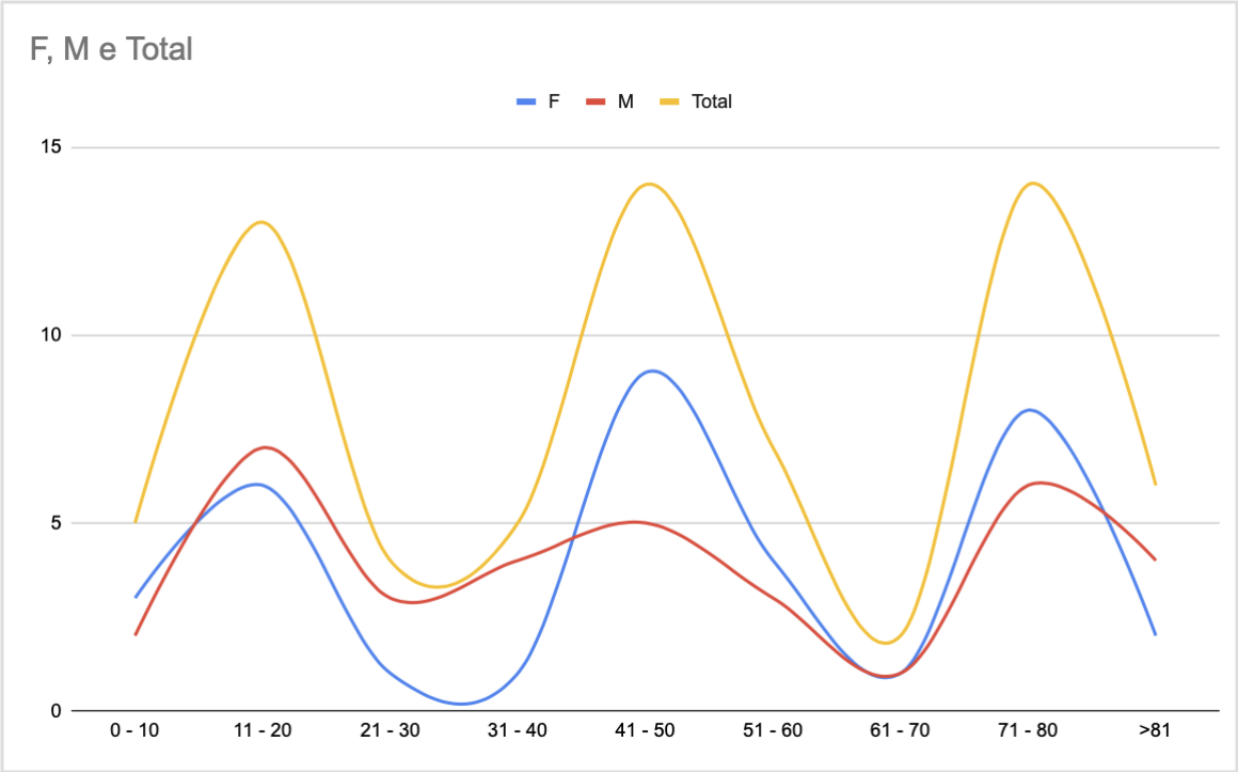

Figure 4: Distribution on men and women by age groups

Analyzing the modularity of each cluster, as shown in Graph 5, it is possible to identify that actually people are not clustered into 3 groups, but 4.

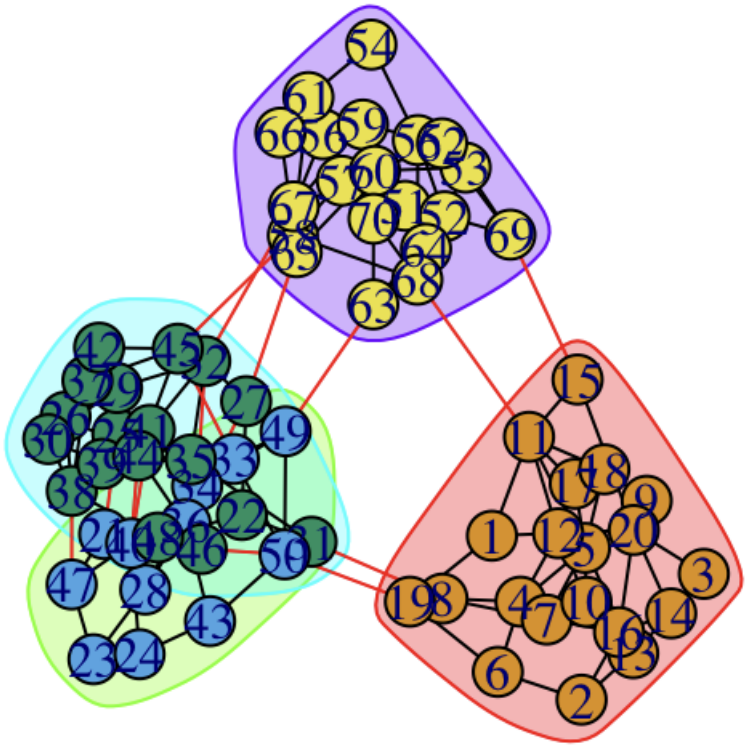

Figure 5: Net graph representation of people relation, cluster distribution

This representation suggests that it is possible to divide the bottom group into different clusters according to the *betweenness method*. The maximum of the modularity function is 0.6150596 and our graph is composed of four different subgraphs.

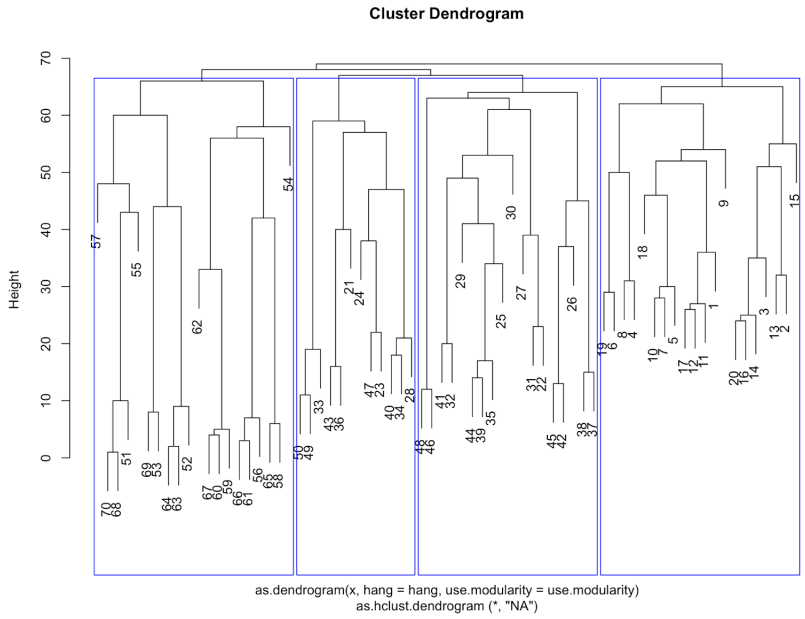

Figure 6: Cluster dendrogram

Here we can clearly observe the main division in two different opposite groups, one on the left and the other on the right side, and a division of the middle group in two clusters, forming with this approach the four groups observed on Graph 4, the key division factor of the middle group is age. That can be also written as the preference for a network's node to attach the others that are similar, i.e. the *assertativity*. Now we can denote each group, based on Graph 4 color code as:

- Mature community 1 - green
- Mature community 2 - blue
- Eldest community - purple
- Teen community - red

To understand the Fake news spread for the population we use the definition of the function 1:

- $V_1 = \{11111111111111111111232233323333223233323323333232244444444444444444444\}$

The groups consist of 20, 11, 19, 20 individuals respectively. Now we should individuate which are the top 3 nodes with more neighbours, that will be a useful information for the second part of the problem, 'cause they will spread the fake news faster than others. the top 3 nodes with more neighbours are the $\{5, 11, 44\}$ with 9, 8 and 9 neighbours respectively. While the closest ones to the center of the graph are $\{45, 58, 68\}$: closeness centrality measures how many steps is required to access every other vertex from a given vertex.

2.1 Spreading the fake news

In the second part of the problem, we analyze the spreading of fake news in the population. As we have seen, our community can be split into four smaller ones. We know that groups are mainly

characterized by their age, and we found some important nodes that will increase and decrease the information flow.

We can now compute some simulations, starting from each of these nodes and analyzing which one can make fake news more viral. to represent the distribution of the fake news for all four communities, we create a simulated spread system with the parameters related to Function 1, initializing the fake news spread for 1 single case considering three neighbours with the same probability. The simulation results for cluster 1 are presented in Graph 7:

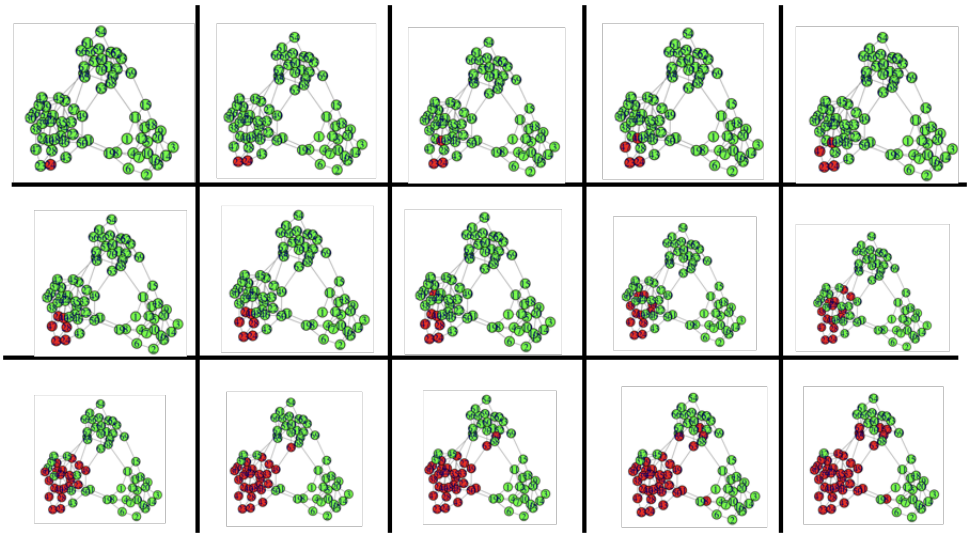

Figure 7: Fake news spread evolution cluster 1 (Red believe, Green do not believe)

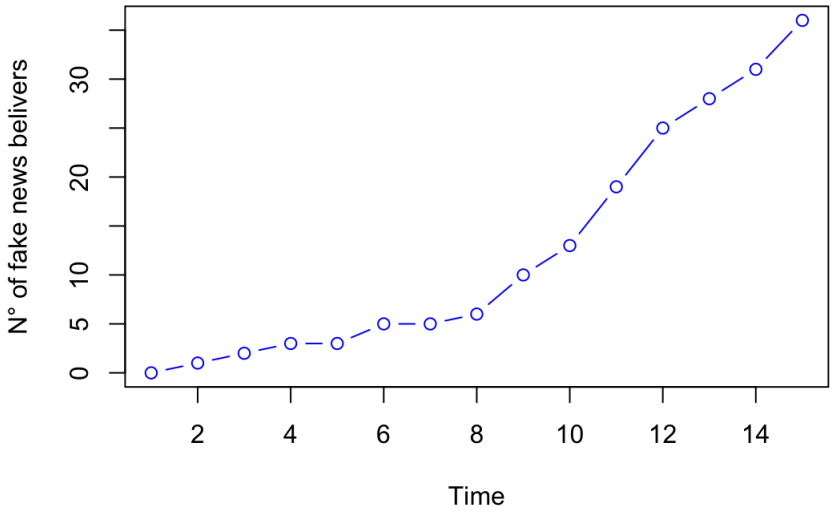

Figure 8: Fake news evolution 1 by iteration

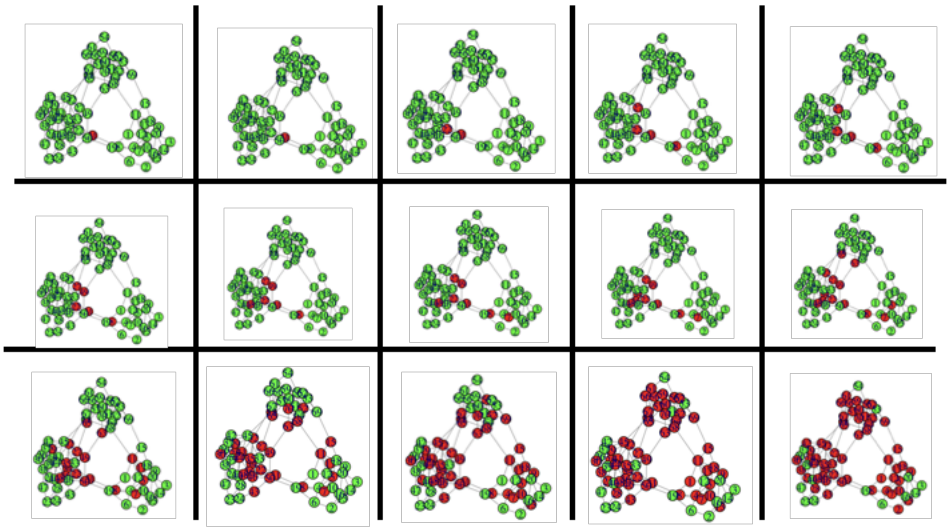

Figure 9: Fake news spread evolution cluster 1

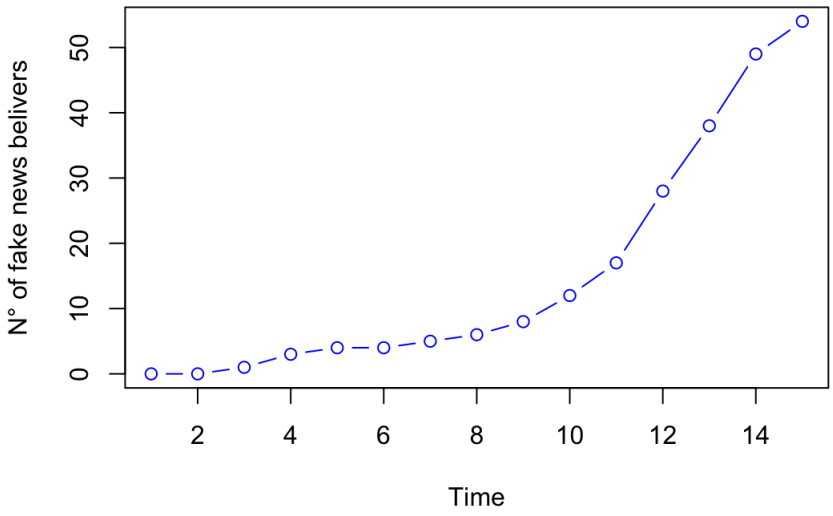

Figure 10: Fake news evolution 2 by iteration

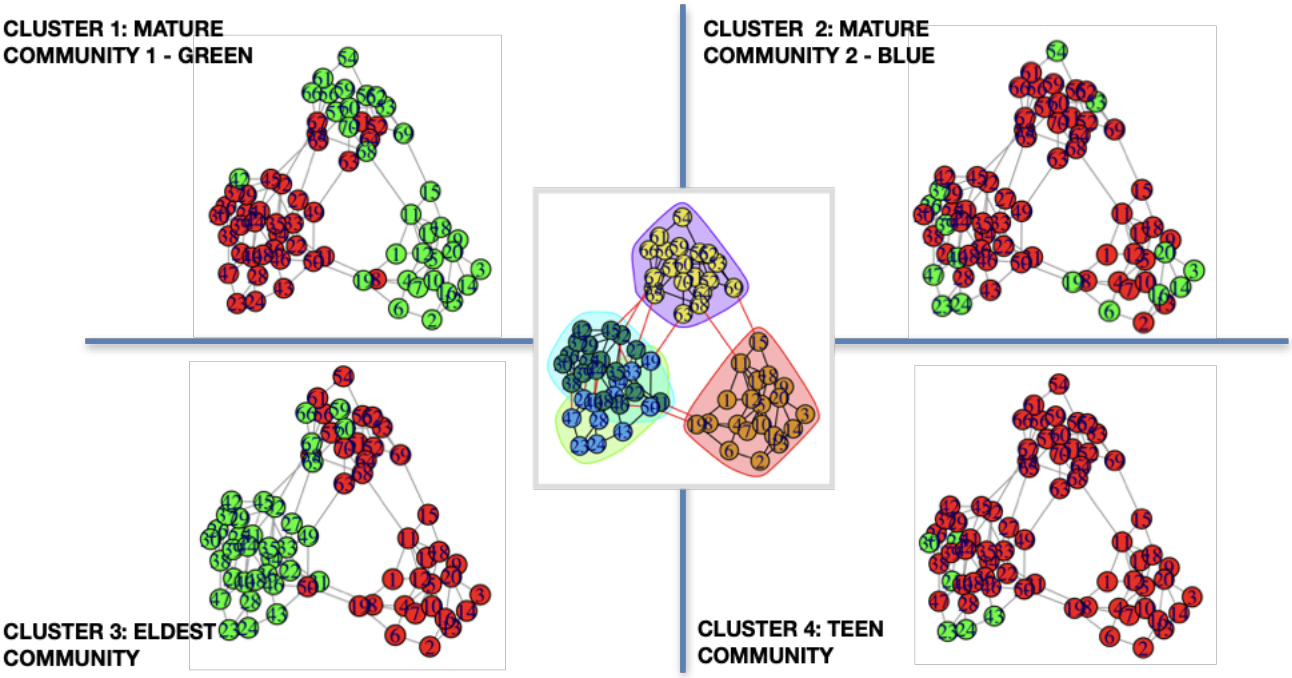

Figure 11: Diffusion of the Fake new starting from different clusters

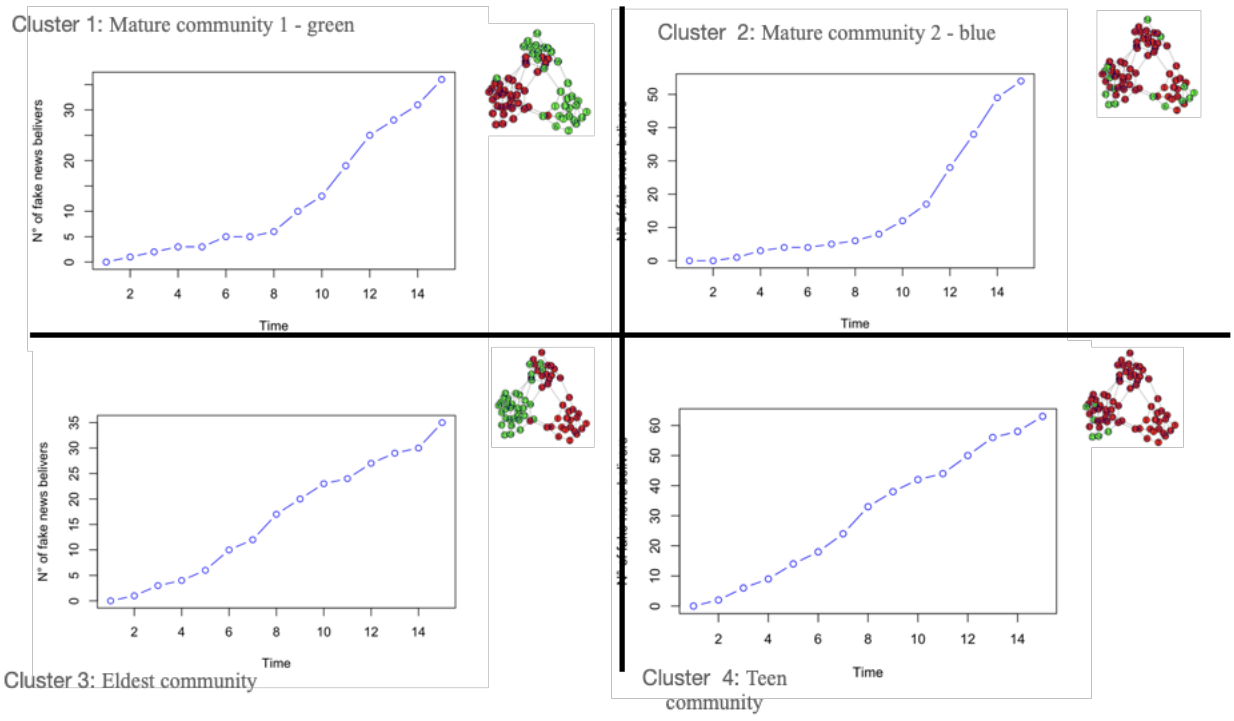

Figure 12: Cluster Comparison spread speed

The conclusion about characteristics in these groups are similar to the one of the chapter above: people usually interact mostly with other people of a similar age, as we can see from the graphs. Eldest, Mature and Mature 2 clusters spread fake news slower because centrality index distribution is lower than group of Teenagers which in fact creates the most widespread situation when fake news starts from their cluster.