Social Network and Fake News Spread Analysis

Luca Graziano

December 2021

Consider the graph stored in the file graph5.gml, containing a sample of a population composed by 70 persons. For each person the age, the gender, and the name (anonymised, identifed 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

$$P\left(infection\ of\ v_i\ at\ time\ t+1\right) = \begin{cases} 0.2\ \cdot \eta_i(t) \text{ if } \eta_i(t) \leq 5 \\ 1 \quad \text{otherwise} \end{cases}$$

Where $\eta_i(t)$ is the number of infected neighbours of v_i 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?

Solution

At first we can take a look at the data. Below we display the full dataset contained in the given graph.

name	age	gender
1	86	F
2	79	M
3	79	F
4	77	M
5	78	F
6	80	M
7	76	F
8	71	M
9	74	F
10	82	M
11	76	F
12	82	M
13	76	F
14	76	M
15	84	F
16	86	M
17	81	F
18	73	M
19	81	F
20	80	M
21	34	F
22	39	M
23	35	F

name	age	gender
24	37	M
25	51	F
26	42	M
27	37	F
28	48	M
29	47	F
30	46	M
31	42	F
32	53	M
33	31	F
34	38	M
35	47	F
36	49	M
37	60	F
38	39	M
39	37	F
40	51	M
41	38	F
42	39	M
43	44	F
44	33	M
46	36	M
47	26	F

name	age	gender
45	39	F
48	51	M
49	48	F
50	49	M
51	15	F
52	5	M
53	18	F
54	5	M
55	20	F
56	6	M
57	16	F
58	10	M
59	12	F
60	11	M
61	16	F
62	8	M
63	17	F
64	11	M
65	6	F
66	15	M
67	7	F
68	11	M
69	11	F

Figure 1: Full dataset

Each person has three different attributes: name, age and gender. After a preliminary look at the dataset, we can hypothesize that the network may be divided in different smaller networks, on the basis of one of the attributes.

The network is composed by 70 vertices. In figure 5 it is displayed the graph of the network.

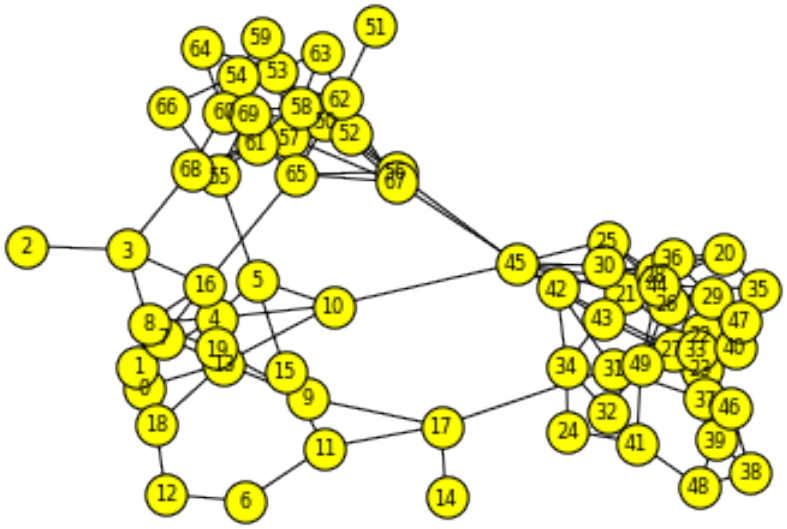

Figure 2: Network graph

To better understand its structure, we can organize the network in different communities. We do this using the Girvan-Newman algorithm, which progressively removes edges from the network based on their betweenness. In Figure 3 it is displayed the graph of the network sorted by communities. In total, we identified 4 different communities, that are identified by color code (community 0 = purple, community 1 = blue, community 2 = green, community 3 = yellow).

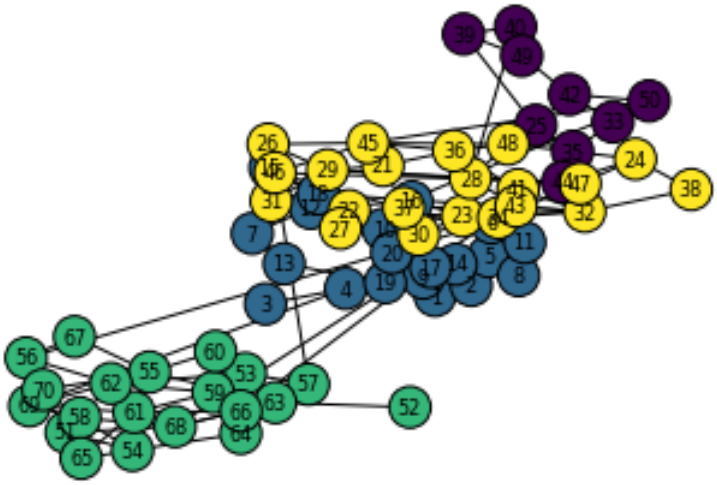

Figure 3: Communities in the network

Each community respectively consists of 9, 20, 20 and 21 nodes. Nodes are grouped as follows:

- $G_0 = 25, 33, 35, 39, 40, 42, 44, 49, 50$
- $G_1 = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20$
- $G_2 = 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70$
- $G_3 = 21, 22, 23, 24, 26, 27, 28, 29, 30, 31, 32, 34, 36, 37, 38, 41, 43, 45, 46, 47, 48$

We might be wondering whether the communities are organized according to specific characteristics of the members (e.g. similar age, same sex). To answer this we can use the assortativity coefficient. Assortativity is defined as the preference for a network's node to link with others that are similar on basis of an attributes. It is a number between -1 and 1, which is computed following how correlated the nodes of a community are regarding a certain characteristic. If we have a positive assortativity value, it means that similar nodes tend to link; a negative value tells us that similar nodes tend to isolate from each other.

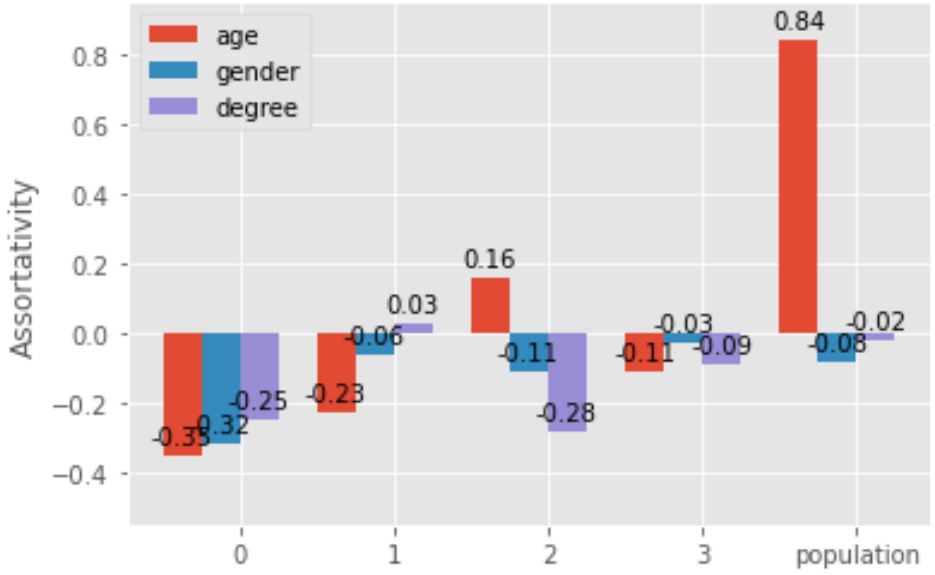

Figure 4: Assortativity plot

The assortativity is computed for each community and on the whole population. The values corresponding to each attribute are the following:

- *Age assortativty* $= -0.35,\ -0.23,\ 0.16,\ -0.11,\ 0.84$
- *Gender assortativity* $= -0.32,\ -0.06,\ -0.11,\ -0.03,\ -0.08$
- *Degree assortativity* $= -0.25,\ 0.03, -0.28,\ -0.09,\ -0.02$

As it's clearly shown in the graph, generally, we don't have significant values of assortativity throughout the different communities; at a whole network level, though, assortativity index on age is equal to 0.84.
We might infer that the nodes of the graph cluster according to age. To verify this, we could check the distribution of age in the four communities.

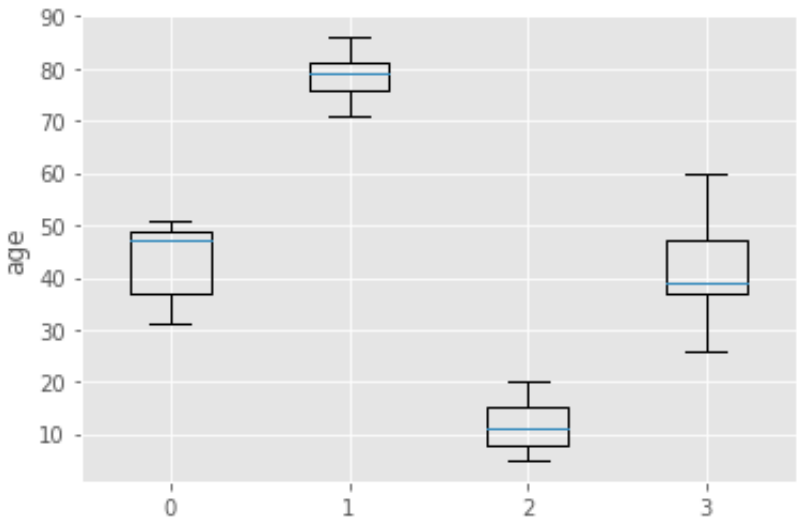

Figure 5: Median and variance of age in the four communities.

As a matter of fact, we can see that communities 0, 1 and 2 are all tightly clustered around specific values; commmunity 3 has more variance, but we

can say that generally the four communities are built according to the age of the members.
The degree and the gender of the nodes have little to no role in the clustering process.

To determine which are the hubs of the network and their degree, we computed the degree of all 70 vertices. From this, we can see that nodes 17 and 46 have the highest degree in the whole network (7) and are therefore the two main hubs. There are also nine other nodes that have degree equal to 6, which can be considered hubs as well since their degree is higher than 5.

Now, suppose that we want to simulate the spreading of a fake news in this graph.
The probability of being infected is defined as follows:

$$P\left(infection\ of\ v_i\ at\ time\ t+1\right) = \begin{cases} 0.2\ \cdot \eta_i(t) \text{ if } \eta_i(t) \leq 5 \\ 1 \quad \text{otherwise} \end{cases}$$

In other words, if the number of the infected neighbors of the node in question at time t is less or equal to five, the probability of being infected at time $t+1$ will be the outcome of a binomial variable with parameters $n = 1$ and $p = 0.2\ *\ n_i(t)$, where $n_i(t)$ is the number of infected neighbors. In case the number of the infected neighbors will be at least six, the probability of becoming infected will be equal to 1.

Knowing this, we can model the problem and simulate the spreading of the fake news, starting from each community to see if we have different results.

We run the simulation for n= 2000 iterations and calculate the average number of steps needed to infect the whole graph. The results for every community are the following:

- $G_0 = 28.56$
- $G_1 = 25.01$

- $G_2 = 26.02$
- $G_3 = 25.32$

We can see how the fake news spreads slightly slower if it originates in community 0, while if it starts from the other three communities it will outcome similar results.

The following boxplot gives a visual representations of the spreading of the fake news starting from each community:

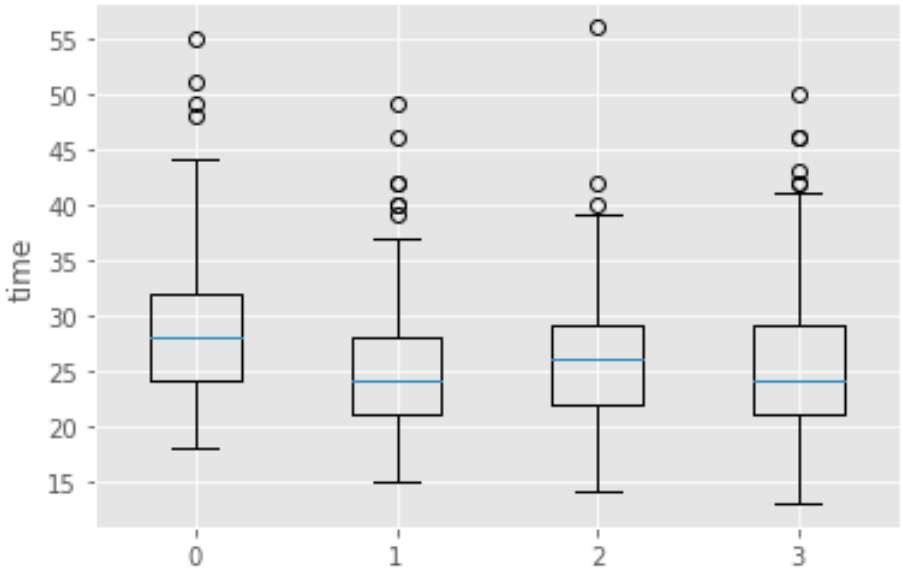

Figure 6: Spreading of a fake news from each community

For a further analysis, it is useful to compute the closeness index of the nodes, since the fake news spreading also depends on which specific node it starts from. After a preliminary analysis, we can identify hubs depending on their closeness: in order to compute an appropriate examination, we run different simulations of the fake news spreading from an hub with degree 7, from the

node with the highest closeness, from an hub of degree 6 and, finally, from an isolated node. They are respectively identified as nodes 17, 46, 9 and 3.

Performance measures are the following:

- $Node_{17} = 22.81$
- $Node_{46} = 23.04$
- $Node_{9} = 24.05$
- $Node_{3} = 30.01$

The corresponding boxplot is the following:

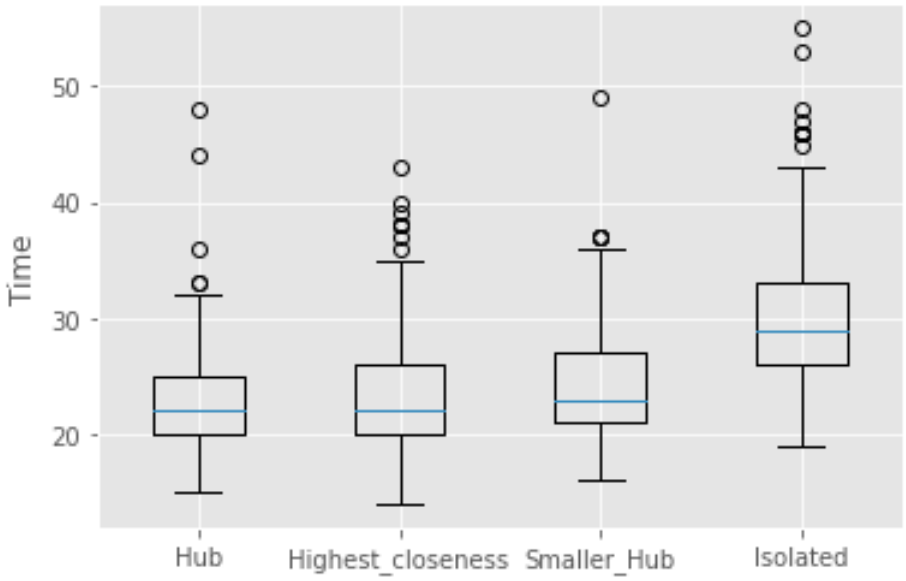

Figure 7: Spreading of a fake news from relevant hubs

As expected, the quickest spread of the news happens when the first infected node is the one with the highest degree, and the slowest is the one corresponding to the isolated node. On the other hand, if the fake news originates from

the smaller hub with 6 connections, it takes a number of steps in between the hubs with the highest degree and highest closeness.

It's also interesting to point out how the first two hubs, that is $Node_{17}$ and $Node_{46}$ have both 7 connections and are respectively considered the one with the highest degree and the one with the highest closeness. The isolated node is both the node with the lowest closeness and the one with the lowest degree (equal to 1). Thus we can see how the computed performances through the simulation make perfect sense once the spatial understanding of the nodes is taken into account.