Homework 4 - Theory/Laboratory

Dainese Fabio, 857661

March 29, 2020

1 Reading and Comprehension

1. The algorithm provided in the *Blondel et al. (2008)* to find the optimal value of $Q(C)$ can be described by the following two steps:

   (a) First, each node in the network is assigned to its own community. Then for each node i, the change in modularity is calculated by removing i from its own community and moving it into the community of each neighbor j of i.
   Then, once the ΔQ value is calculated for all the communities which are directly connected to i, i is placed into the community that resulted in the greatest modularity increase. If no increase is possible, i remains in its original community. This process is applied repeatedly and sequentially to all nodes until no modularity increase can occur.

   (b) In the second phase of the algorithm, it groups all of the nodes in the same community and builds a new network where nodes are the communities from the previous phase. Any links between nodes of the same community are now represented by self-loops on the new community node and links from multiple nodes in the same community to a node in a different community are represented by weighted edges between communities. Once the new network is created and the second phase has ended, the first phase can be re-applied to the new network.

   The previous steps are iterated until there are no more changes and a maximum of modularity is attained.

2. The modularity $Q(C)$ is upper bounded by 1 and lower bounded by -1.

   To prove that $Q(C)$ is lower bounded by -1 we have considered the case of having a complete network, which means having the adjacent matrix 'A' equals to all ones except for the diagonal (of zeros - no self-loop allowed in an undirected graph).
   In this scenario we have:

$$d_i = \sum_{j=1}^{n} = A_{ij} = (n-1)$$

$$d = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij} = \frac{(n^2-n)}{2}$$

Now we are going to use these previous quantities in the modularity formula:

$$\begin{aligned} Q(C) &= \frac{1}{2d}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(A_{ij} - \frac{d_i d_j}{2d}\right)\delta(c_i, c_j) \\ &= \frac{1}{n^2-n}\left[\sum_{i=1}^{n}\sum_{j=1}^{n}\left(A_{ij} - \frac{(n-1)^2}{n^2-n}\right)\delta(c_i, c_j)\right] \\ &\leq \frac{1}{n^2}\left[\sum_{i=1}^{n}\sum_{j=1}^{n}\left(A_{ij} - \frac{n^2}{n^2}\right)\delta(c_i, c_j)\right] \\ &= \frac{1}{n^2}\left[\sum_{i=1}^{n}\sum_{j=1}^{n}\left(A_{ij} - 1\right)\delta(c_i, c_j)\right] \end{aligned}$$

In this particular case of complete graph the value of A_{ij} will be equals to 1 every time $i \neq j$, resulting $(A_{ij} - 1) = (1 - 1) = 0$. The only time in which the multiplication inside the summations is different from zero will be when $i = j$ $((A_{ij} - 1) = (0 - 1) = -1)$ and this happens n times.
Moreover we consider the quantity $\delta(c_i, c_j)$ be equals to 1 since we are dealing with an unique giant partition.
Then we apply these knowledge into the formula, resulting:

$$= \frac{1}{n^2}\left[n(-1)\right] = -\frac{n}{n^2} = -\frac{1}{n}$$

Finally, we consider the following cases:

- $n = 1 : Q(C) = -\frac{1}{n} = -1$
- $n \to \infty$: $Q(C) = \lim_{n\to\infty} -\frac{1}{n} = 0$

Thus, we have proved that the $Q(C)$ is lower bounded by -1.

Meanwhile, to prove that $Q(C)$ is upper bounded by 1 we have considered the case of having a network with 'n' isolated *cliques* each of one having

m nodes (meaning that $|V| = n \cdot m$). To recall also that the number of edges in a clique of m nodes is equal to $\binom{m}{2}$.

As before, we applied these knowledge into the modularity formula, resulting:

$$
\begin{aligned}
Q(C) &= \frac{1}{2d} \sum_{i=1}^{n} \sum_{j=1}^{n} \left(A_{ij} - \frac{d_i d_j}{2d} \right) \delta(c_i, c_j) \\
&= \frac{1}{2d} \sum_{k=1}^{n} \left[\binom{m}{2} - \frac{(m-1)^2}{2d} \right] \\
&= \frac{1}{2d} \sum_{k=1}^{n} \left[\frac{m(m-1)}{2} - \frac{(m-1)^2}{2d} \right] \\
&= \frac{1}{2d} \sum_{k=1}^{n} \left[\frac{(m^2-m)d - (m-1)^2}{2d} \right] \\
&= \frac{1}{2d} \sum_{k=1}^{n} \left[\frac{((d-1)m+1)(m-1)}{2d} \right] \\
&= \frac{((d-1)m+1)(m-1)n}{4d^2} \\
&= \frac{((\frac{nm(m-1)}{2}-1)m+1)(m-1)n}{4\left(\frac{nm(m-1)}{2}\right)^2} \\
&= \frac{((\frac{nm(m-1)}{2}-1)m+1)(m-1)n}{4\frac{n^2m^2(m-1)^2}{4}} \\
&= \frac{nm^2(m-1) - 2m + 2}{nm^2(m-1)} \\
&= 1 - \frac{2m-2}{nm^2(m-1)} \\
&= 1 - \frac{2(m-1)}{nm^2(m-1)} \\
&= 1 - \frac{2}{nm^2} = 1 - \frac{2}{|V|m}
\end{aligned}
$$

Finally, if we consider the case of $|V| \to \infty$, we have:

$$
Q(C) = \lim_{|V| \to \infty} \left(1 - \frac{2}{|V|m} \right) = 1 - 0 = 1
$$

Thus, we have proved that the $Q(C)$ is upper bounded by 1.

3. Consider the new partition defined in the given task, we have to prove that:

$$Q(\tilde{C}) - Q(C) = \frac{k_{2h}}{d} - \frac{k_{1h}}{d} + \frac{d_h k_1}{2d^2} - \frac{d_h k_2}{2d^2}$$

We start by recalling the definition of $Q(C)$:

$$Q(C) = \frac{1}{2d} \sum_{i=1}^{n} \sum_{j=1}^{n} \left(A_{ij} - \frac{d_i d_j}{2d} \right) \delta(c_i, c_j) = \sum_{i=1}^{R} Q(C_i)$$

Where $Q(C_i)$ is defined as follows:

$$Q(C_x) = \frac{1}{2d} \sum_{i,j \in C_x} \left(A_{ij} - \frac{d_i d_j}{2d} \right)$$

Now we apply these definition to prove the desired result:

$$\begin{aligned} Q(\tilde{C}) - Q(C) &= \sum_{i=1}^{R} (Q(\tilde{C}_i) - Q(C_i)) \\ &= Q(\tilde{C}_1) - Q(C_1) + Q(\tilde{C}_2) - Q(C_2) + \sum_{i=3}^{R} (Q(\tilde{C}_i) - Q(C_i)) \end{aligned}$$

Since the summation is equals to 0, the equation becomes:

$$= Q(\tilde{C}_1) - Q(C_1) + Q(\tilde{C}_2) - Q(C_2)$$

Now we substitute the terms with their definitions:

$$= \frac{1}{2d} \left[\sum_{i,j \in \tilde{C}_1} \left(A_{ij} - \frac{d_i d_j}{2d} \right) - \sum_{i,j \in C_1} \left(A_{ij} - \frac{d_i d_j}{2d} \right) + \sum_{i,j \in \tilde{C}_2} \left(A_{ij} - \frac{d_i d_j}{2d} \right) - \sum_{i,j \in C_2} \left(A_{ij} - \frac{d_i d_j}{2d} \right) \right]$$

And finally we apply the given definition of k_r and k_{rh} into the equation, resulting:

$$\begin{aligned} &= \frac{1}{2d} \left[2k_{2h} - 2k_{1h} + \frac{2d_h k_1}{2d} - \frac{2d_h k_2}{2d} \right] \\ &= \frac{k_{2h}}{d} - \frac{k_{1h}}{d} + \frac{d_h k_1}{2d^2} - \frac{d_h k_2}{2d^2} \end{aligned}$$

4. Consider the new partition defined in the given task, we have to prove that:

$$Q(C) - Q(\tilde{C}) = \frac{d_h k_2}{2d^2} - \frac{k_{2h}}{d}$$

We start by multiplying all by -1, obtaining:

$$Q(\tilde{C}) - Q(C)$$

Using the knowledge of the previous point we can substitute the values, resulting:

$$Q(\tilde{C}) - Q(C) = \frac{k_{2h}}{d} - \frac{k_{1h}}{d} + \frac{d_h k_1}{2d^2} - \frac{d_h k_2}{2d^2}$$

Since $\tilde{C} = \emptyset$ we have that $k_{1h} = k_1 = 0$, thus:

$$Q(\tilde{C}) - Q(C) = \frac{k_{2h}}{d} - \frac{d_h k_2}{2d^2}$$

Finally, we multiply all again by -1 and we obtain the expected result, as reported below:

$$Q(C) - Q(\tilde{C}) = \frac{d_h k_2}{2d^2} - \frac{k_{2h}}{d}$$

2 Interlocking Directorate

2.1 Subgraph Extraction

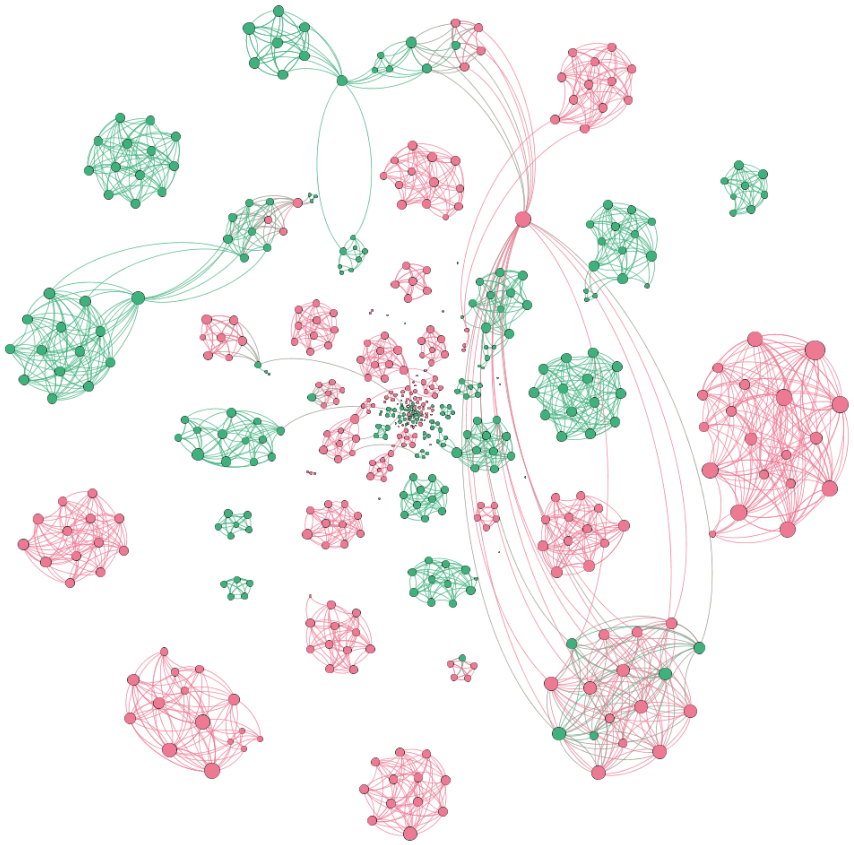


Figure 1: Network

2.2 Connected Components

In the network represented in 'Figure 1' there are:

- 533 nodes;
- 2251 edges;
- 83 weakly connected components.

Moreover, in the 'Figure 2' are highlighted in red the first and the second largest connected components.

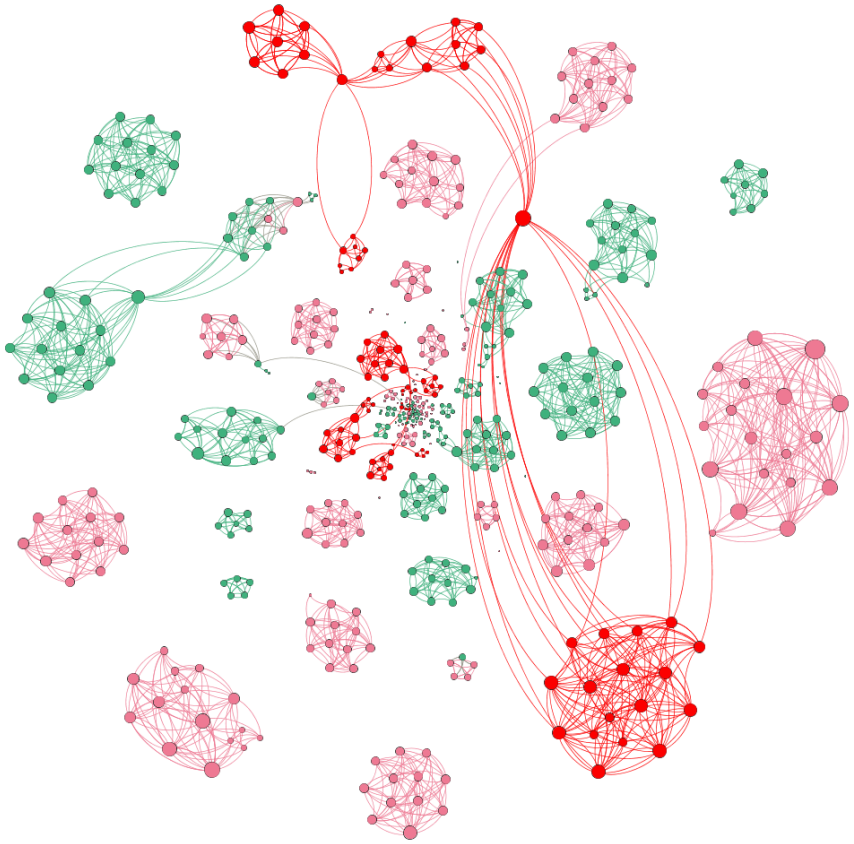
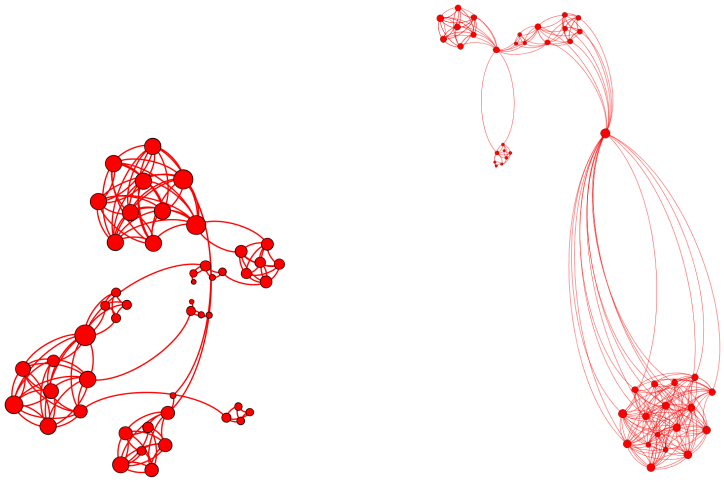


Figure 2

Just to make it more clear, here the isolated two most largest connected components:



(a) Largest connected component (b) Second largest connected component

To isolate the largest connected component it has been used the '*Giant Component*' filter, meanwhile to find the second largest connected component it has been applied the following filters: "Giant Component" → "NOT (Nodes)" → "Giant Component" (where "→" means to applying as a sub-filter).

2.3 Communities

In order to find the communities in the network it has been used the '*Modularity*' statistic with different resolutions (and without *edge weights*).

With resolution 0.5 the results are:

- 89 communities;
- 8 communities belonging to the giant component;
- By using as a node size the *betweenness centrality* (size between 1 and 16) and colouring each community with a different colour, the result is:

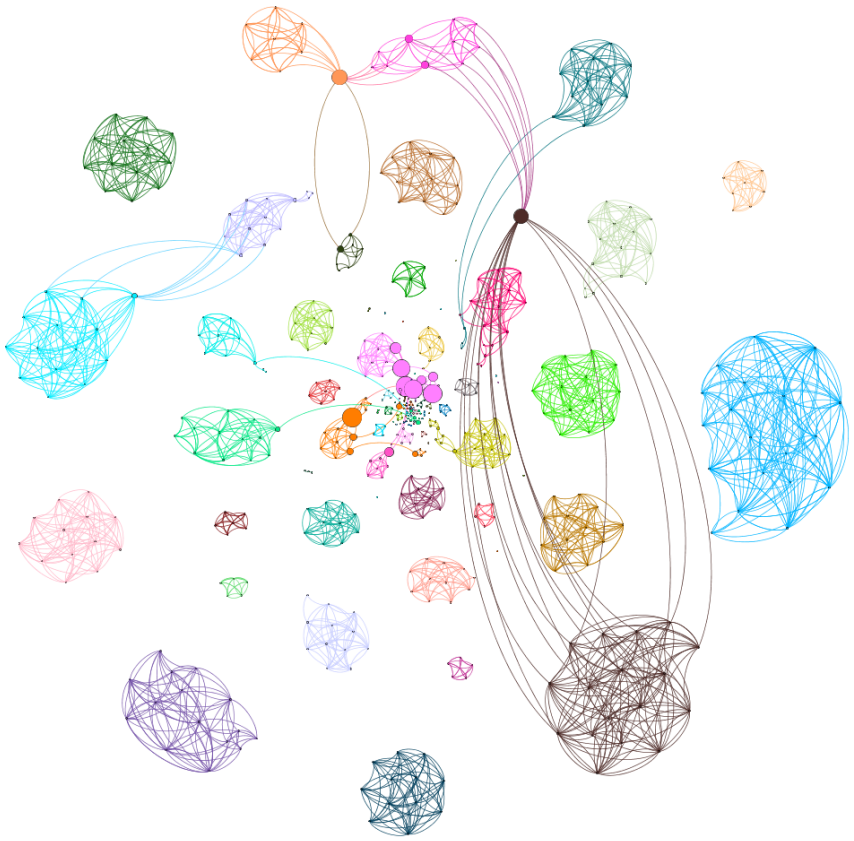


Figure 4

Meanwhile with resolution 1 the results are:

- 85 communities;
- 7 communities belonging to the giant component;
- By using as a node size the *betweenness centrality* (size between 1 and 16) and colouring each community with a different colour, the result is:

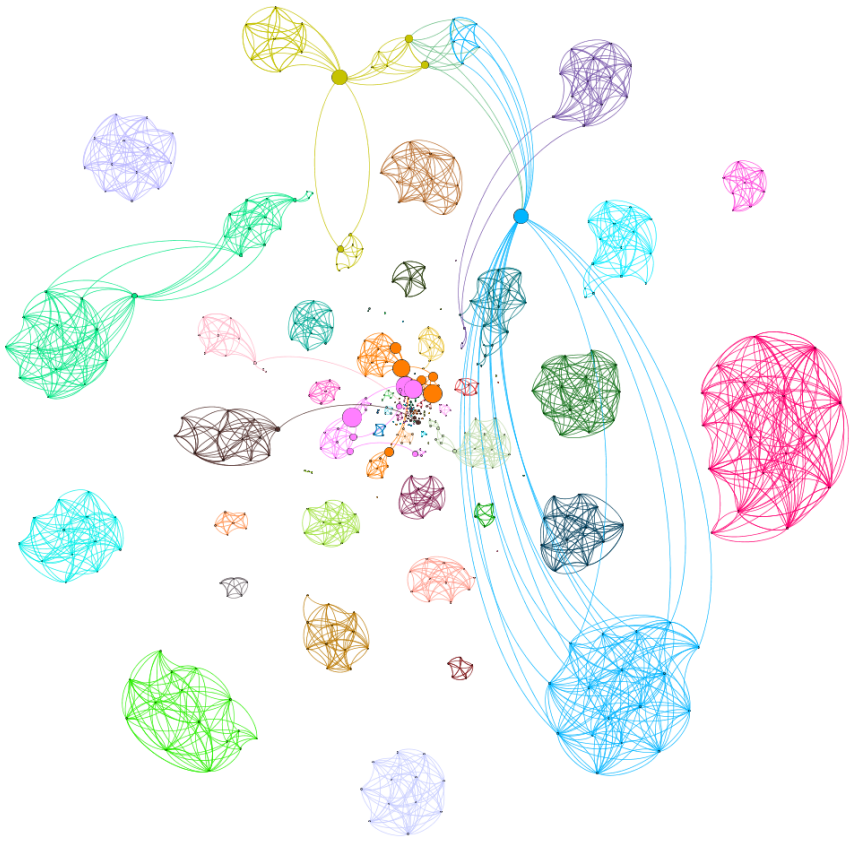


Figure 5

Finally with resolution 2 the results are:

- 85 communities;
- 6 communities belonging to the giant component;
- By using as a node size the *betweenness centrality* (size between 1 and 16) and colouring each community with a different colour, the result is:

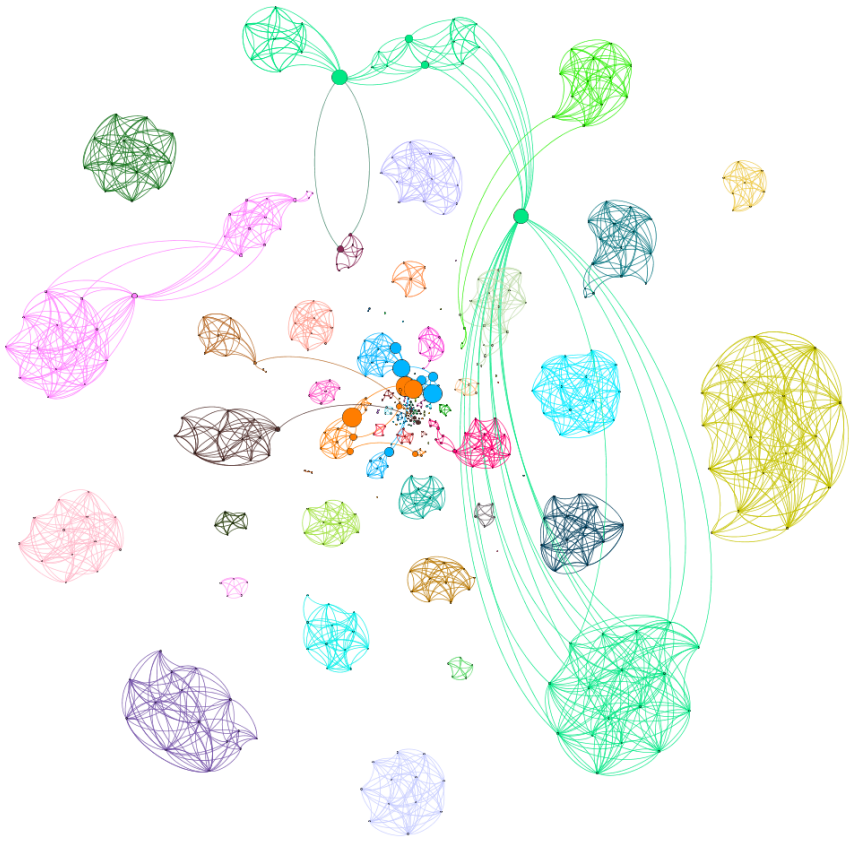


Figure 6

The number of communities and the number of connected components can be different because the definition of *communities* doesn't imply that there must not be any edges between communities, meanwhile that's a requisite to identify a single connected component.

So, for example, in case of a network composed by a series of isolated cliques, the number of communities and the one of connected components are equal. At the same time if the network is composed by a series of cliques weakly connected between each other, the number of communities and the one of connected components are different.