ASN - Assignment 2

Edoardo Gabrielli

November 30, 2021

1 Introduction

The dataset has been taken from UCINET [1] and is distributed by *networksdata* package to be ready to use in R. As described in UCINET, the dataset has no publications and is retrieved from the document Operazione Infinito (pp. 87-110) [2]. Moreover it appears to be also reconstructed from the official warrant issued by an Italian judge. This information is incomplete since the link to the warrant is broken.

Operazione Infinito is a long document describing each person involved in the investigations, including details about their roles in the organization, their involvements in the political and business worlds, crimes committed and finally deep explanations about how all the actors are connected (what they did, with who, etc). Names of places (restaurants, clubs, etc) where they used to meet are also reported, completed with their conversations captured by the audio surveillance.

On pp 87-110, as said above, there is a list of all the summits registered, with the people involved and places where these meetings took place (all in Milan).

A visualization of the network is shown in Figure 1 while a summary of its basic statistics is in Table 1.

N	Density	Mean k (T)	Mean k (F)
203	0.0221	9.681	2.917

Table 1: Table which summarizes the basic statistics about the network. Here T stands for the node type, that is TRUE (summit), while F is the contrary (suspect).

The real community structure of the 'Ndrangheta organization is purely hierarchical and divided in gangs (*locali* is the appropriate term), so the places shown here are restricted to other components and a less influential member can't access some of the them. Moreover it is worth notice that usually these summits were held because they needed to solve fights between *locali* or to confer a new grade (*dote*) to one of the member (pp. 87 of [2]), which is done by means of a special ceremony, a methodology which could reminds the Italian masonry [1].

[1] The rigid hierarchical structure and the secrecy of their rituals, bosses, etc. is, among other things, one of the reason why it is so challenging for the government to get rid of this organization. For example it is still unclear what are all the grades in this structure.

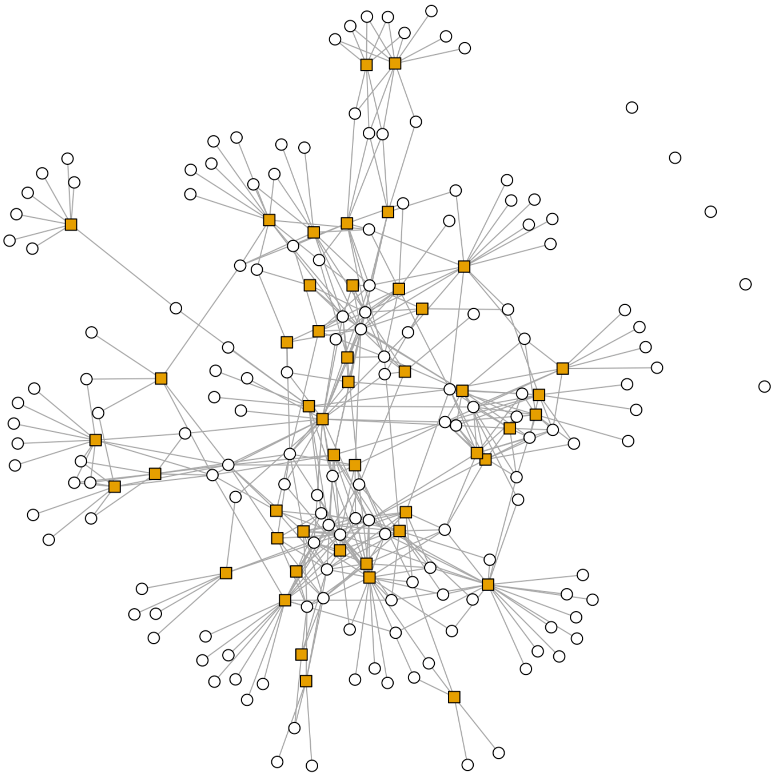


Figure 1: The network where each circle corresponds to a suspect and each square to a place where a summit took place.

Therefore the representation of communities found by an algorithm should divide the most influential and authoritative components from the other less important ones since my assumption is that components which are on the same level of the hierarchy, or in the direct predecessor/successor layers, meet each other more frequently that the others.

2 Preliminaries

Before starting to compute communities it is worth to project the data to a one-mode network, because the methods I am going to use don't work well on two-mode network. Nevertheless, algorithms specifically made for bipartite graphs exist and can be found in [3] and [4] just to cite two.

The one-mode projection (suspect-to-suspect) can be seen in Figure 3 and eliminating places from the network creates a dense graph with a set of separate cliques, with denser regions toward the center of the graph. In Figure 2 the distribution of degrees is shown.

An algorithm that intuitively can perform well on this type of structure is the so called Walktrap [5]. In a nutshell, it assumes that a random walk should get trapped inside a densely connected regions of the graph, so that it is used as a measure for structure similarity and then this measure is employed as the distance for hierarchical clustering.

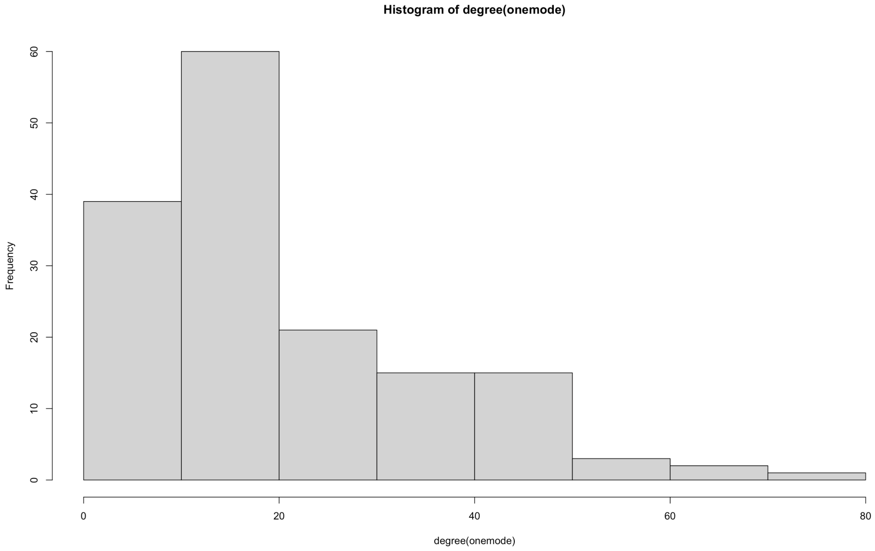


Figure 2: The distribution of degrees in the one-mode network.

3 Analysis

In Figure 3 there is a visualization of the network after running the algorithm. The total number of clusters is 13, where the 5 disconnected nodes are clustered inside their own communities so in the giant component the actual communities are 9. Moreover the distribution of community size is shown in Figure 5 (a), while basic statistics are summarized in 2.

N	Density	Mean k
156	0.13	20.76

Table 2: Table which summarizes the basic statistics about the one-mode network.

Since I don't have a ground truth (can be investigated reading the document Operazione Infinito, but would be a long time-consuming task and it is out of the scope of this assignment), I decided to compare the algorithm with the Louvain's method [6], which is quite different from Walktrap.

Louvain algorithm starts by assigning each vertex to a single community and then it begins to individually consider each community and the neighboring ones. If merging two of them results in a gain of modularity, merge them. In the second phase it proceeds to merge nodes belonging to the same community into a single vertex and then re-apply the first phase. These phases are repeatedly executed until there is no more gain in modularity. Although the methodology described in the paper consider the weight, the igraph implementation accept unweighted graphs as input so it can be used for this case.

The number of communities found by Louvain is 12, one less than Walktrap (the latter assign the purple node (Francesco Cristello) inside an individual community, working as a bridge between the light blue cluster and the rest of the network). In Figure 5 (b) the community size distribution are confronted and in Figure 4 the network is shown. The two modularities found by the algorithms are 0.4549 for Walktrap and 0.4769 for Louvain, with an NMI [7] of 0.8884, which suggests that there are no great differences between the two community structures found.

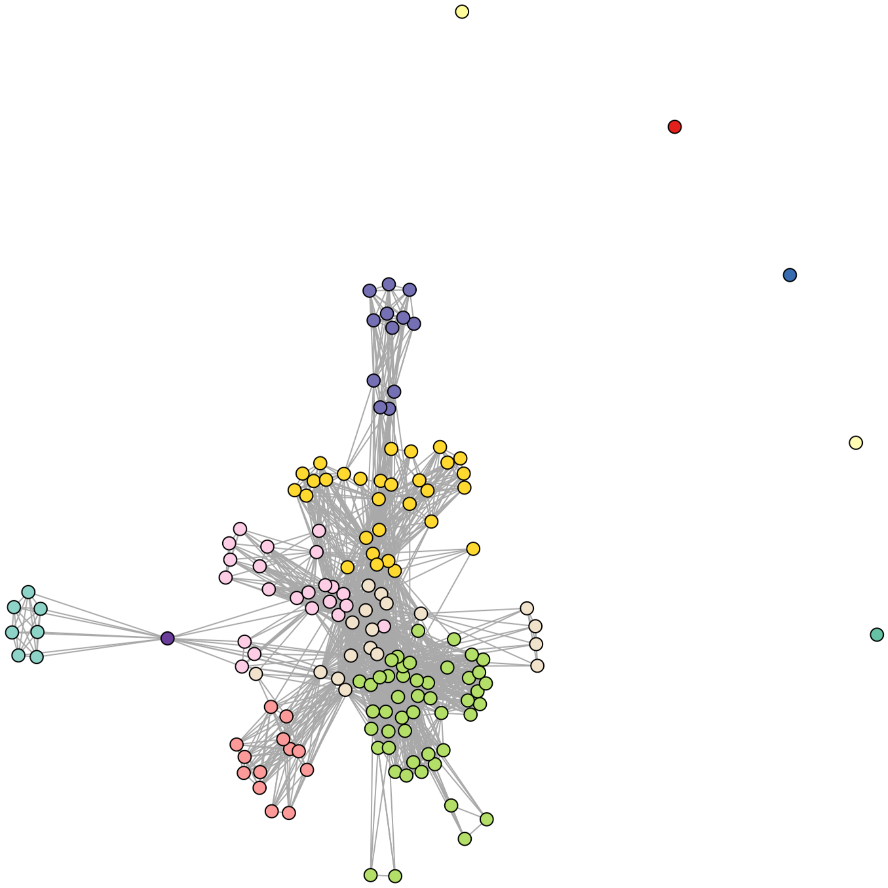


Figure 3: The one-mode projection (suspect-to-suspect) of the original network. Nodes colored according to the community assigned by the Walktrap algorithm.

4 Conclusion

Although the modularity can be computed and gives a measure to understand the quality of the output, the lack of a ground truth makes the analysis of the community structure difficult. The two algorithms seem to differ mostly

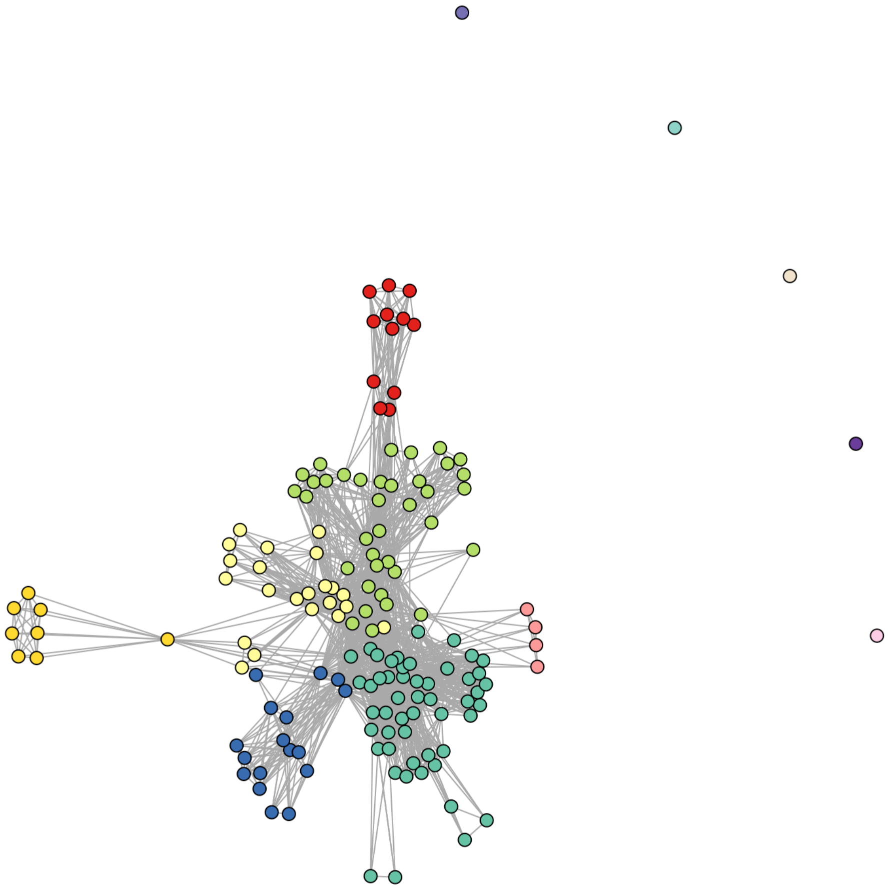
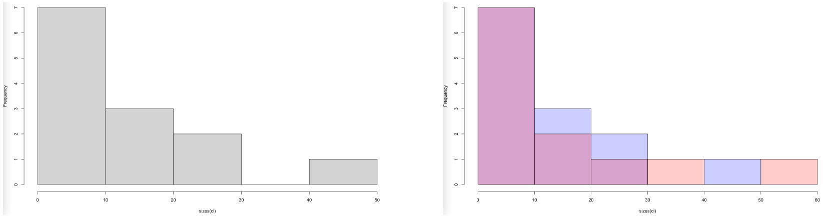


Figure 4: Nodes colored according to the community assigned by the Louvain algorithm.

in the central part of the network where most of the connections take place, nevertheless the Louvain algorithm seems to capture the community structure a little bit better, as the modularity suggests.

As discussed in class, this can be given by the resolution limit, indeed, it can be seen that the two modularities differ by a small value and the Louvain algorithm finds less communities, suggesting that it might fails to detect small structures and tends to merge partition which is better to keep separate. To prove this, I have tried to generate the subgraphs obtained from some of the communities, run the Louvain algorithm again on the communities separately and check the results. It turns out that on each of them the modularity is sub-optimal and plotting the graphs, results in a confused community structure. For the sake of the synthesis, only the result of the subgraph from the biggest community is reported in Figure 6. This may suggest that there is no strong smaller structures, hence disproving my initial thought.

In the end, it is clear that the discussion done at the beginning still holds: we can see at the borders of the networks, where most of the cliques are, that



(a)

(b)

Figure 5: Distribution of community size of the Walktrap algorithm (left). Distribution of community size by the Walktrap algorithm (blue) versus the one generated by the Louvain (red) one (right).

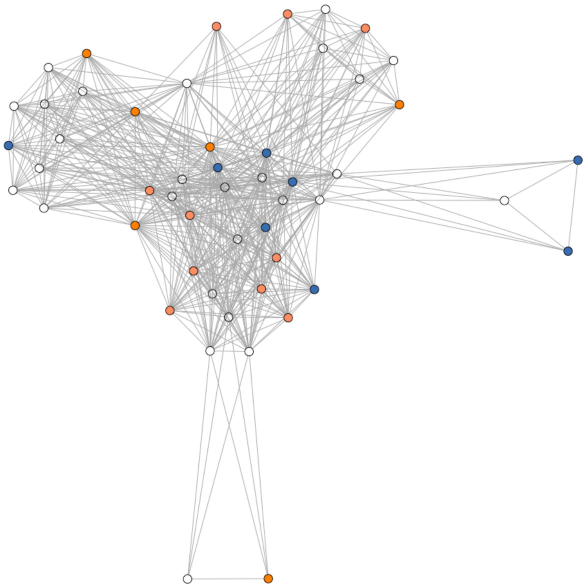


Figure 6: Nodes colored according to the community assigned by the Louvain algorithm in the subgraph obtained from the biggest community found by the same algorithm. Modularity is 0.1798.

the suspects who attend the same events are clustered together so it can be assumed that they are in the same *locale*, with some approximation given by the lack of other data besides the co-attendance to summits (which usually are not restricted to one single *locale* as discussed in the beginning). To be more precise on the confrontation about the separation into clusters and the actual community structure, a deeper knowledge of the domain should be obtained before.

All plots used here have been generated with R Studio and the code is available in the dedicated GitHub repository [8].

References

[1] Covert Network ’Ndrangheta.
https://sites.google.com/site/ucinetsoftware/datasets/covert-networks/ndrangheta-mafia-2

[2] Operazione Infinito, ordinanza di custodia cautelare.
https://liberavco.liberapiemonte.it/operazione-infinito-ordinanza-di-custodia-cautelare/

[3] Calderer, G., & Kuijjer, M. L. (2021). Community detection in large-scale bipartite biological networks. Frontiers in Genetics, 12, 520.

[4] Yen, T. C., & Larremore, D. B. (2020). Community detection in bipartite networks with stochastic block models. Physical Review E, 102(3), 032309.

[5] Pascal Pons & Matthieu Latapy. Computing communities in large networks using random walks.
https://arxiv.org/pdf/physics/0512106.pdf

[6] VD Blondel, J-L Guillaume, R Lambiotte & E Lefebvre. Fast unfolding of communities in large networks.
https://arxiv.org/pdf/0803.0476.pdf

[7] Fortunato, S., & Hric, D. (2016). Community detection in networks: A user guide. Physics Reports, 659, pp 14.

[8] GitHub repository to reproduce the results:
https://github.com/edogab33/asn-ndrangheta-community-detection