Graph nodes clustering: a comparison between algorithms

Clustering di nodi in un grafo: un confronto tra algoritmi

Ilaria Bombelli

Abstract Networks represent an important tool to describe problems and applications in various fields, such as science, technology and economics. Statistics can play a role in a network framework, for example using some clustering techniques to detect clusters of nodes. This work focuses on reviewing the existing algorithms designed specifically for this aim and on suggesting the application of other clustering techniques that require a matrix of distances or dissimilarities between units: a description of how to get such matrix is also provided. A comparison between the aforementioned algorithms is given, by applying them to a benchmark network.

Abstract *I network (o grafi) sono uno strumento importante per rappresentare problemi e applicazioni in vari campi, come quello scientifico, tecnologico e economico. La statistica può giocare un ruolo importante in questo contesto, per esampio applicando alcune tecniche di clustering per identificare clusters di nodi. Tale lavoro passa in rassegna gli algoritmi appositamente costruiti per raggiungere questo scopo e suggerisce anche l'applicazione di altri algoritmi che prendono in input una matrice di distanze o dissimilarità tra le unità: viene fornita anche una descrizione su come ottenere questa matrice. Gli algoritmi suddetti vengono confrontati, applicandoli ad un network di riferimento (benchmark).*

Key words: Clustering, Network, Nodes Clustering, Fuzzy clustering

Ilaria Bombelli
Dipartimento di Scienze Statistiche, Sapienza Università di Roma, Piazzale Aldo Moro, 5 (00185) Roma, e-mail: ilaria.bombelli@uniroma1.it

1 Introduction

Networks (graphs) can be found in many fields: for example, in a society, a network may represent how people interact one another; in technological field, networks may represent email exchange. A graph or a network is a mathematical tool representing connection or relationship between several objects. Formally, a graph G is defined as an ordered tuple of 2 sets, i.e. $G = (V,E)$, where V is the set of n unique nodes, i.e. $V = \{v_1, \ldots, v_n\}$ and E is the set of m edges, i.e. $E = \{e_1, \ldots, e_m\}$.

In a network framework it can be of interest the application of clustering techniques: indeed, we can consider the nodes as the statistical units and the aim is to detect clusters of nodes. Usually in this framework, the starting point of many clustering algorithms, i.e. the distance matrix $\mathbf{D}$, is not provided and therefore, given a network object, a measure of distance between nodes has to be considered.

The paper will be organized as follows: in Section 2 the description of the clustering algorithms that have been used is provided, as well as the explanation of how to build a distance matrix $\mathbf{D}$ from a network object; Section 3 shows an application of the clustering algorithms to a benchmark network. Finally in Section 4 final remarks are given and further possible developments are sketched.

2 Methodology

In this section, an overview on the clustering techniques is presented, followed by an explanation of how to build a distance matrix from a given graph.

2.1 Clustering algorithms

The main two classes of clustering algorithms are hierarchical and non-hierarchical: the former is a class of algorithms that generate n different nested classifications, the latter is a class of algorithms that give rise to a single partition with k (fixed before running the algorithm) groups. Hierarchical clustering algorithms can be either agglomerative or divisive.

In graph framework, agglomerative methods start with the only set of nodes V; the whole network $G = (V,E)$ will be progressively constructed by adding edges between nodes and involving nodes into nested larger and larger *communities* (subsets of the network).
Divisive methods instead start from the whole network and progressively cut edges and divide the network into smaller and smaller communities.

As examples of the two aforementioned different techniques two algorithms designed to be applied to a network in order to detect clusters of nodes are described and then applied in Section 3.
Louvain algorithm [2] belongs to the class of agglomerative algorithms. Accord-

ing to [2], this kind of algorithm finds high quality communities and it is based on modularity optimization: modularity index for clustering evaluation was deeply discussed by [3].
The algorithm consists of two phases, that are repeated alternatively. The starting point is the set of n nodes: it assigns a different membership community to each node of the network, so that in the initial partition there are as many communities as there are nodes. The first phase of the algorithm consists in considering each node i and its neighbors j: it computes the gain of modularity that would be obtained by removing the node i from its community and placing it into the community of neighbor j. After having evaluated all these gains for each node i, the algorithm places node i in the community for which the gain is maximum. This procedure is repeated for each node i in the set of nodes V until no further improvement can be achieved.
The second phase of the algorithm consists in building a new graph whose nodes are the communities detected in phase 1. In order to achieve this goal, the algorithm firstly computes the weights of the edges between any two new nodes (i.e. any two communities identified in phase 1), by summing up all the weights of the links between nodes in the corresponding two communities.
After phase 2 is completed, the algorithm applies again phase 1 to the new network and to iterate. Hence, this type of algorithm is an agglomerative hierarchical procedure, as communities of communities are built during the process: the last community will be the one that involves all the nodes and aggregates all the communities detected in the previous step in only one.

Girvan-Newman algorithm [5] is one of the most known and used alogorithm for communities detection problem. This algorithm is divisive and therefore it starts with the whole network and progressively cuts edges (most likely between communities) and reveals the community structure of the graph.
In order to find such edges, Girvan and Newman generalized the idea of node betweenness centrality, defining the edge betweenness centrality of an edge as the proportion of shortest paths connecting two vertices in the graph and passing through the edge. More formally, the edge betweennes centrality of edge e is

$$C_B(e_i) = \sum_{i \neq j \in V} \frac{\sigma_{jk}(e_i)}{\sigma_{jk}} \tag{1}$$

where σ_{jk} is the number of shortest paths connecting node v_j and node v_k, and $\sigma_{jk}(e_i)$ is the number of shortest paths connecting node v_j and node v_k that run along edge e_i.
The algorithm proceeds as follows: first of all the edge betweenness for all edges in the network are computed; then, the edge with the highest betweennes is removed: indeed, such edge is an inter-communities edge, since all the paths linking any two nodes belonging to different communities go through it. Finally the algorithm computes again the betweenness for all the edges affected by the removal and repeats itself from the second step until no edges remain. As it is clear from the algorithm, the Girvan-Newman procedure belongs to the so-called divisive methods: indeed, it starts by taking into consideration the whole network $G = (V,E)$; then, according

to the decreasing order of the edge betweenness, edges are cut progressively and therefore the whole network is splitted into smaller and smaller communities until we get n communities, as many as there are nodes.

These aforementioned algorithms were built such that they can be applied directly to a network object; it is of interest to notice that actually any clustering algorithm that takes as input a distance matrix $\mathbf{D}$ can be applied, provided that $\mathbf{D}$ can be built from the network $G = (V, E)$.

Clustering algorithms differ from each other also depending on the approach they take. More in details, the *hard* (or *crisp*) approach assigns any single object either to one cluster or to another one. Instead the *fuzzy* approach, introduced by [1], assigns to each object k membership degrees, one for each cluster. Each membership degree takes values in $[0,1]$, instead of in $\{0,1\}$, as occurs in hard approach, and it is such that the membership degrees of each unit sum up to 1.
Among fuzzy clustering algorithms, we focus on the Non-Euclidean Fuzzy Relational Clustering (NEFRC) algorithm, introduced by [4]. [4] proposed a fuzzy clustering algorithm, whose objective function is the following: let i and j identify units, $i, j \in \{1,2,\ldots,n\}$ and c identify clusters, ranging in $\{1,2,\ldots k\}$, where k is the desired number of clusters,

$$F_{NEFRC} = \sum_{c=1}^{k} \frac{\sum_{j=1}^{n}\sum_{i=1}^{n} u_{ic}^{m} u_{jc}^{m} d_{ji}}{2\sum_{t=1}^{n} u_{tc}^{m}} \tag{2}$$

subject to constraints:

$$\sum_{c=1}^{k} u_{ic} = 1, \qquad i = 1,2,\ldots,n \tag{3}$$

$$u_{ic} \geq 0 \qquad i = 1,\ldots,n \qquad c = 1,2,\ldots,k \tag{4}$$

where m is *fuzzifier* or *fuzzyness parameter* that controls how fuzzy the clusters tend to be; u_{ic} is the membership degree of unit i to cluster c. Noteworthy that relational data in $\mathbf{D}$ can be from any dissimilarity measure: indeed, most dissimilarity data are non-Euclidean and, as [6] showed, original relational fuzzy clustering methods that only require Euclidean distances often failed.

2.2 Distance Matrix

In order to apply the NEFRC algorithm, it is necessary to build distance matrix, having dimension $n \times n$ and as generic element d_{ij} the distance between node labeled with i and node labeled with j; $\mathbf{D}$ must be symmetric (i.e. $d_{ij} = d_{ji} \; \forall \; i, j = 1,\ldots,n$), must have null diagonal (i.e. $d_{ii} = 0 \; \forall \; i = 1,\ldots,n$) and must have non-negative entries (i.e. $d_{ij} \geq 0 \; \forall \; i, j = 1\ldots,n$). In order to build such matrix, the *geodesic*

distance, i.e. the length of the shortest path linking any two nodes, is used as measure of distance between the nodes: in this way, the higher the length of the shortest path between any two nodes, the more distant the nodes.

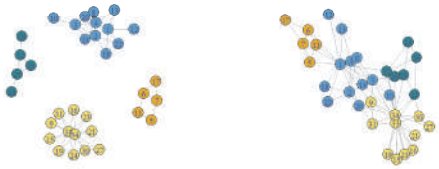
3 Application

The network object of our study is well known in literature and widely used for analysis; it belongs to the category of social networks. The network is called "Zachary karate club network" and it was downloaded from the network repository: it contains social ties among the members of a university karate club collected by Wayne Zachary in 1977. Each member of the club is represented by a node and each tie is represented by an edge. More in details, it is a unweighted, undirected graph, having $|V| = n = 34$ and $|E| = m = 77$: so among the 34 club members there have been 77 bonds of friendship.

This network is well known in literature (see, for example, [9]), since it is of interest for detecting communities: indeed, an argument between the president and the instructor regarding some pay causes actually occurred and divided the group in two parts. The real clustering structure of the problem is therefore known and available (Figure 1 (d)) and hence it is possible to compare the obtained partition with the ground truth one, by using external validation indices to evaluate the performance of the methods used.
The results of the applications of the Girvan-Newman, Louvain and NEFRC algorithms are provided in Figure 1: it shows that the first two, i.e. the hard clustering algorithms, failed to recognize the clustering structure, since they identify four clusters instead of two; the fuzzy clustering algorithm, instead, detects exactly the true partition, leading to an Adjusted Rand Index ([8]) of 1.

4 Conclusion

This contribution aimed to review the most important and known clustering techniques that can be used in order to detect clusters of nodes and to suggest also the use of other clustering algorithms, that may be more successful, as occurred in our application. Other authors focused on applying fuzzy clustering algorithms to a network to detect the underlying community structure: we recall for example [9] that applied Non-Euclidean Relational Fuzzy C-Mean to the distance matrices resulting from the application of hard algorithms and [7] that applied Fuzzy c-Means to the spectral features extracted by spectral clustering from the graph.
Further developments regarding the application of clustering in a network framework is to consider a graph as a statistical unit and look for clusters of networks. This idea open up the field of possible questions regarding which measure of distance between networks can be used and all other instances related to it.



(a) Girvan-Newman algorithm

(b) Louvain algorithm

(c) NEFRC algorithm

(d) True partition

Fig. 1: Zachary Karate Network: Clustering results and true partition

References

1. Bezdek, J.C.: Objective function clustering. In: Pattern recognition with fuzzy objective function algorithms, pp. 43–93. Springer (1981)
2. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment **2008**(10) (2008)
3. Brandes, U., Delling, D., Gaertler, M., Gorke, R., Hoefer, M., Nikoloski, Z., Wagner, D.: On modularity clustering. IEEE transactions on knowledge and data engineering **20**(2), 172–188 (2007)
4. Davé, R.N., Sen, S.: Robust fuzzy clustering of relational data. IEEE Transactions on Fuzzy Systems **10**(6), 713–727 (2002)
5. Girvan, M., Newman, M.E.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA **99**(cond-mat/0112110), 8271–8276 (2001)
6. Hathaway, R.J., Bezdek, J.C.: Nerf c-means: Non-euclidean relational fuzzy clustering. Pattern recognition **27**(3), 429–437 (1994)
7. Havens, T.C., Bezdek, J.C., Leckie, C., Chan, J., Liu, W., Bailey, J., Ramamohanarao, K., Palaniswami, M.: Clustering and visualization of fuzzy communities in social networks. In: 2013 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp. 1–7. IEEE (2013)
8. Morey, L.C., Agresti, A.: The measurement of classification agreement: An adjustment to the rand statistic for chance agreement. Educational and Psychological Measurement **44**(1), 33–37 (1984)
9. Runkler, T.A., Ravindra, V.: Fuzzy graph clustering based on non-euclidean relational fuzzy c-means. In: 2015 Conference of the International Fuzzy Systems Association and the European Society for Fuzzy Logic and Technology (IFSA-EUSFLAT-15). Atlantis Press (2015)