

Density-based clustering of social networks

Giovanna Menardi¹  | Domenico De Stefano² 

¹Department of Statistical Sciences,
University of Padova, Padova, Italy

²Department of Political and Social
Sciences, University of Trieste, Trieste,
Italy

Correspondence

Domenico De Stefano, Department of
Political and Social Sciences, University
of Trieste, Trieste, Italy.

Email: ddestefano@units.it

Accepted: 26 November 2021

Abstract

The idea of the modal formulation of density-based clustering is to associate groups with the regions around the modes of the probability density function underlying the data. The correspondence between clusters and dense regions in the sample space is here exploited to discuss an extension of this approach to the analysis of social networks. Conceptually, the notion of high-density cluster fits well the one of community in a network, regarded to as a collection of individuals with dense local ties in its neighbourhood. The lack of a probabilistic notion of density in networks is turned into a strength of the proposed method, where node-wise measures that quantify the role of actors are used to derive different community configurations. The approach allows for the identification of a hierarchical structure of clusters, which may catch different degrees of resolution of the clustering structure. This feature well fits the nature of social networks, disentangling different involvements of individuals in aggregations.

KEYWORDS

centrality, community detection, modal clustering, weighted networks

1 | INTRODUCTION

1.1 | Background and motivation

Within social communities, individuals sparsely interact with each other and usually set a tight relationship with a limited number of subjects. Interactions favour individuals to aggregate into groups, where the relationships are stronger and the information flow is more intense than outside.

The generating mechanism of these groups, albeit pervasive, is complex and often difficult to be disclosed. On the one hand, different kinds of relationship may be established, from friendship to professional collaboration, each of them possibly with different levels of intensity. On the other hand, aggregation may be driven by diverse, sometimes unobserved, social mechanisms—homophily, popularity, ranking or influence. Depending on the context, cohesive communities may be formed, where even relationships connect each actor with most of other actors. This configuration characterises, for instance, individual interactions, communication system, sport and team relationships (Carron & Brawley, 2000). A different dynamic arises when one or few influential actors drive the aggregation and shape the whole organisation of the community (Ghalmane et al., 2019). Examples of this latter behaviour are opinion or news spreading in communities where followers are attached to influencers (Medo et al., 2009); epidemic diffusion where few prominent actors govern the outbreak (Wang et al., 2017b), scientific relationships built around the so-called star scientists (De Stefano et al., 2013). Here, the nature of leadership may be associated to various roles which actors carve out within the groups, acting for instance as hubs or brokers.

In this context, Social network analysis (SNA) exploits the framework offered by graph theory to translate these ideas into operational tools: any community is suitably described by a graph where nodes represent the actors and the links between them their interactions, possibly of different strength. A wide range of methods, among which centrality or equivalence measures are just simple examples, have been spawned to express notions of social role and position. A standard accounts is Wasserman and Faust (1994).

While the underlying scope to find groups in network may follow different routes, these are usually defined as locally densely connected set of nodes. The correspondence between groups of subjects and their inner connection density, as well as the possible role of influential individuals within communities, suggest us to extend the ideas underlying the density-based approach for clustering non-relational data to the network framework. The *modal* formulation of this approach associates clusters with the domains of attraction of the modes of the density function underlying the data, namely clusters correspond to dense regions of the sample space. Networks prevent the definition of a probabilistic notion of a density function defined on the nodes, yet the two notions of group are in agreement conceptually. Operationally, modal clustering often resorts to graph theory to detect clusters, which further favours the extension of this formulation to network data. As a fortunate side effect the modal approach allows for identifying a hierarchical structure of clusters, which may catch different degrees of resolution of the groups.

Based on these ideas, the aim of this work was to discuss a method to find clusters of nodes within a network structure, while accounting for relationships of different strength. Consistently with the cluster notion shared by the non-relational density-based approach, we focus on aggregation mechanisms driven by the attraction exerted by influential actors, on the basis of different ‘leadership’ roles as detected by means of alternative node-wise measures. Note that this perspective is largely neglected by the inherent literature, most focusing on the concept of mutual cohesiveness within communities.

The paper is organised as follows. After a brief review of clustering approaches for networks, we overview the modal clustering formulation in metric spaces. Then, we propose its extension to network data, in both the unweighted and weighted framework. The procedure is discussed thoroughly and illustrated on some simple archetypal networks characterised by different community configurations, on a number of benchmark examples with a known community structure, and on a comprehensively complex original dataset. A discussion concludes the paper.

1.2 | Overview of the related literature

Community detection refers to the general aim of partitioning networks in subsets of nodes, which share some common properties in the relational structure. Similarly to the non-relational framework, this task is, in fact, far from being accurately defined. Thus, while the general purpose usually translates into the task of identifying assortative groups with dense inner connections, a different perspective would also include the search of disassortative structures with weaker interactions within, rather than between communities. The lack of a precise definition of cluster, along with the unsupervised nature of the problem, have led on one hand to the proliferation of a voluminous amount of literature on this topic and, on the other hand, to confusing taxonomies of methods designed for the scope. A lack of a consistent terminology has determined expressions as *network* or *graph clustering*, *module*, *block* or *community detection* to be either used interchangeably, or carry slightly different, yet ambiguous, connotations. In this panorama, methods are easier classified on the basis of their technical details and algorithmic implementations (e.g. Fortunato, 2010), which yet disguise the more relevant notion of cluster underlying them. Reviewing all these methods is then an awkward task which we cannot engage without crossing over the scope of the paper. For our purpose, we limit to set some boundaries by providing a coarse overview of the main different goals and motivations for finding groups in networks, and refer back to the insightful review of Rosvall et al. (2019), where the reader will find further details and references. At the same time, we use the terms cluster, community, groups interchangeably in the rest of the paper.

A first widespread approach to find clusters in networks aims to identify densely interconnected nodes compared to the other nodes. Due to the generality of this principle, methods differ in the way it is translated into operational implementations. Several formulations rely on detection of actors or edges with high centrality, as the popular method of Girvan Newman (GN, Newman & Girvan, 2004), a divisive algorithm based on the notion of edge-betweenness. Further methods relying on a similar ground build on the optimisation of the cluster modularity (Danon et al., 2005), so that each community will include a larger number of inner edges than expected by chance. The Louvain method is one of the most popular representative of this category (Blondel et al., 2008). These methods result in cohesive communities where transitivity is high and actors are highly connected to each other inside groups. Notwithstanding, the idea of high density within a group may be also intended as the one arising in star-shaped clusters, where density is concentrated in the figure of some hubs attracting less prominent actors. Evidence of such a theoretical mechanism of aggregation has been explained by Goyal et al. (2006) as a combination of small-world behaviour guided by the presence of interlinked stars. This principle, addressed in the current work, has been largely neglected in SNA. Up to some extent, the only exceptions are the work of Falkowski et al. (2007), relying on the detection of core nodes as the building blocks of the groups, and methods based on *seed set expansion* (see, for a review, Kanawati, 2014), where local, centralised clusters are found around a given number of representatives in binary networks.

A further facet of the clustering problem in networks, known as *cut-based* perspective, aims at partitioning networks in a fixed number of balanced groups with a small number of edges between them, and no guarantees about a possible denser structure of inner connection. In this context, networks are often of a mesh- or grid-like form. Methods in this class refer back to the seminal work of Kernighan and Lin (1970) and often build on the spectrum of the data. Examples are Pothén et al. (1990) and Wang et al. (2017a).

The *blockmodeling* approach follows a different purpose, relying on the concept of node equivalence. Aside from density, groups rely on more general patterns that include disassortative

behaviours where nodes serve similar structural roles in terms of their connectivity profile. A first formalisation dates back to Lorrain and White (1971), while Holland et al. (1983) gave rise to the stochastic counterpart, generalised to the weighted framework by Aicher et al. (2015). See Lee and Wilkinson (2019) for a recent review.

2 | CLUSTERS AS DENSE SETS

2.1 | Modal clustering of non-relational data

Modal clustering delineates a class of methods for grouping non-relational data defined on a metric, continuous space, and building on the concept of clusters as ‘regions of high density separated from other such regions by regions of low density’ (Hartigan, 1975, p. 205) Formally, the observed data $(x_1, \dots, x_n)'$, $x_i \in \mathbb{R}^d$, $i = 1, \dots, n$, are supposed to be a sample from a random vector with (unknown) probability density function f . The modes of f are regarded to as the archetypes of the clusters, which are in turn represented by their domain of attraction.

One route, to identify modal regions, associates the clusters to disconnected density level sets of the sample space, without attempting explicitly the difficult task of mode detection. The key idea is that, when there is no clustering structure, f is unimodal, and any section of f , at a given level λ , singles out a connected (upper) level set: $L(\lambda) = \{x \in \mathbb{R}^d : f(x) \geq \lambda\}$. Conversely, when f is multimodal, $L(\lambda)$ may be either connected or disconnected, depending on λ . In the latter case, it is formed by a number of connected components, each of them associated with a region of the sample space including at least one mode of f . Since a single section of f could not reveal all the modes of f , λ is moved along its feasible range, giving rise to a hierarchical structure, known as the *cluster tree*, which provides the number of connected components for each λ . Each leaf of the tree describes a *cluster core*, defined as the largest connected component of the density level sets which includes one mode. Figure 1 illustrates a simple example of these ideas: cluster cores associated with the two highest modes are identified by the smallest λ larger than λ_3 , while the smallest λ larger than λ_1 identifies two connected components whose one is the cluster core associated to the lowest mode.

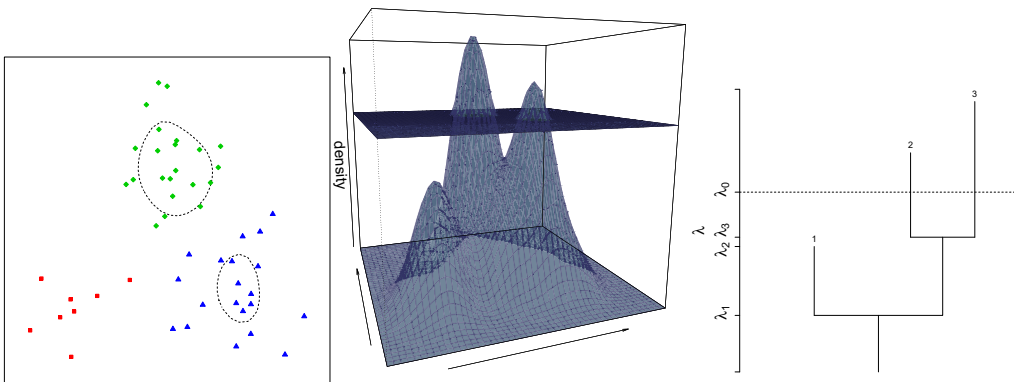


FIGURE 1 A sample from three subpopulations and the associated contour set at a level λ_0 (left). The threshold λ_0 defines a section of the trimodal underlying density function (centre) and identifies two connected regions. On the right, the cluster tree indicates the number of connected components for varying λ and the total number of clusters, corresponding to the leaves [Colour figure can be viewed at wileyonlinelibrary.com]

Note that while the cluster tree resembles a dendrogram, the whole procedure cannot be included in the class of hierarchical techniques. These explore, within the same run, all the partitions with a number of clusters ranging from one to n , by subsequent splits (divisive algorithms) or aggregations (agglomerative algorithms). Conversely, in the cluster tree, the leaves are themselves veritable clusters, instead of single observations, hence their number is the partition cardinality. With respect to a dendrogram, the cluster tree enjoys a different, more insightful interpretation. The height of the leaves corresponds to the density level at which the associated mode appears, thus providing an indication of the cluster prominence. Finally, the hierarchical structure of the tree allows for catching different degrees of resolution of the clustering. In the example illustrated in Figure 1 the number of modes is three, but the two highest ones pertain to the same macro-group, at a lower level of resolution, hence the leaves associated to the two groups collapse to a single branch accordingly.

As the union of the cluster cores is not a partition of the sample space, unallocated points are assigned to the cluster cores according to a supervised scheme of classification, generally accounting for their density.

Operationally, clustering involves two main choices: first, a density estimator is required, typically selected among the nonparametric methods. Second, for each examined threshold λ it is to establish whether the associated level set is connected and what are its components. Since there is no obvious method to identify connected sets in a multidimensional space, graph theory comes to this aid. A graph is built on the sample points and the connected components of the subgraphs induced by the level sets are then detected. The reader is referred to Menardi (2016) for a review on modal clustering.

2.2 | Modal clustering of social networks

2.2.1 | Defining density on networks

For the current formulation, we regard to social networks as undirected graphs $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ consisting of a set $\mathcal{V} = \{v_1, \dots, v_n\}$ of nodes —the actors of the network— and a set $\mathcal{E} = \{e_{ij}\}$ of m links or edges, $i \neq j = 1, \dots, n$, representing relations between pairs of nodes. Depending on the nature of the observed relationships, the elements of \mathcal{E} assume different forms: in binary networks the e_{ij} will take values in $\{0, 1\}$, denoting the absence and the presence of a link, respectively, while real nonnegative values of e_{ij} will account for different strengths of the relationship in weighted networks. In order to represent a given network \mathcal{G} it is possible to define a $n \times n$ adjacency matrix \mathbf{A} whose elements $a_{ij} = e_{ij}$.

The notion of high-density regions highlighted in the previous section suggests a natural counterpart in network analysis, where clusters are often referred to as sets of actors with dense relationship patterns (see, among others, Moody, 2001). However, network objects are subject to an inherent limitation, as their properties can be established in geodesic terms only. In particular, a probabilistic notion of density cannot be defined and shall be intended in a less formal way, reflecting some intuitive meaning of cohesiveness.

We are naturally tempted to borrow the concept of density and akin notions from graph theory. The density of a subgraph $\mathcal{H} \subseteq \mathcal{G}$ is defined as the proportion of all possible edges of \mathcal{H} which are actually observed. In fact, density definition as a node-wise measure is arbitrary as a subgraph \mathcal{H}_v is required to be associated to each node v . For instance, one could set $\mathcal{H}_v = \{\mathcal{V}_v, \mathcal{E}_v\}$ as the subgraph having the nearest neighbours of v as nodes, or focus on the single node $\mathcal{V}_v = v$ and its

incident edges \mathcal{E}_v , thus recasting to the notion of (possible weighted) degree. In fact, consistently with the previous one, a wider set of candidates to quantify local density is represented by measures of connectivity or measures of centrality, which evaluate, somehow, the role as well as the prominence of each actor in a network, but any function defined on the node set \mathcal{V} can be used. It is worthwhile to observe that the choice of a node-wise density measure is not inconsequential with regard to the subsequent interpretation of clusters, and different choices would entail a different concept of cluster. A discussion is provided in Section 3.

2.2.2 | Clustering of unweighted networks

Consider a binary network $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{E} = \{e_{ij}\}$ and $e_{ij} \in \{0, 1\}$. To perform clustering, we select a node-wise measure of density $\delta : \mathcal{V} \mapsto \mathbb{R}^+ \cup \{0\}$ as discussed in the previous section. Afterwards, we may proceed to cluster the nodes according to the modal formulation illustrated in Section 2.1, that is, actors are clustered together when they have density above the examined threshold and they are connected. With respect to the non-relational framework above, we further benefit of the fact that the connected components of the high-density level sets may be identified as the connected components of the induced subgraphs, namely the maximal set of nodes such that each pair of nodes is connected by a path. An operational route is represented by the following scheme:

- (a) Compute the density of the relationships of each actor: $\delta(v_1), \dots, \delta(v_i), \dots, \delta(v_n)$. Clusters will be formed around the modal actors, namely actors with the densest relationship patterns.
- (b) For $0 < \lambda < \max_i \delta(v_i)$:

- Determine the upper level set $\mathcal{V}_\lambda = \{v_i \in \mathcal{V} : \delta(v_i) \geq \lambda\}$,
- Build the subgraph $\mathcal{G}_\lambda = (\mathcal{V}_\lambda, \mathcal{E}_\lambda) \subset \mathcal{G}$ where $\mathcal{E}_\lambda = \{e_{ij}(\lambda)\}$ and

$$e_{ij}(\lambda) = \begin{cases} e_{ij} & \text{if } (v_i, v_j \in \mathcal{V}_\lambda) \\ 0 & \text{otherwise} \end{cases}$$

- Find the connected components of \mathcal{G}_λ .
- (c) Build the cluster tree by associating each level λ to the number of connected components of \mathcal{G}_λ .
 - (d) Identify all the lowest λ for which the branches of the tree represent the leaves, and form the cluster cores as the connected components of the different associated \mathcal{G}_λ .

Essentially, at each threshold λ we evaluate the connected components of \mathcal{G}_λ , the subgraph formed by the nodes with density above λ and the only connections between them. The scheme usually leaves unallocated a number of actors with low-density patterns, when they do not univocally pertain to a modal actor. Depending on the aim of clustering and on subjects-matter considerations, part, or all of them may be either left unallocated or assigned to the cluster for which they present the highest density $\delta(\cdot)$.

The described way of proceeding entails the early identification of clusters as formed by actors with the highest density, that is, the leaders of the community, and the subsequent aggregation to the formed clusters of actors with less prominent role. In this sense, and consistently with the

non-relational version of modal clustering, the final clusters are then described by the domains of attraction of the community leaders.

2.2.3 | Clustering of weighted networks

Let us now consider a weighted network $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{E} = \{e_{ij}\}$ and $e_{ij} \in \mathbb{R}^+ \cup \{0\}$, i.e. the link weight is proportional to the strength of the relationship between the two incident nodes and it is set to zero when the two nodes are not linked.

As a first natural ploy to account for real-valued edges, we consider density measures for weighted networks. Indeed, the generalisation of these measures to weighted networks has been historically a somewhat controversial matter which cannot be tackled without considering the nature of the data, the goal of the analysis, and subject-matter knowledge. However, for most of the mentioned candidate measures δ , there exist a reasonable weighted counterpart. The degree, for instance, is easily extended to measure centrality in weighted networks by summing up the weights incident with each node. This allows considering prominent an actor not only when he has many connections, but also when the strength of these connection is large. We refer the reader to the existing literature for a discussion about the specification of descriptive measures for weighted networks (Opsahl et al., 2010).

In the presence of relationships of different strengths, we need to further adjust the presented procedure. Indeed, a possible weak connection between two high density actors does not appear as a sufficient condition for them to be clustered. Thus, we account for the weights on the basis of the following simple idea: two actors are clustered together when they have density above the examined threshold and they are *strongly* connected. Actors presenting a weak relationship with their neighbours are merged into the same cluster at a lower level of density. Here, the strength of the connection is intended as relative to the set of connections of each node. While this is consistent with the natural idea that prominent actors exercise more influence over their strong connections and less influence over their weak connections, its implementation may take various forms. The following scheme provides two options of possible operational routes:

- (a) Compute the density of each actor, $\delta(v_1), \dots, \delta(v_i), \dots, \delta(v_n)$, with δ an appropriate measure of node-wise density accounting for the weights of the edges;
- (b) For each node v_i , $i = 1, \dots, n$, identify the incident edge with maximum weight $e_{im} = \max_{j: e_{ij} \in \mathcal{E}} e_{ij}$;
- (c) For $0 < \lambda < \max_i \delta(v_i)$:
 - Determine the upper level set $\mathcal{V}(\lambda) = \{v_i \in \mathcal{V} : \delta(v_i) \geq \lambda\}$
 - Build the subgraph $\mathcal{G}_\lambda = \{\mathcal{V}_\lambda, \mathcal{E}_\lambda\}$, where $\mathcal{E}_\lambda = \{e_{ij}(\lambda)\}$ and $e_{ij}(\lambda)$ can be defined according to the two alternative options, denoted by ‘AND’ and ‘OR’ respectively.

option AND

$$e_{ij}(\lambda) = \begin{cases} e_{ij} & \text{if } (v_i, v_j \in \mathcal{V}_\lambda) \cap ((e_{im} = e_{ij}) \cap (e_{jm} = e_{ij})) \\ 0 & \text{otherwise} \end{cases}$$

option OR

$$e_{ij}(\lambda) = \begin{cases} e_{ij} & \text{if } (v_i, v_j \in \mathcal{V}_\lambda) \cap ((e_{im} = e_{ij}) \cup (e_{jm} = e_{ij})) \\ 0 & \text{otherwise} \end{cases}$$

- find the connected components of \mathcal{G}_λ
 - update $e_{im} = \max_{j: e_{ij} \in \mathcal{E} \setminus \mathcal{E}_\lambda} e_{ij}$;
- (d) Build the cluster tree by associating each level λ to the number of connected components of \mathcal{G}_λ .
- (e) Identify all the lowest λ for which the branches of the tree represent the leaves, and form the cluster cores as connected components of the different associated \mathcal{G}_λ .

Essentially, at each λ , we identify the connected components of \mathcal{G}_λ which are formed by the nodes with density above λ . According to ‘AND option’ the additional condition for aggregation is that these nodes represent their reciprocal strongest connection among those not examined yet; conversely, according to ‘option OR’ the condition is loosen by requiring that such connection is the strongest for just one of the actors. The two options, albeit not exhaustive, correspond to different ways of disentangling network complexity and defining the underlying network group structure. With the tight AND option, aggregation is harder to occur, hence leading to a large number of highly homogeneous clusters. The resulting partition is mostly driven by the importance of the relations among nodes rather than by their relative importance within the whole network. According to the ‘OR option’, where the aggregation condition is more frequently satisfied, more parsimonious partitions are created, with clusters mostly driven by the attraction hold by the high density nodes, namely the leaders, on the lower-density ones. Note that this way of proceeding does not guarantee that all the weights are scanned while scanning the density values, that is, at the lowest considered λ , the weakest connections between some pairs of actors might not be accounted for. Since in practice these connections are negligible as, by construction, the weakest ones, we simply circumvent this problem by identifying, at the end of the density scanning, the connected components of the network disregarding the weights of the connections.

The clustering procedure eventually entails the formation of singleton clusters: suppose that three connected nodes u , v , and z have all density above a given λ , but while the strongest relationship of u is with v , the strongest relationship of v is with z and *viceversa*. Then, with the AND option, v and z will fall in the same cluster while u will be a singleton cluster which will be aggregated to the other at a lower λ .

Unallocated actors are finally classified to the cluster core at which they present highest density, like in the unweighted setting.

3 | DISCUSSION

By borrowing ideas from the non-relational version of modal clustering, the procedure illustrated so far allows for disclosing communities as the domains of attraction of network leaders, that is, communities are formed by both high-density actors and the lower density ones on which the attraction is exerted. The operational implementation of such ideas requires the early identification of the leaders, to be intended as the actors with a relatively prominent role in terms of some user-defined density measure of interest. Whenever disconnected (or weakly connected), the leaders will dominate their own community; conversely, (strongly) connected leaders will be assigned to the same community, disregarding the relationships among their followers. The subsequent aggregation of less prominent actors occurs incrementally, by identifying, for each leader, the (strongly) connected components of sequential subgraphs, where all actors have density higher than a moving threshold. Peripheral actors are eventually allocated to the already formed communities, depending on their density, in a further step of the procedure.

Unlike other methods, where the number of communities is either set by the user or determined via some optimality criterion, the cardinality of the partition is here intrinsic to the aggregation mechanism and depends on the local distribution of the density measure among the actors. On the one hand the number of communities is related to the number of leaders, with leadership being a relative notion which depends on the neighbourhood of each actor, even extensible to peripheral actors as long as these are separated from other locally high density actors by lower density ones. On the other hand, connected leaders are in turn aggregated to the same cluster, thus exerting either oligarchic or more democratic forms of domain. The alternation between higher and lower density actors also defines the extent of separation between clusters, in the same guise as the valleys of a density separate modal clusters in non-relational data. Communities will be more or less separated depending on the presence of a large or, respectively, small number of (weakly) connected low density actors separating the leaders.

Beyond the outlined pattern, the resulting partition strongly relies on the choice of the node-wise density that, by measuring the actor prominence, defines the leaderships and hence the cluster membership. The choice cannot neglect subject-matter considerations and the goals of the analysis, and in principle any node-wise function providing indications about the actor roles may be defined. To aid the user selection, we provide here some guidelines about the use, the meaning, and the cluster notion associated to some common measures of centrality, selected for the different community configuration they can disentangle.

Degree centrality evaluates the actor importance in terms of number of relations within the network; in this sense, it is perhaps the most sensible selection to identify central leaderships resulting in star-shaped clusters. Local density measures the fraction of ties one actor sets in his/her neighbourhood, over the maximum number of settable ties. Hence, it favours actors involved in dense relationships and depowers hubs of centralised structures, with a large, yet poorly connected neighbourhood (indeed local density generally exhibits a negative correlation with degree). Its selection makes sense to reveal dense communities, and is not advisable in star-shaped centralised structures where, by highlighting peripheral actors, risks to identify small peer clusters or even singletons. Dense, centralised, structures would be rather detected by measures like the average degree of an actor neighbourhood. Betweenness centrality counts the number of times actors work as bridges to connect other members, and hence evaluates their strategic role in terms of brokerage; revealing leaderships based on brokers may be useful to identify the network ability to spread, for instance, epidemics or unhealthy behaviour (Borgatti, 2006). Groups are built around such bridges which, depending on the network structure may just have the key role of connecting hubs or communities or be themselves hubs as well. In the former case, cross communities are built with respect to hubs, up to their possible merging in a single community. In the latter case the ensuing partitions are similar to the one produced by degree, yet with a possibly lower number of groups.

To clarify the aggregation mechanism caught by the outlined density measures, we illustrate their behaviour on some binary archetypal networks whose structure relies upon the presence of high-density nodes. The simple network displayed in the top row of Figure 2 highlights 4 hubs standing out among the actors. Each hub drives the information flow from and towards six actors having a less prominent role. Density-based clustering built on degree reflects the hub dominance by identifying 4 clusters headed by the leaders (Figure 2, a1). In fact, if the leaders were connected —row(b) of the Figure 2— the clustering configuration would change accordingly, and a single group would be formed by all the followers of the leader dyad (Figure 2, b1). In the lack of hubs —row(c)— modal clustering built on the degree fails to identify groups (Figure 2, c1), which are better identified by alternative node-wise measures accounting for a decentralised leadership.

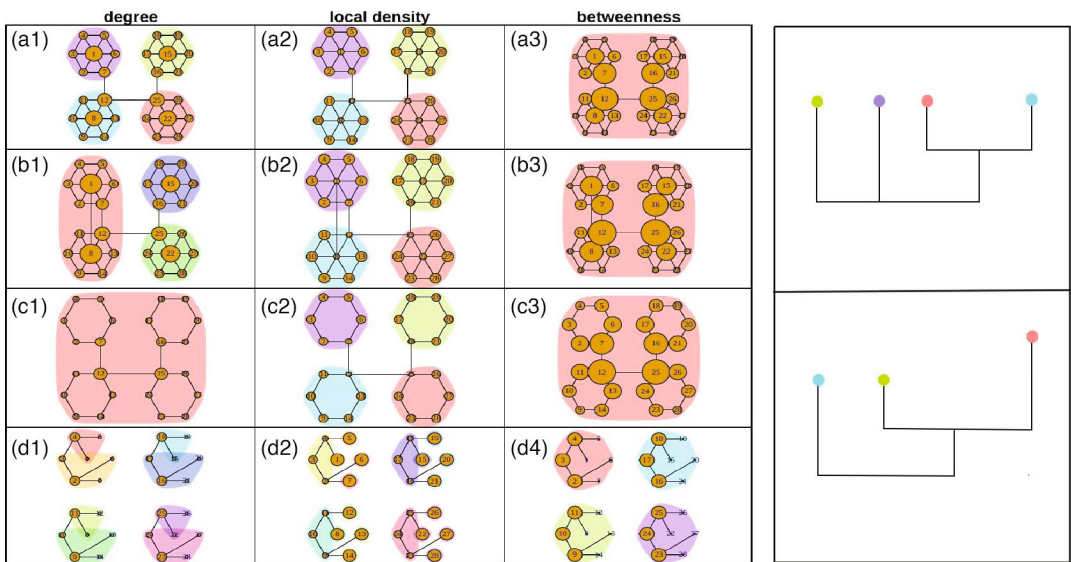


FIGURE 2 At each row a slightly changed version of the same toy network: (a) four hubs not linked directly; (b) two of them are connected by a link; (c) the hubs have been removed; (d) actors are arranged in trees. At each column the clustering built on different measures, with groups marked in different colours and actor size proportional to their density. On the right, the cluster trees associated with the partitions of the connected networks (a, b, c), into 4 and 3 clusters [Colour figure can be viewed at wileyonlinelibrary.com]

By considering local density based on the nearest neighbourhood, modal clustering detects four clusters in all the three versions of the network (second column of Figure 2). Networks arranged in the form of trees —row(d)— are better caught by the betweenness, because of the relevance of bridge actors in hierarchical structures, which would collapse without them. Within the structures highlighted in the first three rows of Figure 2, conversely, the whole community is compact around the brokers, which connect nodes otherwise disconnected in the network. Consistently, density-based clustering built on betweenness detects just one cluster in all the first three versions of the network, led by the connected brokers (panels a3, b3, c3).

The cluster trees of connected networks—the only ones where they can describe the whole structure—provide further information by identifying the hierarchy of the communities. Thus, in the four clusters configurations, the more central communities aggregate first, whereas in the three clusters configuration the first merge occurs between the largest cluster and the closest one (right panel of Figure 2).

In general, since modal clustering relies on the attraction exerted by leaders towards less prominent actors, it catches disassortative mixing (in terms of the considered node-wise measure); in fact, such bias in favour of communities of dissimilar nodes falls whenever the leader themselves are connected, that is, establish homophilic relationships between them. To illustrate these aspects, a further example is displayed in Figure 3, where two mirror sub-communities apparently arise, yet actors exhibit different levels of assortativity. In the top row, a grid-like network connects all pairs of actors within each sub-community, thus showing an assortative behaviour of all members. The homophily exhibited by the two hubs, which connect the two subcommunities, entails both degree- and betweenness-based modal clustering to identify one cluster only (panels a1 and a3). Conversely, local density results in depowering the leaders (and hence their connection) in favour of the more peripheral nodes. In this context, the dense structure of the ties

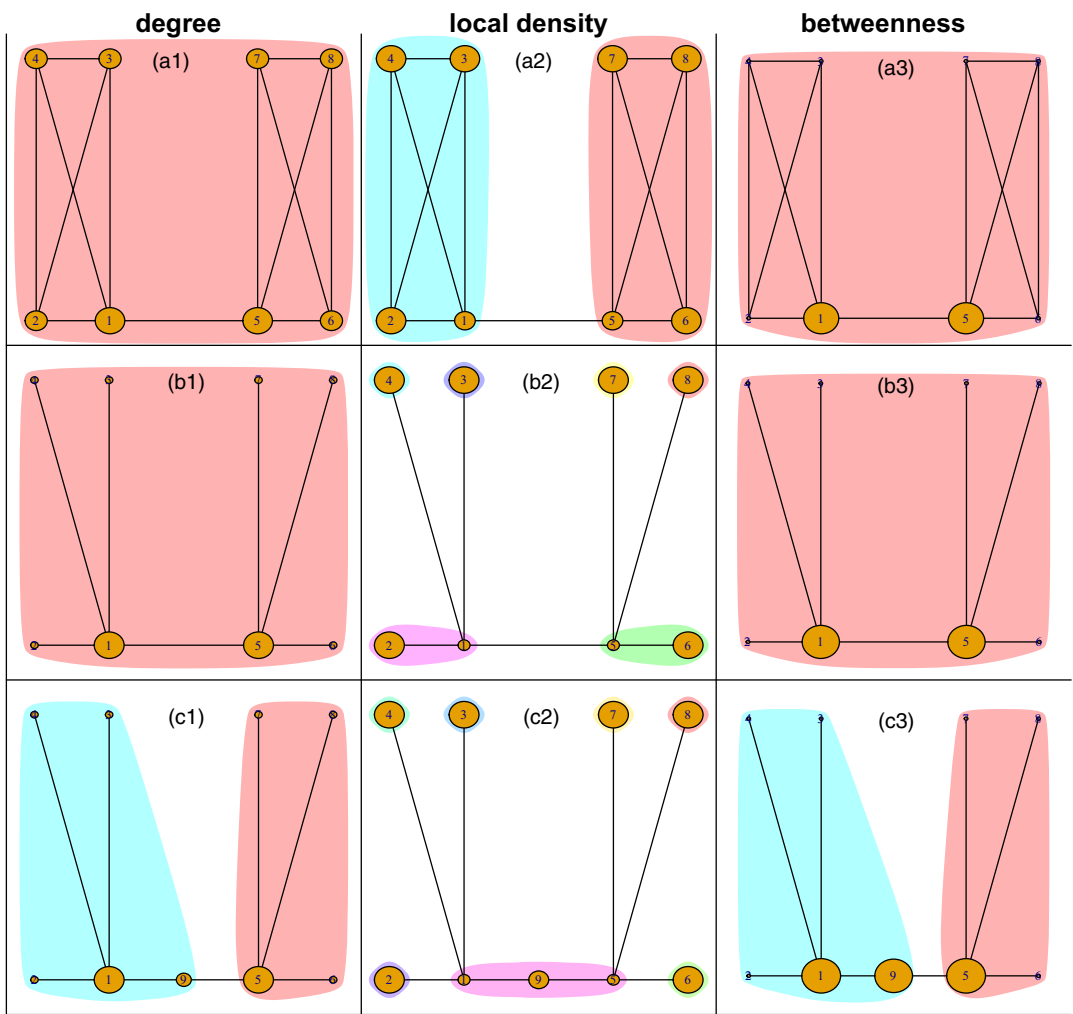


FIGURE 3 At each row a slightly changed version of the same toy network, highlighting different levels of assortativity: the first row displays an assortative network with assortative hubs, the second row displays a disassortative network with assortative hubs, the third one displays a disassortative network with disassortative hubs. Actor size is proportional to their density [Colour figure can be viewed at wileyonlinelibrary.com]

allows to identify the two mirror sub-communities. In row (b) of Figure 3, a general assortativity is prevented by getting rid of peer to peer connections so that actors connect to the hubs only, and form two star-shaped communities. A disassortative behaviour is then shared by all actors but the two connected leaders, which again entail degree and betweenness-based density clustering to produce one cluster only, similarly to the first row. Again, local-density acts by defusing the hubs; however, enhancing the role of peripheral actors in star-shaped structures results in their assignment to small or to singleton clusters (panel b2). The example clearly witnesses the inappropriateness of local-density to describe star-based communities, as it therein tends to enhance actors with limited neighbourhoods. In row (c), a further actor enters the network, acting as a bridge between the two hubs which then change their reciprocal behaviour to be disassortative like all other actors. Due to the unchanged star-shaped structure, local density still produces a

number of singleton clusters; conversely, the lack of a direct link between the hubs allows us to identify the two mirror sub-communities when degree and betweenness are considered (in fact, the latter measure is high for the additional actor as well, yet lower than for the two original brokers). See panels c1 and c3.

4 | REAL DATA ANALYSIS

4.1 | Aims and implementation details

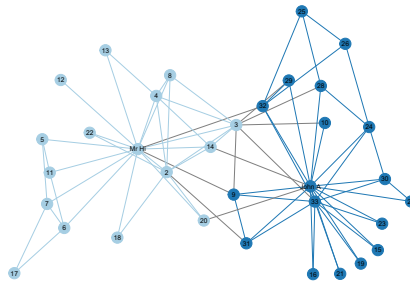
The current section aims to illustrate the aggregation mechanism at the basis of the proposed method, when applied to a number of real networks characterised by different community configurations, also with respect to the selected node-wise measure and in comparison with some other methods.

Consistently with early considerations, we compare results deriving from the application of degree, local density and betweenness as node-wise measures. The considered measures have been adjusted for their use in weighted networks (Opsahl et al., 2010).

The empirical analysis is run also by considering a few diverse community detection methods. We compare with some popular methods, selected because they are frequently considered as a benchmark in community detection: the Girvan Newman (GN) method and its extension to weighted networks, the Louvain method, relying on modularity optimisation to determine the number of communities, and Stochastic Block Models (SBM). Here, Bernoulli and Poisson laws have been assumed on the edges of binary and weighted networks respectively, as in the latter cases weights result from the count of some event. The number of communities is selected via the maximisation of the integrated classification log-likelihood (ICL), as usually employed for the clustering task (Côme & Latouche, 2015). The selected approaches rely on different rationales and community concepts, useful to highlight which framework best suits the application of each method. Among methods relying on similar principle with respect to our approach, we run the DenGraph algorithm (Falkowski et al., 2007) based on the concept of neighbourhood of core nodes (that is a number of nodes that must be reachable by the core nodes within a given distance) around which communities are built, and the Licod algorithm (Yakoubi & Kanawati, 2014) as a representative of the seed-centric approaches, with seeds of communities selected as nodes with degree higher than the $\sigma\%$ of their direct neighbours. Both competitors require a non-trivial specification of some user-defined parameters, such as the number of core nodes (η) and the distance (ϵ) to define their neighbourhood in DenGraph, and the threshold σ in Licod. For DenGraph we vary η and ϵ whose starting values are chosen according to the characteristics of the analysed network, whereas to apply Licod we set σ to its default value.

Note also that DenGraph leaves some nodes unclustered and treated as noise and allows for multiple membership of distinct nodes. Licod, instead, is not designed for weighted networks so we limit the comparison only when edge weights can be neglected.

As a first step, we explore the behaviour of our method in some popular datasets where a ground truth community membership is assigned. The choice of evaluating results in term of a true labelling, rather common in clustering, is motivated by our will of not being biased towards specific community configurations. The agreement between the true and the detected membership has been measured in terms of normalised mutual information (NMI, Danon et al., 2005) which increases for improved quality and associates the maximum value 1 to a perfect agreement. Note, however, that the possible identification of community structures diverse from the defined



binary network							
G-N	Louvain	SBM	DenGraph	Licod	Density-based clustering		
					degree	loc.dens.	betw.
0.58	0.59	0.01	0.56	0.32	1	0.36	1

weighted network										
G-N	Louvain	SBM	DenGraph	Licod	Density-based clustering					
					OR option			AND option		
					degree	loc.dens.	betw.	degree	loc.dens.	betw.
0.56	0.69	0.58	0.17	-	1	0.44	0.85	0.61	0.36	0.42

FIGURE 4 Zachary Karate Club network with true communities marked with different colours. Below, normalised mutual information results of different community detection methods [Colour figure can be viewed at wileyonlinelibrary.com]

true labels would not necessarily imply a failure of the applied method. Such result just reflects that the true clusters have a configuration different from the one that each method is designed to detect, namely different aspects from those explained by the actual labels might, indeed, be highlighted, depending on the network structure, as will be in the following remarked.

As a second step, a more complex, original case study is analysed, to disclose possible clustering structure within the community of the Italian academic statisticians.

The analyses are run in the R computing environment (R Core Team, 2020) with the aid of libraries *igraph* (Csardi & Nepusz, 2006), *sna* (Butts, 2020), *sbm* (Chiquet et al., 2020), and *MUNA* (Falih & Kanawati, 2015). For DenGraph we use the Python library available at <https://pypi.org/project/dengraph/>. The proposed method has been implemented within the DeCoDe R package (Density-based Community Detection), available on the author webpage¹.

4.2 | Benchmark examples

Zachary Karate Club network The well-known Karate Club data (Zachary, 1977) describes the network of friendship between 34 members of a karate club at a US university in the 1970s. The network is in principle weighted, with the strength of connections given by the number of common activities of the club members. In fact, we run the empirical analysis on both the weighted network and on its binary version, built by neglecting the strength of connections. Due to a dispute between the administrator ‘John A’ and the instructor ‘Mr Hi’, the club split into two factions, here representing the benchmark membership. The two factions are then built around the leadership of John A and Mr Hi, which play a special role in terms of both direct influence on the club members, and influence on the information flow to and from the actors. See Figure 4.

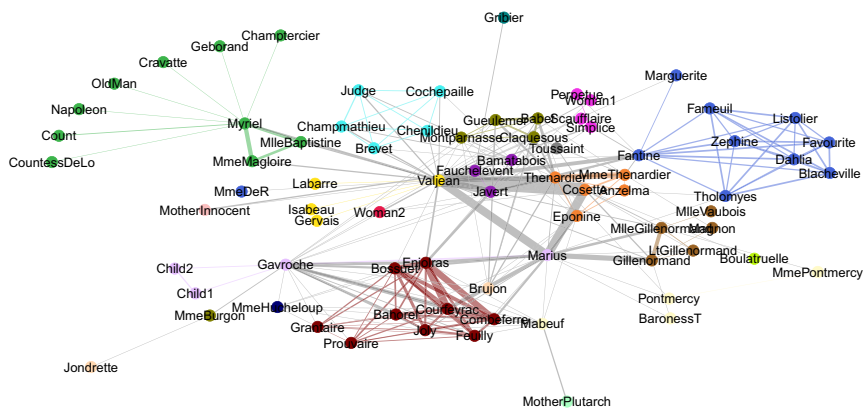
¹<https://homes.stat.unipd.it/giovannamenardi/content/software>

In agreement with these considerations, in the binary setting, an essentially perfect agreement is found between the two factions and the density-based partition detected with both degree and betweenness as node-wise measures. Conversely, local density identifies an unsensibly high number of small communities, hence proves not to be adequate as a node-wise measure to recover star-shaped factions, consistently with the early discussion. Both Louvain and GN methods, identify 4 clusters, yet homogeneous by faction and hence leading to fair values of the NMI. Conversely, due to disassortative mixing, SBM does not reconstruct the benchmark factions, and it allocates the two leaders in the same community. The Licod algorithm identifies 6 clusters each centred around a local leader. Indeed, none of these methods is designed to detect hub headed communities. Instead, similarly to our method, DenGraph (with $\eta = 11$ and $\varepsilon = 1$) detects 2 main clusters but their composition recovers the original benchmark partition in a slightly lesser extent with respect to our approach. We also notice the presence of 3 unallocated nodes.

When considering the weighted network with the OR option, the benchmark factions are again perfectly recovered with the degree used as a node-wise measure. Betweenness overall gives good results as well, although it identifies three community instead of two. Accounting for the link weights, in fact, allows to distinguish a new leader beyond Mr Hi and John A, namely actor 32, with a high prominent bridging role. The inadequacy of local density to find clusters arising from a leadership is confirmed also in the weighted setting, where the 12 detected clusters, albeit accurate in distinguishing members of the two factions, arise again in a too high number. The AND option gives rise, by construction, to a larger number of homogeneous clusters, with the highest density ones still led by John A and Mr Hi. For this reason the NMI stands at decreased values. Disregarding the employed density, in general, the presence of more peripheral actors is enhanced, as with the AND option individual connections are accounted for in clustering formation, rather than the leader influence. In fact, despite the true cluster membership, driven by a forced choice of each actor to line up with one of the leaders, the observed relationships among the peripheral actors are generally stronger than the ones they have with the leaders.

Accounting for the weights enhances the different relationships established by the two leaders with the club members, thus resulting in SBM to assign them to different clusters and produce a partition more faithful to the benchmark. Also the Louvain method produces improved results with respect to the binary case in terms of NMI, as the surplus clusters are therein less populated, while GN stands at about the same level than in the binary counterpart. Dengraph (with $\eta = 4$ and $\varepsilon = 2$) instead finds 4 clusters and the number of unallocated nodes is much larger than in the binary case (12 nodes).

Les Misérables character network This popular network describes the interactions between 77 characters of the Victor Hugo’s novel *Les Misérables* (Knuth, 1993). The network is in principle weighted, with edge strength set to the number of co-appearance of characters in one or more scenes of the novel. Like in the previous example, we also analyse its binary version. With the aim of an objective evaluation, we pursue the assignment of a ground truth membership by associating each character to the book of his/her early appearance. This eventually results in 20 small communities having an assortative attachment mechanism, with some communities formed around more relevant characters and other more cohesive communities (Figure 5). The partition provides an overall fair summary of the novel plot, yet we shall account with some limitations. Beyond three ambiguous references to unnamed actors, the membership of a few main characters should rather have overlapping nature. Hence, results evaluation requires some further insights beyond the mere observation of the NMI values.

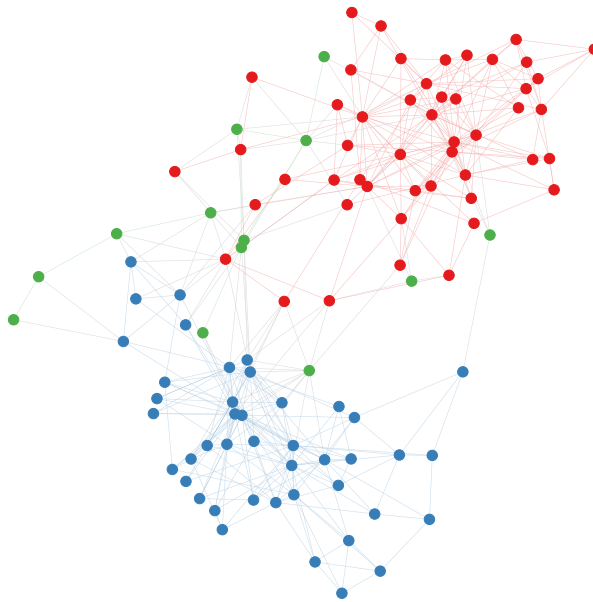


binary network					Density-based clustering			
G-N	Louvain	SBM	DenGraph	Licod	degree	loc. density	betw.	
0.76	0.63	0.54	0.19	0.75	0	0.76	0	

weighted network					Density-based clustering					
G-N	Louvain	SBM	DenGraph	Licod	OR option		AND option			
					degree	loc. dens.	betw.	degree	loc.dens.	betw.
0.63	0.63	0.59	0.27	-	0.48	0.43	0.61	0.76	0.78	0.78

FIGURE 5 Les Misérables characters network. Cf. Figure 4 [Colour figure can be viewed at wileyonlinelibrary.com]

Neglecting the strength of relationships has little impact on the minor characters, generally claiming a small number of weak interactions. Thus, in the binary version of the network, cohesive communities are anyway easily detected, whereby GN, Louvain, Licod and modal clustering based on local density stand out from the other methods at high values of NMI. Conversely, the loss of information on the weights affects the classification of the main characters, all being connected to each other, yet with a different extent which pinpoints their role. Hence, modal clustering with degree and betweenness identifies in the binary network just one community, built around the protagonist Jean Valjean. This is also true for the DenGraph algorithm which moreover classifies as noise 27 out of 77 characters. This is observed with parameter values equal to $\eta = 20$ and $\varepsilon = 1$. It is worth noting that, even changing parameter values, DenGraph keeps recovering one cluster with a slight change on the number of noisy nodes. On the opposite side, SBM places the main character in a singleton cluster, and due to disassortative mixing, creates a large cluster of not necessarily connected supporting characters. When the weight strength is accounted for, this latter cluster is not identified, thus leading to a partition slightly more faithful to the benchmark. Modal clustering works remarkably well with the AND option, which tends to inflate the segmentation and highlights minor groups. The OR option underperforms the AND version compared to the true labels, yet results are anyway highly interpretable. The use of both degree and betweenness gives rise to 6 clusters. In the former case most communities are built around one main character, with a resulting partition very close to those produced by the weighted versions of GN and Louvain. Conversely, when betweenness centrality kicks in, being its values larger for those characters having a protagonist role in multiple books, modal clustering is able to isolate all the main characters in just one group (interpreted as the ‘main plot’ cluster) together with other 5 smaller sized groups of actors whose story is standalone within the whole



G-N	Louvain	SBM	DenGraph	Licod	Density-based clustering		
					degree	loc. density	betw.
0.56	0.51	0.45	0.58	0.37	0.60	0.31	0.07

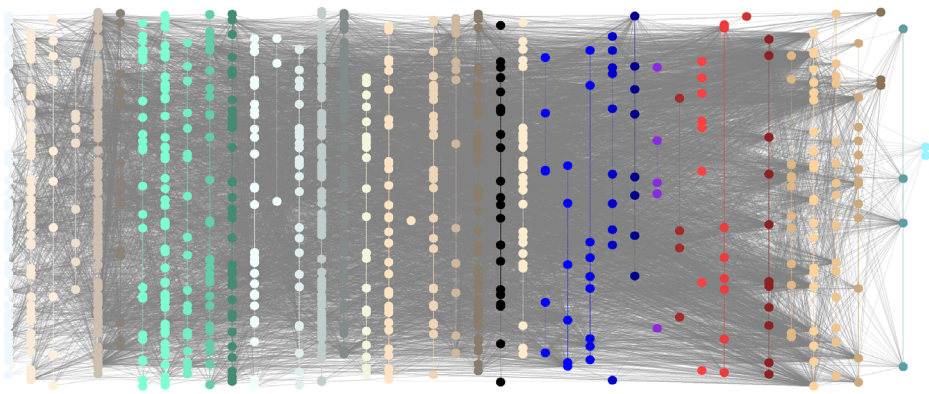
FIGURE 6 US politics books co-purchasing network. Cf. Figure 4 [Colour figure can be viewed at wileyonlinelibrary.com]

plot. DenGraph (with $\eta = 4$ and $\varepsilon = 2$) aggregates a large proportion of characters in 2 clusters (with a residual of 11 noisy nodes), one of them still formed by the large majority of characters associated with Jean Valjean (about 75% out of the total nodeset). Also in this case letting the parameter values vary, we observe a slight change in the number of unallocated nodes.

US politics books co-purchasing network As a further example of star-shaped communities in networks, the US politics books co-purchasing data² include 105 books about US politics published around the presidential election in 2004 and sold online at Amazon.com. The 441 ties between them represent co-purchasing of books by the same buyers. Community membership is given by the book political alignment: liberal, neutral, or conservative. Within communities there exists a slightly centralised organisation of links, especially among liberal and conservative thinkings, with bestsellers representing high-density nodes, often bought in bundle with a variety of less popular other books.

Results (Figure 6) reflect such behaviour, as the density-based partition built on degree centrality outperforms both the other node-wise measures as well as the other methods in terms of NMI and identifies two groups, associated to liberal and conservative alignments. DenGraph (with $\eta = 11$ and $\varepsilon = 2$) stands at a just slightly lower level of NMI. Betweenness density creates a main central cluster which gathers best-sellers from all the political alignment, in the guise of the toy example in Figure 2, and a number of peripheral small clusters of books. Both local-density and the other methods are able to discriminate rather accurately liberal from conservative books,

²<http://www.orgnet.com/>



G-N	Louvain	SBM	DenGraph	Licod	Density-based clustering		
					degree	loc.density	betw.
0.56	0.53	-	0.02	0.57	0.26	0.58	0.26

FIGURE 7 Email-EU-core network network. Cf. Figure 4 [Colour figure can be viewed at wileyonlinelibrary.com]

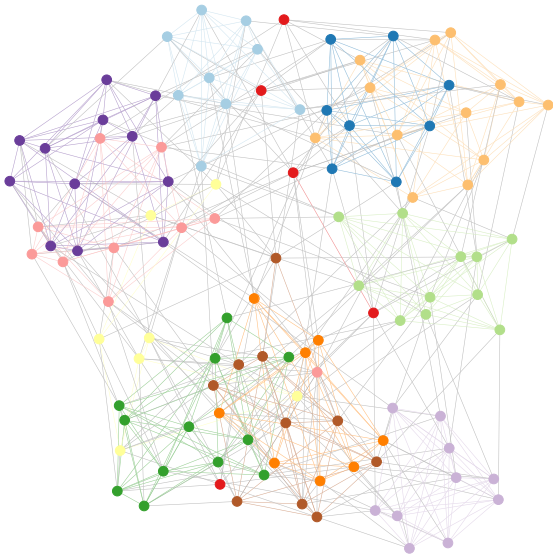
yet with a tendency to oversegment the network (this is especially remarkable with local density). The sole exception is represented by the Licod algorithm which fails to discriminate between liberal and conservative books because it detects 45 tiny sized communities (the largest composed of 10 books and 14 singletons). Without exceptions, methods are not able to identify the least characterised neutral books.

Email-EU-core network The Email-EU-core network (Leskovec et al., 2007; Yin et al., 2017) describes the email exchange between the members of 42 departments of an European research institution. The network is regarded to as undirected by setting an edge whenever there has been at least one either outgoing or incoming email between two members. True clusters are the Departments of affiliation. Distribution of actors among Departments is rather unbalanced, ranging from 1 to 107 individuals. Since the network includes a few isolated nodes, we focus on the giant component only, consisting of 986 individuals (98% of the total) connected by 25552 ties (99.9%). The network is far more complex than the ones examined above. While of difficult inspection, Figure 7 shows that the community configuration is hardly caught by the link description. Research collaborations, indeed, are possibly conducted also by email, and likely not to be limited to the members of a Department. Additionally, there is little evidence of some attachment mechanism guided by the presence of prominent individuals in terms of their degree; also due to the unavailability of weights, conversely, it is likely to expect quite a homogeneous distribution of links within each Department and possible clusters not built around some leaders.

Results confirm the expectations, as local density is the only centrality measure able to catch the gross community structure via density-based clustering, yet detects a number of additional isolated groups. GN and Louvain methods stand on about the same level of accuracy with respect to the benchmark classification. The same holds for Licod which, in fact, consistently with the lack of leadership, identifies one large community and several small than 5 members clusters, partially overlapping with the Department clusters. DenGraph, even varying parameter values, keeps detecting one huge community and few noisy nodes. The application of SBM is computationally unfeasible on this network, due to an inner limitation of the R routines included in the

package `sbm`, which requires the joint estimation of models for any number of communities and the subsequent selection of the best model. Hence, networks with a large number of clusters, like in this case, run into a memory error.

American college football network The American college football network, described by Girvan and Newman (2002), represents the schedule of Division I American football games for the 2000 season. Nodes represent teams and ties between two teams represent regular-season games they dispute. The 115 teams are divided into 12 conferences, representing the benchmark community memberships. In most conferences, inner games are more frequent than games with external teams, with an average of about seven intraconference games and four interconference games in the reference season (Girvan & Newman, 2002, p. 7824). The example is here explored to show the inadequacy of density-based community detection in the lack of leadership. Games configuration, indeed, leads to a grid-like organisation of links within communities. In this situation GN, Louvain and SBM identify 10 out of the 12 benchmark conferences, yet with a notable accuracy where only a few teams are misclassified. In spite of a rather high NMI, Licod algorithm oversegments the network in 24 communities with some singletons. Consistently with a lack of leadership, DenGraph detects just one community without the presence of noisy nodes. This is true with $\eta = 10$ and $\epsilon = 1$, however the result remains substantially unchanged if we let vary the parameter values. Our proposal fails by setting either degree or betweenness as node-wise measure. The former identifies two groups only, each associated to six teams; the latter detects 7 groups whose one gathers most of the teams. Local density, in this case, allows for a slight improvement, where 7 detected clusters are grossly homogeneous by conference or merge pairs of conferences, and three additional groups include a few peripheral teams. See Figure 8 for details.



GN	Louvain	SBM	DenGraph	Licod	Density-based clustering		
					degree	loc.dens.	betw.
0.88	0.89	0.89	0	0.71	0.33	0.55	0.13

FIGURE 8 American college football network. Cf. Figure 4 [Colour figure can be viewed at wileyonlinelibrary.com]

4.3 | Finding clusters within the community of Italian academic statisticians

The aim of the case study here considered is to characterise the scientific community of the Italian academic scholars in Statistics and related fields, via the identification of the clusters formed on the basis of the relationships between them, possibly of different nature and strength, and of the leading aggregation mechanism. This can be useful, for instance, for the creation of new projects and synergies, or more generally, to understand who are, within the community, the leading actors with respect to specific topics.

The main hypothesis underlying the data collection is that, to characterise a researcher within the community, we broadly answer to the questions: *Where does he/she work? What is his/her macro-area of research? Who does he/she work with? What does he/she work on?* As a consequence, we have built a weighted network having in principle a multiplex structure, divided in four layers associated with the questions above: (1) affiliation adjacency matrix (AFF): two actors are connected when they share the same university department affiliation; (2) macro-sector adjacency matrix (MS): two actors are connected when they belong to the same macro-sector, within the area of Statistics and related fields, and as defined by the Italian Ministry of Education, Universities, and Research MIUR (statistics, economic statistics, demography and social statistics, mathematical methods for economy, actuarial and financial sciences); (3) co-authorship network (PUBS): two actors are connected with a link weighted as the number of publications they co-authored; (4) common keywords adjacency matrix (KW): two actors are connected with a link weighted as the number of common keywords in their publications.

Data have been collected in November 2019 and refer to 1160 among professors and researchers of the academic community of statisticians, as recorded by the MIUR database³ where information about the university affiliation and the scientific macro-sector have been drawn. Information about the publications and the keywords have been extracted from the WoS database⁴. Handling the latter one has been troublesome, due to an awkward operation of author matching, especially in the case of homonymy or when a researcher has changed his affiliation at some time and the WoS database does not recognise it. In fact, we shall live with the likely, hopefully not relevant, distortion in the assessment of both the publications and the inherent keywords.

A summarising description of the single layers is provided in Table 1. All networks at the individual layers are composed of the 1160 nodes representing the members of the scientific community under study. Given the exclusivity of the affiliation, the associated network is composed by as many components as the number of observed University departments (namely, 194), within which every actor is connected with all the other actors. The number of researchers within departments is pretty heterogeneous, ranging from 1 to 54. A similar behaviour is observed in the network associated to the macro-sector, where each actor is connected with all other researchers in the same macro-sector. The number of connected components in this layer is equal to the number of considered macro-sectors and these components have diverse sizes (both the statistics and mathematical methods for economy, actuarial and financial sciences areas count more than 400 researchers, while each of the two further sectors count about 150 researchers). The co-authorship layer represents an updated, enriched version of one of the databases employed by De Stefano et al. (2013). We observe 255 isolated researchers, either because they have not published on WoS

³<http://cercauniversita.cineca.it>.

⁴<https://apps.webofknowledge.com>

TABLE 1 Italian academic statisticians network: descriptive statistics for the individual layer networks (AFF - Department affiliation, MS - Macro-sector, CA - Coauthorship network, KW - common keywords) and overall weighted network

	AFF	MS	PUBS	KW	Overall
# of isolated nodes	67	0	255	109	0
# of components	194	4	292	110	1
Network Density	0.014	0.308	0.002	0.523	0.734
Global transitivity	1	1	0.306	0.820	0.833
Degree centralization	0.032	0.082	0.016	0.346	0.240

journals, or because their publications have never been co-authored by any other Italian academic statistician currently on the MIUR list. Their publications have been in any case considered to extract the keywords for the fourth layer of the network, where the number of isolated researchers reduces to 109.

In order to aggregate the four layers into a single, weighted network, we have first normalised the edge weights, measured on different scales, depending on the represented relationship. In principle, there are many procedures to choose among for the purpose. We opt for the simple idea of dividing each weight by the sum of weights within the layer. Then, stemming from the four normalised networks, we have built the associated *overlapping* network (Battiston et al., 2014) by simply summing up the edge weights associated to the same actor across different layers.

The overall network is relatively dense and cohesive with no isolates since all nodes are comprised in the unique component (see Table 1). The strength of the links in the overall network is largely governed by the sparsest co-authorship layer because of the used weighting system.

Among the community detection methods, we have been able to run the Louvain method only. The DenGraph algorithm has not been able to split the network in distinct communities and only returns one single cluster containing all authors and few unallocated nodes (we also tried with different combinations of the η and ϵ parameters but they lead to the same solution), whereas as already mentioned, the Licod algorithm cannot deal with weighted networks. The application of both GN and SBM has turned out computationally unfeasible. It is worth reporting that we run SBM up to the maximum number of communities allowed by our computational resources, that is, 64. The detected partition is unarguably suboptimal, with one community gathering the 2/3 of the actors but the remaining 63 clusters do not differ much from the ones obtained with our procedure. The Louvain algorithm identifies 23 communities, of size ranging from 13 to 102 researchers. Cluster homogeneity with respect to the scientific macro-sector and the affiliation has been evaluated via the complement to one of the Gini index. As for the publications, for each researcher, the proportion of works coauthored by members of the same cluster has been evaluated and the cluster average used as a summarising measure of cluster homogeneity. The same index has been computed for the keywords. To look at the assortative mixing within the detected communities and find if actors within clusters tend to exhibit dense connections among them rather than with actors in different clusters, modularity of the clusters has been also evaluated. Results are summarised in Figure 9. Due to the large size of the detected clusters, clusters are somewhat homogeneous for the scientific sector and the affiliation only, while communities are scarcely associated to co-authorship and research topics. While, by construction, modularity of the Louvain-based partition is maximised, it does not show a remarkably high value. To this respect, it is worth noting that even if the detected partition corresponds to the global maximum

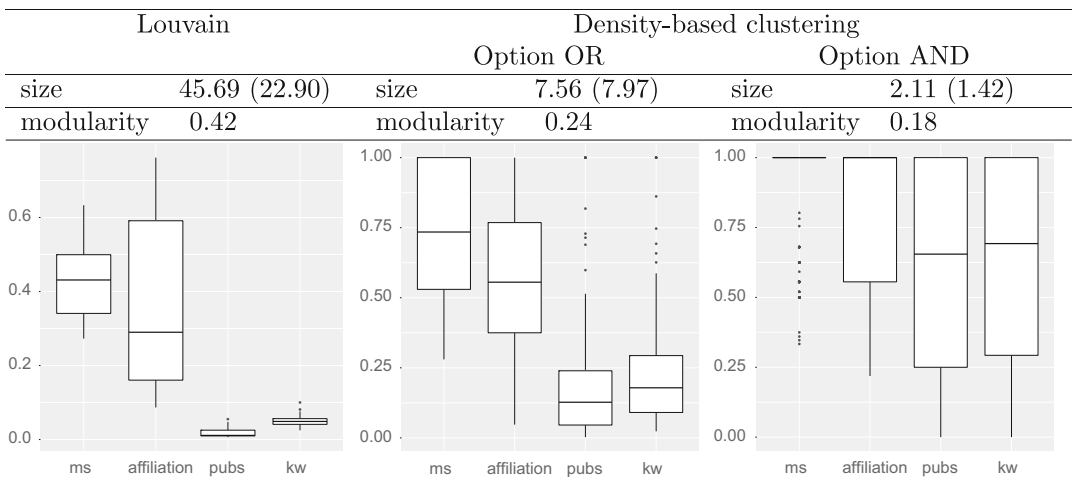


FIGURE 9 Italian community of statisticians. Top panel: average size of clusters (and standard deviation) found via the Louvain method and both the options OR/AND of the density-based method and modularity. The boxplots display the homogeneity of actors across clusters with respect to the considered relationships

of the modularity, in scientific applications this solution is not guaranteed to be more meaningful than the ones obtained by local maxima (Good et al., 2010). Furthermore, results are affected by the so-called resolution limit for which small, plausible communities cannot be identified if the network is large and heterogeneous clusters tend to be formed (Fortunato & Barthélemy, 2007).

Modal clustering has been run building on the degree of the actors, as it appears the most sensible and easiest to interpret choice in such a complex application. Summarising results in terms of size of clusters, modularity, and homogeneity with respect to the considered relationships are reported in Figure 9. The AND option gives rise to 499 clusters, most of them fully homogeneous with respect to the scientific macro-sector and affiliation, and gathering researchers who are known to belong to the same research group. Given the high number of detected communities, an overall limited parsimony is then provided by this partition in terms of summarising description of the network. The partition is yet fairly realistic in the overview of the statistical community where, excluded applied interdisciplinary scholars, researchers tend to work within very small sized teams, and publications are generally co-authored by 2/3 researchers at most. Note that this result is largely acknowledged in social contexts, where groups are limited in size even if social actors are embedded in relatively large networks (Dunbar, 1992; Leskovec et al., 2009).

In fact, if the interest focuses on identifying a more parsimonious partition, the OR option is better suited to the aim, as it identifies 139 groups of size ranging from 4 to 49 scholars. Some heterogeneity with respect to the considered relationships is unavoidable, but clusters are far more homogeneous than those identified by the Louvain method. In fact, while actors working either together or on similar topics tend to be aggregated into the same cluster, the same membership is often shared by other researchers. In this case, cluster aggregation is mostly driven by the attraction hold by a few leaders towards minor actors which often exhibit pretty diverse characteristics.

Clustering is not to be interpreted solely in terms of final group membership - a not that different partition could be trivially obtained by aggregating pairs of maximally connected actors. In fact, the generation process of collaboration among researchers is pretty peculiar: there may be

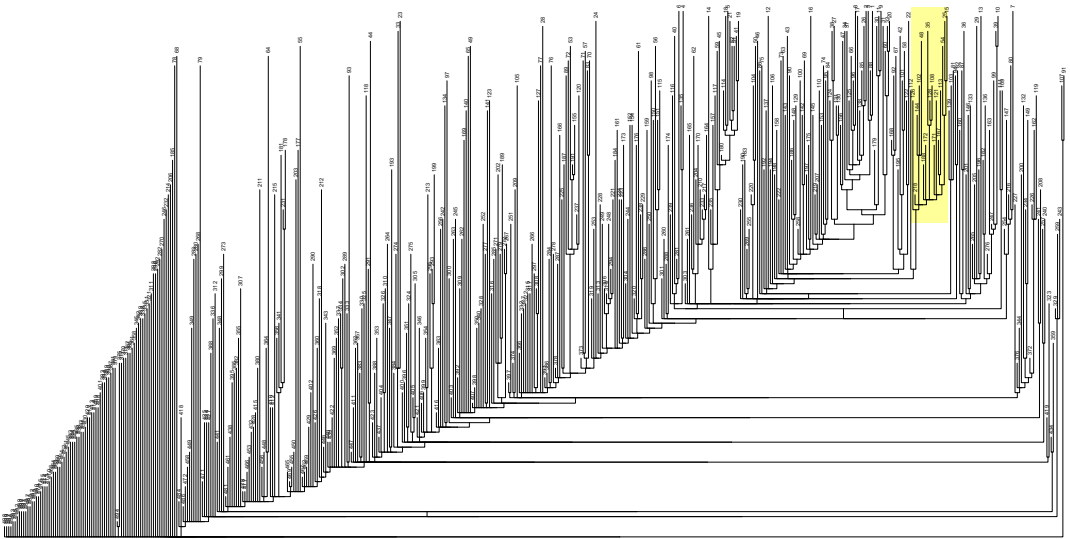


FIGURE 10 Cluster tree of the Italian academic statisticians, obtained with degree as node-wise measure and the AND option. An insight of the highlighted area is provided in Figure 11 [Colour figure can be viewed at wileyonlinelibrary.com]

solitary researchers, sparsely collaborating with other scholars; also, there are researchers who mostly focus on a specific research topic, but also collaborate with different groups of people on a variety of different areas. Both keeping these researchers separated or merging them into the same group may be a stretch. To this aim, a relevant interpretation derives from the exploration of the cluster tree, where clusters are subsequently aggregated at lower levels of the hierarchy, to form larger clusters with a lower resolution (Figure 10, referring to the AND option). For the sake of interpretation, one of its branches, including 17 clusters and a total of 30 researchers, is detailed in Figure 11, along with the associated subnetwork highlighting cluster aggregations at the different levels of the cluster tree. The forming leaves of the branch mostly include either researchers affiliated to the Department of Statistical Science at University of Padova, or scholars who have spent at that Department part of their academic career. Actor aggregation in clusters mostly relies on the strength of connections, hence lead by co-authorship which weights most on the overlapping network. At a lower level of the tree, cluster merging is driven by research topics, with the largest branch on the left associated to likelihood theory, and the other branches including scholars working on its more applied declinations. A link between branches derives from the eclecticism of some of the researchers, working on different research topics. The size of the tree prevents an overall interpretation, but similar traits of homogeneity can be easily identified by picking any branch of the tree. Of course, the lower the level of aggregation of the branches, the lower the homogeneity of the branch.

5 | FINAL REMARKS

Due to the unsupervised nature of the problem, and to the further lack of a ground truth against which to evaluate the quality of a partition, clustering is an ill-posed task, which cannot be performed fully automatically, that is, without some amount of human intervention and disregarding subject-matter considerations.

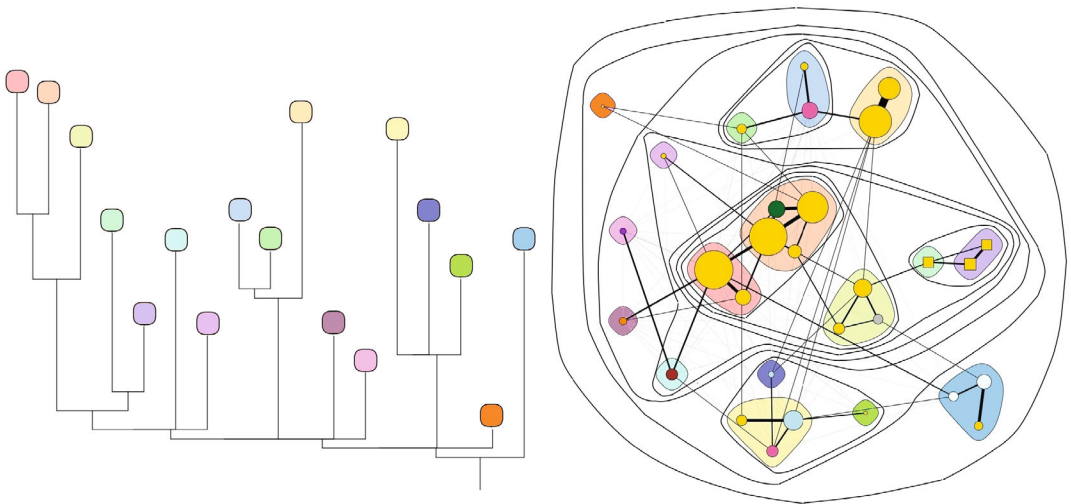


FIGURE 11 Detailed visualisation of the subtree highlighted in Figure 10 and associated subnetwork with clusters marked in different colours and superimposed the cluster aggregations at the different levels of the cluster tree. Actor size is proportional to the their density and different shapes are associated to different macro-sectors. Actor colour is associated to the affiliation. Edge width is proportional to the number of common keywords and coauthored publications [Colour figure can be viewed at wileyonlinelibrary.com]

The methodology here presented makes no exception in the clustering panorama, as it both has required during its planning and still requires the user to make a few thorny choices. A first choice concerns the density measure. The lack of a probabilistic notion of density at a node-wise level implies the loss, for network data, of the probabilistic framework of the original approach defined for non-relational data. Hence, the proposed procedure cannot enjoy the mathematical rigour of other well-known stochastic procedures. On the other hand, the freedom of selecting the measure of density among a wide set of candidates, which quantify connectivity or centrality roles of the actors, ensues. Different group structures arise according to the chosen density measure and those structures account for different aspects of subnetworks cohesiveness, consistently with the intrinsic ill-posedness of the clustering problem. As a downside of such high customisability we acknowledge a certain level of uncertainty, at least when community detection is pursued for exploratory purposes, without having in mind a clustering aim. This has motivated our effort in driving the user choice via a thorough discussion about the community features entailed by different choices of the density measure.


A second choice concerns the way to handle relationships of different strength. Unlike the unweighted framework, there is no obvious way to extend modal clustering in the presence of weighted links. Our strategy aggregates strongly connected individuals at a higher density level than individuals which are weakly connected. While this choice is consistent with the considered aggregation mechanism, based on the most prominent actors exerting influence over their neighbours, the actual implementation of this idea may take various forms. The AND option aggregates two actors with density above a threshold, when they represent their reciprocal strongest connection among those not examined yet. Alternatively, the condition may be loosen via the OR option, by requiring that such connection is the strongest for just one of the actors. A further alternative route would consist in proceeding in a block-sequential manner, aggregating several actors with density above the threshold at a time, as long as their relationship has, at least, a given strength.


The possibility of choosing among these options in cluster formation allows for looking at a given network structure from a different granularity of the representation. As showed in the proposed applications, the OR mechanism tends to minimize the network partition in a smaller number of internal densely connected clusters with loose connections with other clusters. On the other hand, the AND mechanism maximizes the internal homogeneity of clusters detecting a larger number of smaller groups. Here again the choice of the mechanism to handle weights depends on the purpose of the analysis. For instance in the Karate network, the choice of the OR option would reflect an interest in the big picture after collapsing the relations in the community. Conversely, in the Italian statisticians network, the choice of the AND option would entail small scale groups of actors and reflect the purpose of looking for cohesive research clusters.

Although featuring these different options of analysis, the proposed density-based procedure does not suffer from the arbitrariness matters which are typical of standard clustering procedures and does not require to set additional arguments like in DenGraph or seed-centric algorithms. While the number of clusters is determined within the procedure, the partitioning accounts for different levels of cluster resolution, via the group hierarchy provided by the cluster tree. In this sense, the cluster tree represents a somewhat formal instrument to emulate the human cognitive system and allows for getting over the resolution limit of modularity-based methods.

From a computational point of view, the algorithm requires to run $O((n + m)n)$ operations on a binary network, in addition to the ones needed to compute the density, which depend on the selected node-wise measure. While the quadratic growth discourages the use of the procedure with huge networks, we have not experienced system crashes in any of the examples run in the manuscript, and the procedure has proven its feasibility on networks having n in the order of the thousands. According to our experience, detecting communities in networks of such size is conversely prevented by the application of popular methods like Girvan Newman and SBM.

ORCID

Giovanna Menardi  <https://orcid.org/0000-0003-0429-3034>

Domenico De Stefano  <https://orcid.org/0000-0003-1120-2524>

REFERENCES

- Aicher, C., Jacobs, A.Z. & Clauset, A. (2015) Learning latent block structure in weighted networks. *Journal of Complex Networks*, 3, 221–248.
- Battiston, F., Nicosia, V. & Latora, V. (2014) Structural measures for multiplex networks. *Physical Review E*, 89, 32804.
- Blondel, V.D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. (2008) Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008, P10008.
- Borgatti, S.P. (2006) Identifying sets of key players in a social network. *Computational & Mathematical Organization Theory*, 12, 21–34.
- Butts, C.T. (2020) SNA: Tools for Social Network Analysis. Available from: <https://CRAN.R-project.org/package=sna> R package version 2.6.
- Carron, A.V. & Brawley, L.R. (2000) Cohesion: conceptual and measurement issues. *Small Group Research*, 31, 89–106.
- Chiquet, J., Donnet, S. & Barbillon, P. (2020) SBM: Stochastic Blockmodels. R package version 0.2.2. Available from: <https://CRAN.R-project.org/package=sbm>
- Côme, E. & Latouche, P. (2015) Model selection and clustering in stochastic block models based on the exact integrated complete data likelihood. *Statistical Modelling*, 15, 564–589.
- Csardi, G. & Nepusz, T. (2006) The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695, 1–9. Available from: <https://igraph.org>

- Danon, L., Díaz-Guilera, A., Duch, J. & Arenas, A. (2005) Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 2005, 9008.
- De Stefano, D., Fucella, V., Vitale, M.P. & Zaccarin, S. (2013) The use of different data sources in the analysis of co-authorship networks and scientific performance. *Social Networks*, 35, 370–381.
- Dunbar, R. (1992) Neocortex size as a constraint on group size in primates. *Journal of Human Evolution*, 22, 469–493. <http://www.sciencedirect.com/science/article/pii/004724849290081J>
- Falih, I. & Kanawati, R. (2015) MUNA: A multiplex network analysis library. In: *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015, ASONAM '15*. New York, NY, USA: ACM, pp. 757–760. Available from: <http://doi.acm.org/10.1145/2808797.2808804>
- Falkowski, T., Barth, A. & Spiliopoulou, M. (2007) Dengraph: A density-based community detection algorithm. In: *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI'07)*. Washington, DC: IEEE Computer Society, pp. 112–115.
- Fortunato, S. (2010) Community detection in graphs. *Physics Reports*, 486, 75–174.
- Fortunato, S. & Barthélemy, M. (2007) Resolution limit in community detection. *Proceedings of the National Academy of Sciences of the United States of America*, 104, 36–41. Available from: <http://www.jstor.org/stable/25426046>
- Ghalmene, Z., El Hassouni, M., Cherifi, C. & Cherifi, H. (2019) Centrality in modular networks. *EPJ Data Science*, 8, 15.
- Girvan, M. & Newman, M.E.J. (2002) Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99, 7821–7826.
- Good, B.H., de Montjoye, Y.-A. & Clauset, A. (2010) Performance of modularity maximization in practical contexts. *Physical Review E*, 81, 46106.
- Goyal, S., Van Der Leij, M.J. & Moraga-González, J.L. (2006) Economics: an emerging small world. *Journal of Political Economy*, 114, 403–412.
- Hartigan, J. (1975) *Clustering algorithms*. New York: John Wiley & Sons.
- Holland, P.W., Laskey, K.B. & Leinhardt, S. (1983) Stochastic blockmodels: first steps. *Social Networks*, 5, 109–137.
- Kanawati, R. (2014) Seed-centric approaches for community detection in complex networks. In: Meiselwitz, G. (Ed.), *Social computing and social media*, Lecture Notes in Computer Science, vol. 8531. Cham: Springer International Publishing, pp. 197–208.
- Kernighan, B.W. & Lin, S. (1970) An efficient heuristic procedure for partitioning graphs. *The Bell System Technical Journal*, 49, 291–307.
- Knuth, D.E. (1993) *The Stanford GraphBase: a platform for combinatorial computing*. New York: ACM Press.
- Lee, C. & Wilkinson, D.J. (2019) A review of stochastic block models and extensions for graph clustering. *Applied Network Science*, 4, 122.
- Leskovec, J., Kleinberg, J. & Faloutsos, C. (2007) Graph evolution: densification and shrinking diameters. *ACM Transactions on Knowledge Discovery Data*, 1, 2–es.
- Leskovec, J., Lang, K.J., Dasgupta, A. & Mahoney, M.W. (2009) Community structure in large networks: natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics*, 6, 29–123.
- Lorrain, F. & White, H. (1971) Structural equivalence of individuals in social networks. *The Journal of Mathematical Sociology*, 1, 49–80.
- Medo, M., Zhang, Y.-C. & Zhou, T. (2009) Adaptive model for recommendation of news. *EPL (Europhysics Letters)*, 88, 38005.
- Menardi, G. (2016) A review on modal clustering. *International Statistical Review*, 84, 413–433.
- Moody, J. (2001) Peer influence groups: identifying dense clusters in large networks. *Social Networks*, 23, 261–283.
- Newman, M.E.J. & Girvan, M. (2004) Finding and evaluating community structure in networks. *Physical Review E*, 69, 26113.
- Opsahl, T., Agneessens, F. & Skvoretz, J. (2010) Node centrality in weighted networks: generalizing degree and shortest paths. *Social Networks*, 32, 245–251.
- Pothen, A., Simon, H.D. & Liou, K.-P. (1990) Partitioning sparse matrices with eigenvectors of graphs. *SIAM Journal on Matrix Analysis and Applications*, 11, 430–452.
- R Core Team. (2020) *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Available from: <https://www.R-project.org/>

- Rosvall, M., Delvenne, J.-C., Schaub, M.T. & Lambiotte, R. (2019) Different approaches to community detection. In: Doreian, P., Batagelj, V. & Ferligoj, A. (Eds.) *Advances in network clustering and blockmodeling*. New York: John Wiley & Sons, Ltd., pp. 105–119.
- Wang, T.-S., Lin, H.-T. & Wang, P. (2017a) Weighted-spectral clustering algorithm for detecting community structures in complex networks. *Artificial Intelligence Review*, 47, 463–483.
- Wang, Z., Moreno, Y., Boccaletti, S. & Perc, M. (2017b) Vaccination and epidemics in networked populations—an introduction. *Chaos Solitons & Fractals*, 103, 177–183.
- Wasserman, S. & Faust, K. (1994) *Social network analysis: methods and applications*. Cambridge: Cambridge University Press.
- Yakoubi, Z. & Kanawati, R. (2014) Licod: a leader-driven algorithm for community detection in complex networks. *Vietnam Journal of Computer Science*, 1, 241–256.
- Yin, H., Benson, A.R., Leskovec, J. & Gleich, D.F. (2017) Local higher-order graph clustering. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pp. 555–564.
- Zachary, W.W. (1977) An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33, 452–473.