Applied Intelligence (2021) 51:8067–8087
https://doi.org/10.1007/s10489-021-02250-4

Overlapping community detection using core label propagation algorithm and belonging functions

Jean-Philippe Attal[1] · Maria Malek[2] · Marc Zolghadri[3]

Accepted: 28 January 2021 / Published online: 24 March 2021

Abstract
The community detection in complex networks has become a major field of research. Disjoint community detection deals often with getting a partition of nodes where every node belongs to only one community. However, in social networks, individuals may belong to more than one community such as in co-purchasing field, a co-authorship of scientist papers or anthropological networks. We propose in this paper a method to find overlapping communities from pre-computed disjoint communities obtained by using the *core detection label propagation*. The algorithm selects candidates nodes for overlapping and uses *belonging functions* to decide the assignment or not of a candidate node to each of its neighbours communities. we propose and experiment in this paper several belonging functions, all based on the topology of the communities. These belonging functions are either based on global measures which are the density and the clustering coefficient or on average node measures which are the betweenness and the closeness centralities. We expose then a new similarity measure between two covers regarding the overlapping nodes. The goal is to assess the similarity between two covers that overlap several communities. We finally propose a comparative analysis with the literature algorithms.

Keywords Complex networks · Community detection · Overlapping communities · Covers · Topological metrics · Centralities · Similarity measure between covers

1 Introduction

Networks are powerful tools to model complex systems in many fields such as biology (protein-protein interaction), anthropology, sport, etc. Most of the networks representing complex systems contain *communities*. A community is a group of nodes highly interconnected together, but loosely linked to the rest of the graph.

From a general point of view, clustering methods aim to synthesise and summarise observations (or objects) by grouping them in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (or clusters) [1]. Several similarity or dissimilarity measures have been proposed in the literature in order to test the quality of a partition. Clustering techniques are very diverse and they have been continuously developed for over a half century depending upon the optimisation techniques. These algorithms are generally classified in two principle categories: partitional clustering and hierarchical clustering. On the other hand, when objects are connected via a network represented by a graph, a community structure consists of several nodes which shows dense internal connections compared to the rest of the network. The identification of communities hidden within the structure of large network is a challenging problem which has attracted a considerable amount of interest.

Despite the ambiguity in the definition of community, many methods have been proposed for both efficient and effective community detection. Reviews on disjoint community detection are presented in [2–5].

However, some nodes may belong to several communities at the same time. For example, a person usually has connections to several social groups like family, friends, and colleagues; a researcher may be active in several areas.

The research in this area is referred to as *overlapping community detection problem*. In the case of overlapping

✉ Maria Malek
maria.malek@cyu.fr

1 Quartz laboratory, Cergy, France

2 ETIS Lab, UMR 8051, CY Cergy Paris University, ENSEA, CNRS, CY TECH, Cergy, France

3 Quartz laboratory SUPMECA, Saint-Ouen, France

community detection, the set of found communities is called a cover $C = \{c_1, c_2, \ldots, c_k\}$ [6], in which a node may belong to more than one community.

We propose a method to find overlapping communities from pre-computed disjoint communities obtained by using the *core detection label propagation* (CDLP) described in [7]. The algorithm selects candidates nodes for overlapping and uses *belonging functions* to decide the assignment or not of a candidate node to each of its neighbours communities. we propose and experiment in this paper several belonging functions, all based on the topology of the communities. These belonging functions are either based on global measures which are the density and the clustering coefficient [8] or on average node measures which are the betweenness [9] and the closeness centralities [10]. We use then a similarity measure between two covers representing two sets of overlapping communities in order to a draw up experimentally a similarity matrix. This helps to compare and assess the four belonging functions in relations with the properties of the studied network.

The node betweenness centrality is a measure of centrality in a graph based on shortest paths. It represents the degree of which nodes stand between each others. It is measured by the ratio of pairs shortest paths that pass through this node among all shortest paths in the graph. The betweenness centrality is used to identify nodes that control the flows of information between separate parts of the network. It allows also to identify causal nodes that have influence on other entities behaviour, such as genes in genomics or customers in marketing studies. Nodes inside a community have a weaker betweenness centrality than nodes which link several communities together. The betweenness centrality can be used to find out the communities inside a graph. In this paper, we will show that by applying a belonging function to a subset of neighbours communities, it is possible to find those nodes that belong to several communities and to find their covers.

The closeness centrality can be understood as a measure of how long it will take to spread sequentially out information from a node v to all other nodes. The more central a node is, the lower is its total distance to all other nodes; therefore the higher is its closeness centrality. It might be concluded that a node which links several communities has a higher closeness centrality measure that those ones which are deep inside the communities.

We propose in Section 2 a state of the art on overlapping community detection. Section 3, describes the core label propagation algorithm for disjoint communities detection. Section 4, shows our method for finding overlapping communities based on the core label propagation algorithm and the belonging function coupling with the edge density or the average clustering coefficient. There, we expose our new belonging functions based on betweenness and closeness. At the end of this section, we provide a synthetic view of all these belonging functions. In Section 5, we show experimental results on different graphs as well as a comparative analysis with other overlapping community detection algorithms. Finally, Section 6 provides conclusions and perspectives on our current work.

2 Overlapping community detection algorithms

Many previous studies [11–20] showed that the overlap is a significant characteristic of many real-world social networks.

In [18], algorithms for overlapping community detection are reviewed and categorised into five classes related to how communities are identified which are: the clique percolation algorithms, link partitioning, local expansion and optimisation, fuzzy detection and agent based, and dynamical algorithms.

2.1 The clique percolation algorithms

The clique percolation algorithms (CPM) are based on the hypothesis that a community consists of overlapping sets of fully connected subgraphs. Communities detection is thus done by searching adjacent cliques. Firstly, all cliques of size k are identified. Then, a new graph is constructed such that each node represents one of these k-cliques. CPM is suitable for networks that have dense connected parts.

One of the first overlapping community detection algorithms was proposed by Palla et al. based on the search for local patterns, by clique percolation method (CPM) [21, 22]. The algorithm named CFinder [22] is the implementation of CPM, the complexity of the algorithm is polynomial.

However, the algorithm does not terminate for large networks. CPM-like algorithms are more like pattern matching rather than finding communities, since they aim to find specific, localised structure in a network. Shen et al. [23] have proposed EAGLE to detect both the overlapping and hierarchical properties of complex community structures together. This algorithm deals with the set of maximal cliques and adopts an agglomerative framework.

2.2 Link partitioning

This category of algorithms is based on the idea of partitioning links instead of nodes to discover community. A node in the graph is called overlapping if links connected to it are put in more than one cluster. In [24] links are partitioned via hierarchical clustering of edge similarity. A similarity can be computed via the Jaccard Index. Although the link partitioning for overlapping detection seems intuitive, the quality detection is not well argued [25] since these algorithms are based on an ambiguous definition of community.

2.3 Local expansion and optimisation

These algorithms are based on local expansion and optimisation of growing a natural or a partial community [6]. They use a local benefit function that evaluates the quality of a densely connected group of nodes.

Baumes [26] proposed a two-step process. First, the algorithm RankRemoval is used to rank nodes according to some criterion. Then, the process iteratively removes highly ranked nodes until obtention of small, disjoint cluster cores. These cores are used as seed communities for the second step of the process, named Iterative Scan (IS). In this step the cores are expanded by adding or removing nodes until a local density function cannot be improved.

LFM [6] expands a community from a random seed node to form a community by using a fitness function based on the internal and external degree of the community as well as a resolution parameter controlling the size of the communities. After finding one community, LFM randomly selects another node not yet assigned to any community to expand a new community. LFM results depend on the resolution parameter. The complexity in the worst-case complexity is $\mathrm{O}\left(n^2\right)$.

OSLOM [25] tests the statistical significance of a cluster with respect to a global null model during community expansion. To grow the current community, the r value is computed for each neighbour, which is the cumulative probability of having the number of internal connections equal or larger than the number of connections from a neighbour into this community in the null model. The worst-case complexity in general is $\mathrm{O}\left(n^2\right)$.

In [27], authors propose to find a set of good seeds by using two strategies called Graclus centers and Spread hubs. The Graclus centers seeding is based on a kernel distance based on kernel k-means with graph clustering objectives. This function allows to locate a good set of nodes as seeds. The idea of Spread hubs seeding is to select an independent set of high node degrees. This strategy is based on real observations about clusters formation around high degree nodes in real-world networks with a power-law degree distribution. The used algorithm to grow a seed set is based on personalised PageRank (PPR) clustering and is named NISE Neighbourhood-Inflated Seed Expansion. Experimental results are good in finding good overlapping communities in real-world networks. Authors show that he NISE algorithm outperforms others overlapping community detection methods.

In [28], the network is transformed into a line graph, then it is divided into communities based on normal node partitioning. The algorithm named ESCA based on edge strength is proposed. ESCA uses the edge strength and the belonging degree to resolve the problems of unreasonable initial community selection. It can also be used to determine overlapping communities in weighted and unweighted networks. ESCA resolves the issues of unreasonable initial community selection and missing nodes by using the concepts of edge strength and belonging degree. Experiments demonstrate that for both unweighted and weighted networks, ESCA does not miss nodes, and the detected communities are closer to the real network community structure.

In [29], the concept of backbone that consists of an edge and the two nodes connected to the edge is proposed. Backbone degree measure three factors related to an edge and two nodes which are the strength of the edge and the similarity of nodes. This helps to characterise the internal structure of the community. A community forest model based on the backbone degree and community expansion is then proposed to detect the internal structure and external boundary of community. The expansion should gradually decreases from the centre of the community to the boundary of the community, backbone degree allows to add new nodes to the community gradually from the centre of the community, until the expansion of the community began to grow bigger, the complexity is O(n + m) approximately, n is the number of nodes and m is the number of edges.

In [30] a new local expansion method for uncovering overlapping communities based on structural centrality is proposed. The idea is to locate structural centres of communities with the structural centrality, and then expand these structural centres with a weighted strategy and a local search procedure. Structural centres are defined as the nodes that have higher density than their neighbours and have a relatively larger distance from nodes with higher densities.

2.4 Fuzzy detection

Fuzzy community detection algorithms use the strength of association between all pairs of nodes and communities. In these algorithms, a soft membership vector [31] is computed for each node. A drawback of such algorithms is the need to determine the dimensionality k of the membership vector. This value can be either provided as a parameter to the algorithm or computed from the data. Wang et al. [16] combined disjoint detection methods with local optimisation algorithms. First, a partition is obtained from any algorithm for disjoint community detection. Communities attempt to add or remove nodes. The algorithm uses the computed difference (called variance) of two fitness scores on a community, either for including a node or removing it.

On the other hand, spectral theory was used to detect overlapping communities in [32]. The principle is based on computing a number of vectors related to the Laplacian matrix representing the graph, and applying on this eigenspace the *Fuzzy C means* (FCM) clustering algorithm. The results are then retranscripted on the graph to get the covers.

2.5 Agent based and dynamical algorithms

This family of algorithms concerns label propagation algorithms [18, 33], in which nodes with same label form a community, has been extended to overlapping community detection by allowing a node to have multiple labels.

The first known method that used label propagation was proposed by [15], namely COPRA. The authors propose to use a vector to maintain the most common labels with the intervention of a probability threshold. The result, is that a node may belong to one or more communities.

In COPRA [15], each node updates its belonging coefficients by averaging the coefficients from all its neighbours at each time step in a synchronous fashion. A parameter v is used to control the maximum number of communities with which a node can associate. The time complexity is O(v*m*log(v*m/n)) per iteration, n is the nodes number and m is the edges number.

SLPA [18] is a general speaker-listener based information propagation process. It spreads labels between nodes according to pairwise interaction rules. Unlike [15, 33], where a node forgets knowledge gained in the previous iterations, SLPA provides each node with a memory to store received information (i.e., labels). The membership strength is interpreted as the probability of observing a label in a nodes memory. One advantage of SLPA is that it does not require any knowledge about the number of communities. The time complexity is O(t*m), linear in the number of edges m, where t is a predefined maximum number of iterations.

Many techniques using label propagation were then defined, such as SPAEM (dynamic label propagation) in [34], BMLPA *balanced multi-label propagation algorithm* in [35], MLPA *multi-label propagation algorithm* in [36], etc.

In [37], authors point out that connecting degree can reflect the community tendency for a node to its neighbour communities, they thus proposed a COPRA based on connecting degree, named COPRA-CD. In COPRA-CD, all nodes are initialised with a unique community identifier and a belonging coefficient setting to 1, each node updates its community identifier by the union of its neighbours labels, the corresponding belonging coefficient is obtained by normalising the sum of the belonging coefficients of the communities over all neighbours. After several iterations, communities that are totally contained by others are removed and disconnected communities are splitted. Experimental results show improvements in quality and best stability for fuzzy networks.

2.6 Other methods

Nepusz et al. [38] modelled the overlapping community detection as a nonlinear constrained optimisation problem which can be solved by simulated annealing methods.

Gregory et al. [39] extends Girvan and Newman's divisive clustering algorithm (GN) [40] by allowing a node to split into multiple copies with CONGA.

Rees et al. [41] proposed an algorithm to extract the overlapping communities from the egonet, which is a subgraph including a center node, its neighbours, and the links around them. Kovacs et al. [42] proposed an approach focusing on centrality based influence functions. Others methods have been developed as link community detection with a nonnegative matrix factorisation method in [43], Evolutionary algorithms were used to find overlapping partition in [44, 45].

In [46], novel method for detecting new overlapping community in complex evolving networks based on node vitality for modelling network evolution constrained by multi-scaling and preferential attachment. First, according to a node's dynamics such as link creation and destruction, node vitality is found by comparing consecutive network snapshots. Then, it is combined with a fitness function to obtain a new objective function. Next, by optimising the objective function, maximal cliques are expanded, overlapping nodes are reassigned, and overlapping community that matches not only the current network but also the future version of the network are found. experiments results show good results for detecting an overlapping community in a real-world evolving network.

2.7 Our proposition

We propose a method to find overlapping communities from pre-computed disjoint communities obtained by using the *core detection label propagation* (CDLP) described in [7]. The algorithm selects candidates nodes for overlapping and uses *belonging functions* to decide the assignment or not of a candidate node to each of its neighbours communities. These belonging functions are either based on global measures which are the density and the clustering coefficient [8] or on average node measures which are the betweenness and the closeness centralities. Figure 1 shows the block diagram of our method. Even if our algorithm is based on label propagation algorithms we can not classify

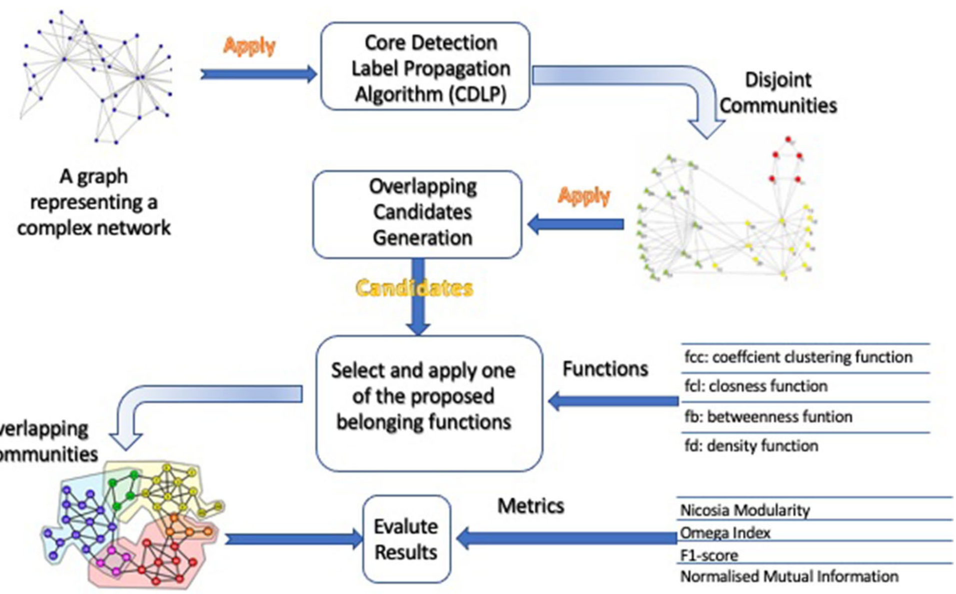

Fig. 1 Block diagram representing our method

it in the category of *agent based and dynamical algorithms* (see Section 2.5), because the process of identifying overlapping communities is not done simultaneously with the propagation of labels. It is based on using proposed belonging functions applied to candidates nodes which are located in the boundaries of disjoint communities. These candidate nodes could make disjoined communities to overlap.

3 Core detection label propagation algorithm(CDLP)

Label propagation algorithm (LPA) is an iterative algorithm based on local information of neighbouring nodes [33]. Let us consider an undirected graph $G = (V, E)$, with V the set of nodes and E the set of edges. The neighbours of the node x are gathered in $V(x) = \{x_1, \ldots, x_k\}$. Through an iterative approach, at each step, every node updates its label according to its immediate neighbours, by a voting mechanism. The label of x is then changed to the label of the majority of the labels of its neighbours. More formally, if c_x stands for x's label and $N^l(x)$ for the set of x's neighbours having the label l, then the new label of x is obtained by $c_x = \arg\max_l |N^l(x)|$. At the end of the process, nodes with the same label form a community. LPA algorithm is similar to some clustering algorithms in the literature and more particularly to the KNNCLUST algorithm [47] where neighbours are computed by using a k-nearest neighbour (knn) density-based rule where the number of clusters is automatically determined. KNNCLUST is based on the combination of nonparametric k-nearest-neighbor (KNN) and kernel density estimation (KNN-kernel). Using the KNN-kernel density estimation allows to model clusters of different densities in high-dimensional data sets. These properties were illustrated via a segmentation application concerning a multispectral image of a floodplain in The Netherlands.

LPA method is applied in a *synchronous* (label propagation is performed in parallel on all nodes) or an *asynchronous* (label propagation step is performed sequentially) way. This method has the disadvantage of bad propagations. A bad propagation usually occurs at the initialisation of the method, when a visit order is given to make the vote on the labels. If there is an equidistribution of the majority labels for a node and if this node connects different communities, the result can either give a giant community (gathering several smaller communities that are detected), or no connected communities (several disjoint communities with the same label). Moreover, the method does not give the same partition at each execution, this is to say that the results of this method are not deterministic. In [33], experimental studies have shown that the asynchronous propagation has a better stability.

One way to stabilise the algorithm consists of applying several times the LPA and observing the nodes that appear often together. In [7], authors proposed a method to stabilise the label propagation with some variations. The method consists of launching $\mathcal{N}$ times the non-deterministic algorithm and creating a matrix $P_{ij}^{\mathcal{N}} = \left[p_{ij}\right]_{n\times n}^{\mathcal{N}}$ such that p_{ij} represents the frequency the nodes i and j appear in the same communities. A new graph $G' = \left(V, E'\right)$ is then created using a threshold $\alpha \in [0, 1]$. Pairs of nodes of the matrix $P_{ij}^{\mathcal{N}}$ having a weight smaller than α are excluded from the set of edges of G'. The connected components created in the graph G' represent the communities. Figure 2 shows a simple example of the obtention of the graph $G' =$

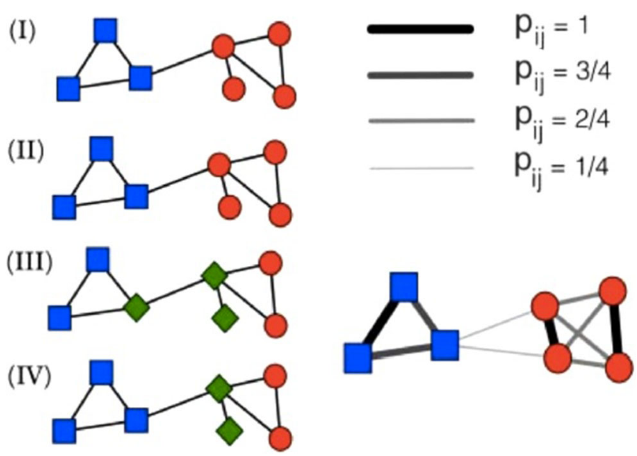

Fig. 2 Example of CDLP algorithm results: communities obtained with $N = 4$ and $\alpha = 0.5$

(V, E') where $\mathcal{N} = 4$ and $\alpha = 0.5$. The choice of α is given by taking the highest modularity score of the lowest conductance score. This algorithm is called *core detection label propagation* (CDLP).

Modularity is one measure of the structure of networks or graphs [48]. It was designed to measure the strength of division of a network into communities. Networks with high modularity have dense connections between the nodes within communities but sparse connections between nodes in different modules. Modularity is often used in optimisation methods for detecting community structure in networks. Given a partition P of a set nodes of the graph $G = (V, E)$ into k Communities, $Q : P \rightarrow [-1, 1]$, is defined by: $Q(P) = \frac{1}{2m}\sum_i \left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta\left(l_i, l_j\right)$, where A_{ij} is the adjacency matrix, k_i is the degree of the node i, m is the number of links in the graph, l_i and l_j are the identifiers of the communities to which belong respectively nodes i and j and $\delta\left(l_i, l_j\right) = 1$ if the nodes i et j are in the same community, 0 otherwise.

On the other hand, the conductance measurement [49] is based on the density of the communities and the number of links leaving them. A community structure is supposed to have a lot of links within it and a low number of outgoing links. The number of outgoing links is denoted by l^c_{out}, and that of inner links l^c_{int} for a community c.

Let c be a community of a graph G, the conductance of this community is defined by $\varphi_{(c,G)} = \frac{l^c_{out}}{[2l^c_{int}+l^c_{out}]}$. Considering a partition $P = \{c_1, \ldots, c_k\}$ in k parts of disjoint nodes, the conductance of G is defined as follows: $\Phi_G = \frac{1}{k}\left[\sum_{c=1}^{k}\varphi_{(c,G)}\right] = \frac{1}{k}\left[\sum_{c=1}^{k}\frac{l^c_{out}}{[2l^c_{int}+l^c_{out}]}\right]$.

The conductance can have a value between 0 and 1. The closer this value is to 0, the more the communities have a high density with few outgoing links.

One of the advantages of CDLP is that by varying α, the hierarchy of the communities can be found. This hierarchy can be shown by a dendrogram. Authors in [7] made experimentations on social networks and showed that low values of α give large communities size and high values of α generate very small communities. Communities in which the nodes remain always together are called the *cores* of communities. The matrix stabilisation method was empirically used in [50] on the Louvain method [51] which is an agglomerative method using a local optimisation of the modularity.

4 Proposed belonging functions to detect overlapping communities

We propose a methodology to detect overlapping community based on the graph G' obtained thanks to the stabilisation matrix method presented in the previous section. G' is first projected on the original graph G. This means that those edges present in G' but not in G are excluded to preserve the topological structure of the original graph [8]. We use the information stored in $P^{\mathcal{N}}_{ij}$ to weight the graph G. This allows to find out the nodes with a higher probability to be together in communities. At the same time, we obtain the *edge between communities*, noted EBC (the set of the edges linking communities) and nodes connected to different communities. These nodes, connected to several communities by their edges, are the possible *candidates* for forming the overlaps of communities. To know if these nodes can be overlapped, we proposed several belonging functions, all based on the topology of the communities.

The topology of the graphs depends on the studied systems. A network of scientific collaboration will not have exactly the same characteristics as a social network related to music or movies. However, complex networks have some common characteristics.

Many studies, including [52–54], have tried to find all the features related to complex networks. They showed common characteristics concerning the distribution of the degrees of the nodes: a weak number of nodes with strong centrality, a large number of triangles, a weak average distance between each pair of nodes, and the existence of groups of nodes strongly connected together and weakly with the rest of the graph (i.e. the communities).

By using characteristics of complex networks and using belonging functions, our goal is to be able to observe whether a node can belong to one or more communities. In what follows, we will first define four belonging functions and then apply them on an illustrative example.

To illustrate the different methods, we propose to use a small graph $G = (V, E)$, with $V = \{v_1, v_2, v_3, v_4, v_5, v_6,$

$v_7, v_8, v_9, v_{10}, x\}$ and E, the set of edges, see Fig. 3. We use the notation c to designate a community and C to designate a set of communities. The weight on the edges represents the value extracted from the matrix $P_{ij}^{\mathcal{N}}$ after computing 100 label propagations (we chose the value 100 according to our experimental studies). By choosing $\alpha > 0.5$ (witch corresponds to the highest overlapping modularity score), we obtain three communities, $c_1^x = \{v_1, v_2, v_3, v_4\}$, $c_2^x = \{v_9, v_{10}\}$, $c_3^x = \{v_5, v_6, v_7, v_8\}$ with the node x whose membership to different communities is investigated. We focus on node x, which is the most likely to overlap different communities.

Let us consider the set $\{c_1^x, \ldots, c_K^x\}$ of the neighbouring communities of the node x, that is to say all the communities which have an edge with x.

To generate all possible overlap combinations between the node x and its K neighbouring communities, we define a function $gen-candidate\left(\{c_1^x, \ldots, c_K^x\}, j\right) \rightarrow C_j$, where C_j is a set of j communities representing all the possible combinations of j communities among the K neighbouring communities of x The cardinality of C_j is $\begin{pmatrix} k \\ j \end{pmatrix}$.

We obtain the following combinations from the graph G (Fig. 3):

$$\begin{aligned} C_1 &= \left[c_1^x, c_2^x, c_3^x\right] \\ &= [\{v_1, v_2, v_3, v_4\}, \{v_9, v_{10}\}, \{v_5, v_6, v_7, v_8\}] \end{aligned}$$

because $\begin{pmatrix} 3 \\ 1 \end{pmatrix} = 3$ possibilities

$$\begin{aligned} C_2 &= \left[\{c_1^x, c_2^x\}, \{c_1^x, c_3^x\}, \{c_2^x, c_3^x\}\right] \\ &= [\{\{v_1, v_2, v_3, v_4\}, \{v_9, v_{10}\}\}, \{\{v_1, v_2, v_3, v_4\}, \\ &\quad \{v_5, v_6, v_7, v_8\}\}, \{\{v_9, v_{10}\}, \{v_5, v_6, v_7, v_8\}\}] \end{aligned}$$

because $\begin{pmatrix} 3 \\ 2 \end{pmatrix} = 3$ possibilities.

$$\begin{aligned} C_3 &= \left[\{c_1^x, c_2^x, c_3^x\}\right] \\ &= [\{\{v_1, v_2, v_3, v_4\}, \{v_9, v_{10}\}, \{v_5, v_6, v_7, v_8\}\}] \end{aligned}$$

because $\begin{pmatrix} 3 \\ 3 \end{pmatrix} = 1$ possibility.

We use in the rest of the document the function *gen−candidate* and the results of the various possible combinations to illustrate our proposals on the belonging functions.

4.1 Function 1: Belonging function based on the density

The idea is to propose a belonging function based on the following intuition: a node connected with a set of communities having high densities can overlap them.

The density of a community c is given by $d : c \longmapsto [0, 1]$, with $d(c) = \frac{2*|E|}{|V|*(|V|-1)}$, where V is the set of nodes of the community and E is the set of edges relating pairs of V.

In Fig. 3 communities $c_1^x = \{v_1, v_2, v_3, v_4\}$, $c_2^x = \{v_9, v_{10}\}$ and $c_3^x = \{v_5, v_6, v_7, v_8\}$ have a high density: $d\left(c_1^x\right) = 1$, $d\left(c_2^x\right) = 1$ and $d\left(c_3^x\right) = 0.83$.

We consider the density of the communities associated to the weight of the edges which links the node x to them to find overlapping communities [8]. Looking for combinations that maximise densities of communities multiplied by the weights of the stabilisation matrix $\left(P_{ij}^{\mathcal{N}}\right)$, we propose the following belonging function based on the density $f_d\{x, C\} \longmapsto \mathbb{R}_+$:

Definition 1

$$f_d(x, C) = \frac{1}{|C|} \sum_{c \in C} w_{c,x} \times d(c) \quad \bullet$$

where $\sum_{c \in C} w_{c,x}$ represents the weights extracted from the stabilisation matrix $\left(P_{ij}^{\mathcal{N}}\right)$ of the edges linking the node x to the different communities in C, C being a combination in C_j, $d(c)$ being the density of the community c.

Using the information contained in the stabilisation matrix, a node with a strong weight linked to a set of high density communities has more chance to overlap that a node linked to a set of weak density communities with a weak weight in the above formula. Overlapping may also be refused. This could be the case if the edges linking node x to the other communities have a weak weight or if f_d is weak, with a small density.

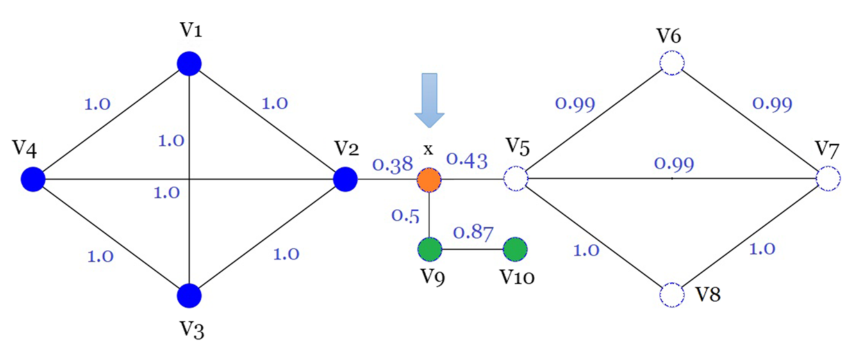

Fig. 3 A graph with 3 obvious communities and a node (x) whose membership in a specific community is questionable

4.2 Function 2: Belonging function based on the clustering coefficient

The clustering coefficient (CC) [55] is a social network measure which deals with the nodes clustering. It computes the probability that two individuals linking to another one are also linked together.

A node connected with a set of communities having high average clustering coefficient could overlap them.

In Fig. 3, the communities:$c_1^x = \{v_1, v_2, v_3, v_4\}$ and $c_2^x = \{v_{10}, v_{11}\}$ have a high average clustering coefficient. By considering a graph $G = (V, E)$ of the Fig. 3 and by noting $CC : G \longmapsto [0, 1]$, the average clustering coefficient function, we get $CC\left(c_1^x\right) = 1$, $CC\left(c_2^x\right) = 0$, $CC\left(c_3^x\right) = 0.833$

We considered the clustering coefficient of the communities associated with the weight of the edges which link the node x to them [8]. The idea consists of assigning nodes to communities with nodes linked to several triangles.

We define the *belonging coefficient clustering measure* $f_{cc} : x \times \{C\} \longmapsto \mathbb{R}_+$ as follows :

Definition 2

$$f_{cc}(x, C) = \frac{1}{|C|}\sum_{c\in C} w_{c,x} \times CC(c) \quad \bullet$$

where $w_{c,x}$ represents the sum of the weights of the stabilisation matrix $\left(P_{ij}^{\mathcal{N}}\right)$, of the edges linking the node x to the different communities in C, C being a combination in C_j, $CC(c)$ being the average clustering coefficient of the community c.

Again, overlapping may be refused. This can be the case if the edges linking a node x to the other communities have a weak weight or clustering coefficient.

4.3 Function 3: Belonging function based on nodes betweenness centrality

The node betweenness centrality is a measure of centrality in a graph based on shortest paths. It represents the degree of which nodes stand between each other. Considering a graph $G = (V, E)$ with V the set of nodes of the graph G and E the set of edges of the graph G, the betweenness centrality of a node v is the sum of the fraction of all-pairs shortest paths that pass through v:

$$g(v) = \sum_{s,t\in V} \frac{\sigma(s, t|v)}{\sigma(s, t)}$$

with $\sigma(s, t)$ the number of shortest paths linking nodes s to t, $\sigma(s, t|v)$ the number of those paths passing through some node v other than s.

Figure 4 shows the nodes betweenness values of the example. The node x, which connects the different communities, has the highest node betweenness.

We propose to use these values to give a high score both to the nodes inside the communities but also to the communities themselves. We compute then the value $1 - b_v$ for each node v, where b_v is the node betweenness of the node v. In Fig. 5, node x has the lowest score.

Overlapping is done over nodes with a low score regarding the sum of the values of the nodes considered for each of the surrounding communities.

We propose a new function based on the reverse of nodes betweenness centrality. The nodes betweenness centrality belonging function for a node x, denoted by f_b, with $f_b : x \times \{C\} \longmapsto \mathbb{R}_+$ and defined as follows:

Definition 3

$$f_b(x, C) = \frac{1}{|C|}\sum_{c\in C} w_{c,x} \times [|c| - g(c)] \quad \bullet$$

in which $g(c) = \sum_{u\in c} g(u)$ $g(S)$ is the sum of the individual nodes betweenness centrality values of the community c, $|c|$ the number of nodes in c. C is a set of the precomputed disjoined communities and $w_{c,x}$ is the weight of the edge linking the node x to the community c.

As for the precedent belonging functions, we allow to refuse an overlapping if the sum of the individual normalised nodes betweenness centrality values in the list of permutation of communities c is not enough strong, i.e. without a real community structure.

4.4 Function 4: Belonging function based on closeness centrality

Closeness centrality [56] measures the mean distance from a node to other nodes. It is a measure of the global centrality of nodes based on the intuition that a node has a strategic or important position in a graph if it is close to the others. In a social network, this measure translates the idea that an actor is important if he is able to easily contact a large number of actors with a minimum of effort (the effort here is relative to the paths length). In practice, the closeness centrality of a node is obtained by computing its average proximity to the other nodes.

Considering a graph $G = (V, E)$, with $|V|$ the number of nodes of the graph G, the closeness centrality of the node $v \in V$ is defined by :

$$cl(v) = \frac{|V| - 1}{\sum_{j=1}^{|V|} dist\left(v, v_j\right)}$$

where v_j is the jth node of the graph.

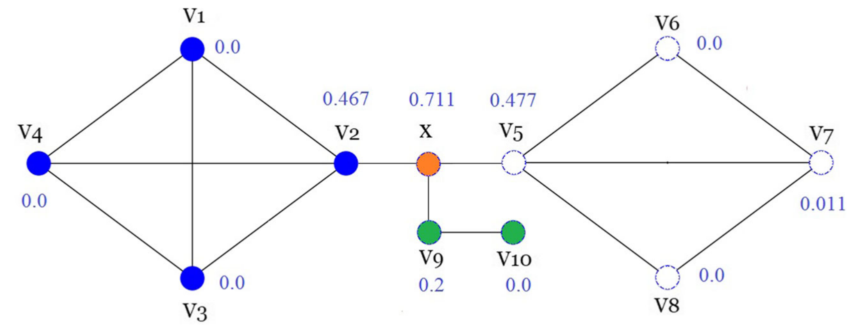

Fig. 4 Node betweenness values

We propose to use the position of the node within the graph and the community topology to which it can be linked to deduce a possible overlap.

Figure 6 shows the nodes closeness centrality values of the example. The node x which connects the different communities and is placed in the centre of the graph has again the highest closeness value.

The position of nodes and the position of communities within the graph are considered to find some overlaps. We compute $1 - cl_v$ for each node v, where cl_v is the node closeness of the node v. A high score is assigned to communities linked to possible nodes for the overlaps (Fig. 7).

We propose the nodes closeness centrality belonging function for a node x, denoted by f_{cl}, with $f_{cl} : x \times \{C\} \longmapsto \mathbb{R}_+$ defined by:

Definition 4

$$f_{cl}(x, C) = \frac{1}{|C|} \sum_{c \in C} w_{c,x} \times [|c| - cl(c)] \quad \bullet$$

where $cl(c)$ is the sum of individual closeness centrality in the community c and $|c|$ the number of nodes in c. C is a set of the precomputed disjoined communities and $w_{c,x}$ is the weight of the edge linking the node x to the community c.

As for the other belonging functions, we allow to refuse an overlapping if the sum of the individual normalised nodes betweenness centrality values in the list of permutation of communities c is not strong enough, i.e. without a real community structure.

For each of the above described functions and for a given node x and a set of communities C, we propose that the overlapping is done if and only if the value of the belonging function is greater or equal than the average values of the topological mesure used in the definition of the belonging function computed for the set of communities C*, where $C* = argmax_{C \in C_j \wedge j \in \{1,\ldots,K\}}(f_{TM}(x, C))$. TM means topological measure and can be one of the following metrics: density, clustering coefficient, betweenness, closeness; $f_T M(x, C)$ is one of the following functions: $\{f_d(x, C)$, $f_{cc}(x, C)$, $f_b(x, C)$, $f_{cl}(x, C)\}$. This means that overlapping is done according to one of the following hypothesis in relation to the used belonging function:

- $f_d(x, C) \geq \frac{1}{|C*|} \sum_{S \in C*} d(S)$ if the density belonging function is used, where $d(S)$ is the density of the community S.
- $f_{cc}(x, C) \geq \frac{1}{|C*|} \sum_{S \in C*} CC(S)$ if the clustering coefficient belonging function is used, where $CC(S)$ is the average clustering coefficient of the community S.
- $f_b(x, C) \geq \frac{1}{|C*|} \sum_{S \in C*} [|S| - g(S)]$ if the betweenness belonging function is used, where $g(S)$ is the sum of individual betweenness centralities in the community S.
- $f_{cl}(x, C) \geq \frac{1}{|C*|} \sum_{S \in C*} [|S| - cl(S)]$, if the closeness belonging function is used where $cl(S)$ is the sum of individual closeness centralities in the community S.

It is also possible to force overlapping. To do so, the domain of variations of j which is greater than 1 (i.e.

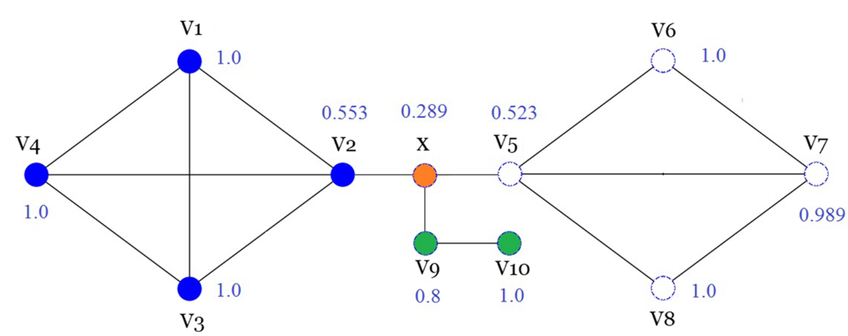

Fig. 5 Nodes score computation using $1 - b_v$ for each node v

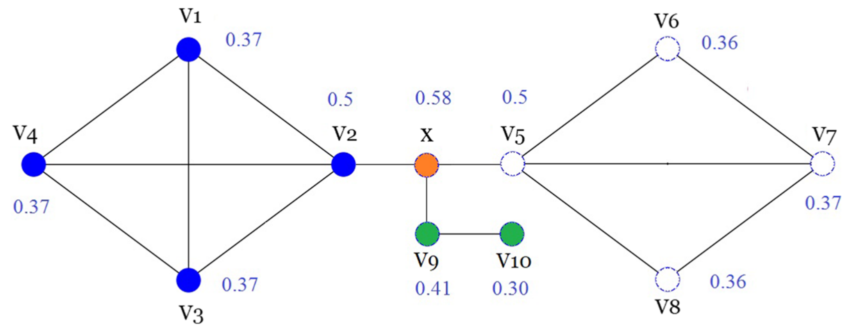

Fig. 6 Node closeness computation

$j \in \{1, \ldots, K\}$) is modified to L, i.e. $j \in \{L, \ldots, K\}$. This will force the node x to belong simultaneously to at least L communities, respecting the topological measures constraint.

4.5 Illustration of the belonging functions

Using the example of the Fig. 3, we propose to compute the four different belonging functions on the set of communities resulting from the core label propagation with the frequency matrix (CDLP) algorithm. The question is to know whether x can belong to several communities.

In Fig. 8, we apply the four different belonging functions. We decompose the computation of each belonging function by calculating separately products between the weights of the edge connecting the node x to the considered community and the used social measures. The best configuration to obtain an overlapping community detection on x is given in bold for each belonging function. Regarding the social measures, the node x will be replicated where the considered social measure is the highest and the relation between x and the other communities, given by the weight $w_{c,x}$ is strong (c being a community linked to x).

5 Evaluation measures, benchmarks, experiments and discussion

To evaluate our algorithms, we use measures exclusively defined for overlapping community detection problem. There are two kinds of measures : the internal measures and the external ones when knowing the ground−truth communities. As internal measure, we use an overlapping version of the *modularity* [57]. As external measures, we use the *normalised mutual information* (NMI) [58] with its extended overlapping version proposed by Lancichinetti et al., the *omega-index*, an overlapping version of the adjusted rand index [59] and the F_1 score. We give also the *edge between communities* (EBC) in percentage and the relative number of communities #.

We also compute the similarity between two covers regarding the nodes which overlap several communities. Considering two covers $C_1 = \{c_1^1, \ldots, c_K^1\}$ and $C_2 = \{c_1^2, \ldots, c_{K'}^2\}$, we define the similarity measure based on overlapping nodes as follows:

Definition 5

$$\sigma : C_1 \times C_2 \to \mathbb{R}$$

$$\sigma(C_1, C_2) = \frac{\left|over_{i=1}^{K}\left(c_i^1\right) \cap \left(over_{i=1}^{K'}\left(c_i^2\right)\right)\right|}{\left|over_{i=1}^{K}\left(c_i^1\right) \cup \left(over_{i=1}^{K'}\left(c_i^2\right)\right)\right|} \quad \bullet$$

where the term $over_{i=1}^{K}\left(c_i^1\right)$ represents the overlapping nodes in the cover C_1.

The closer σ is to 1, the more the overlapping nodes between the two covers are similar. We recall that the computation of the different belonging functions is applied on the results of the CDLP algorithm (see Section 3). As mentioned above, the CDLP algorithm consists of

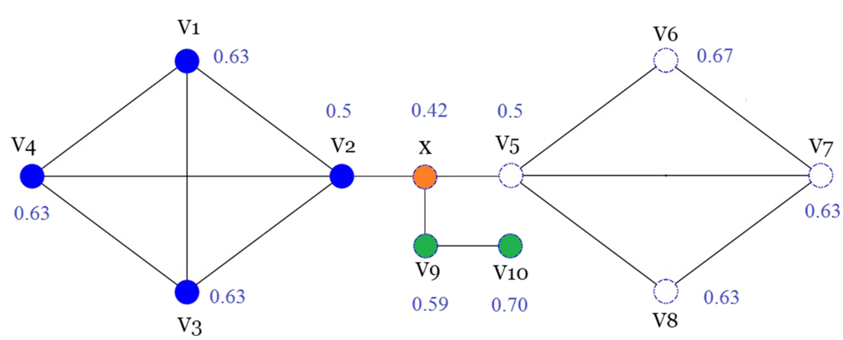

Fig. 7 Scores computation using $1 - cl_v$ for each node v

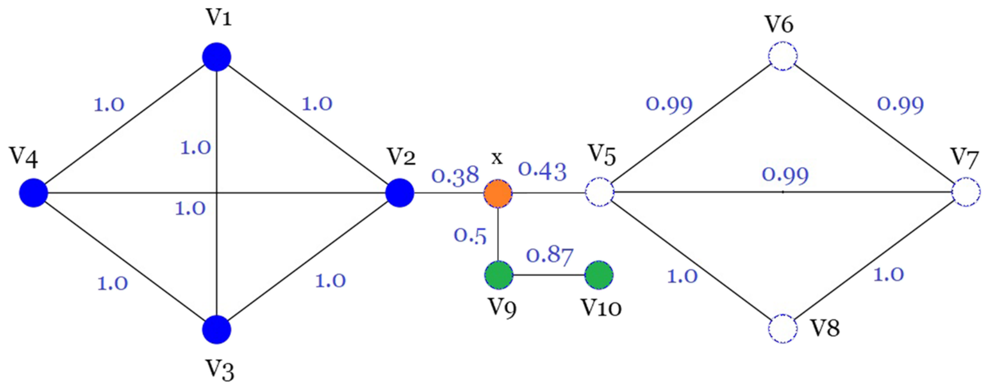

Fig. 8 After computing the functions f_d, f_{cc}, f_b and f_{cl} on the node x, x belongs to different communities depending on the belonging functions. The results in bold give the assignment to the different communities

Combinations	f_{cc}	f_d	f_b	f_{cl}
$\{x, \{c_1^x\}\}$	**0.38**	0.38	1.346	0.908
$\{x, \{c_2^x\}\}$	0	0.5	0.9	0.645
$\{x, \{c_3^x\}\}$	0.357	0.357	**1.512**	**1.0621**
$\{x, \{c_1^x, c_2^x\}\}$	0.19	0.44	1.226	0.78
$\{x, \{c_1^x, c_3^x\}\}$	0.368	0.368	1.429	0.98
$\{x, \{c_2^x, c_3^x\}\}$	0.178	0.428	1.206	0.85
$\{x, \{c_1^x, c_2^x, c_3^x\}\}$	0.246	**0.618**	1.253	0.87

launching $\mathcal{N}$ label propagations and computing a matrix called frequency or stabilisation matrix, $P_{ij}^{\mathcal{N}}$. From the stabilisation matrix, a new graph G' is created using a threshold α, which consists of taking only the edges of the matrix $P_{ij}^{\mathcal{N}}$ whose weight is greater than this threshold. In other words, an edge will be put in the G' graph between two nodes if the frequency of occurrence in the same communities is bigger than the threshold α (see Fig. 2).

To compare two functions with different α values according to the CDLP algorithm, we consider the average of the different α value.

Let $C_{\alpha_i}^{f_X}$ be the resulting cover using the function f_X when applying the CDLP with the parameter α_i and A the set of the different thresholds which can be used with the CDLP, we define the average similarity between two sequences of covers $C_1^{f_X}$ and $C_2^{f_Y}$ as follows: $\bar{\sigma}\left(C_1^{f_X}, C_2^{f_Y}\right) = \frac{1}{|A|}\sum_{\alpha_i \in A} \sigma\left(C_{\alpha_i}^{f_X}, C_{\alpha_i}^{f_Y}\right)$. This measure has a value between 0 and 1. The closer this value is to 1, the more the overlapping nodes of the two sets of covers are similar. This measure does not take into account the replication rate (the number of communities an overlapping node can belong to). In the rest of the paper, we denote C_{f_X} as the cover resulting of the belonging function f_X coupling with the CDLP algorithm. We use the notation $\bar{\sigma}$ to designate the similarity matrix between the possible pairs of the four belonging functions.

We are interested in the characteristics that a node needs regarding its neighbourhood communities to be replicated and the difference in term of results between the different belonging functions. For each of our experiments, we lunch the CDLP algorithm with $\mathcal{N} = 100$ to approach deterministic communities.

Table 1 shows the networks we use for our experimentation: the Zachary Karate Club network [60] (Zac), the football club network [61] (Foot), the political book network [62] (Pol), the dolphins network [63] (Dol) ans coauthorship network of scientists [64] (NS).

5.1 Experiments

Zachary Karate Club We obtain the following results with the Zachary Karate Club.

Figures 9, 10, 11 and 12 represent the visual results obtained for the different belonging functions. For each of the functions, the CDLP gives 2 communities for $\alpha \geq 0.6$, 4 communities for $\alpha \geq 0.7$ and $\alpha \geq 0.8$. In Figs. 9 and 10, the

Table 1 Networks Characteristics

Network	Networks characteristics $\lvert V\rvert$ and $\lvert E\rvert$	Density	D	AT
Zachary	34 \ 78	0.139	5.0	0.256
Foot	115 \ 615	0.094	4.0	0.407
NS	1589 \ 2742	0.002	17.0	0.693
Dol	62 \ 159	0.084	8.0	0.309
Pol	105 \ 441	0.081	7.0	0.348

D is the diameter of the graph and AT is the average transitivity

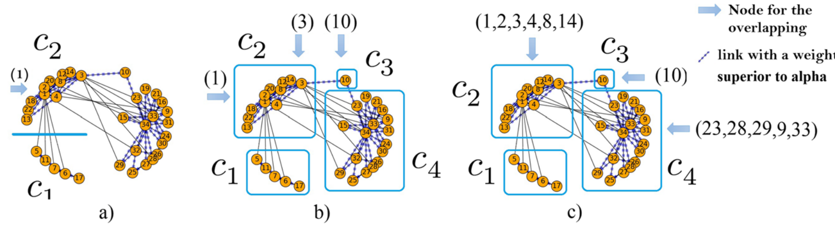

Fig. 9 Graph with different values of α using f_d, **a** $\alpha \geq 0.6$, **b** $\alpha \geq 0.7$ **c** $\alpha \geq 0.8$

node 10 overlaps two communities with f_d and none with f_{cc} for $\alpha \geq 0.7$. This is for $\alpha \geq 0.8$, that node 10 becomes overlapping with f_{cc} function.

The node 3 which is known to be in overlapping communities in the literature, belongs to two communities and is replicated in two communities with f_d in c_3 and c_4, but just one time with f_{cc} (c_4). The node 3 is also detected by f_b and f_{cl} but with a high α value. The node 1 is replicated in one community in c_2 with f_d and f_{cc}.

Table 2 shows that the higher α is, the bigger the number of candidates for the overlapping. Even if the number of candidates is the same until $\alpha \geq 0.9$, the quality of results is better using f_d than f_{cc}. The highest score of the modularity is obtained for $\alpha \geq 0.7$ for each of the methods, with the highest NMI and the highest Ω index. The highest modularity score is obtained with the method based on closeness centrality with 3 communities. The results in terms of quality using all the functions is represented in Table 2.

The similarity matrix between the different covers found according to the four belonging functions is:

$$\sigma = \begin{array}{c} \\ f_d \\ f_{cc} \\ f_b \\ f_{cl} \end{array} \begin{array}{c} \begin{array}{cccc} f_d & f_{cc} & f_b & f_{cl} \end{array} \\ \begin{pmatrix} 1 & 1 & 0.1 & 0.092 \\ 1 & 1 & 0.1 & 0.092 \\ 0.1 & 0.1 & 1 & 0.67 \\ 0.092 & 0.092 & 0.67 & 1 \end{pmatrix} \end{array}$$

We notice that the maximum similarity is reached between covers when using the belonging functions based on density and clustering coefficient. We notice also that the similarity is high enough between covers when using the belonging functions based on betweenness and closeness measures.

On the range [0.6; 0.9], the overlapping nodes are the same with a different rate of assignments to the communities.

Dolphins network The graph is composed of two communities. When $\alpha \geq 0.5$, the system finds the two communities with f_{cc} and f_d, without overlapping community with an NMI, an Omega index and a F_1 score of 1.0 (see Table 3). This is not the case with f_b and f_{cl} where two overlapping are found: those having the highest number of links to the other community they don't belong to. By increasing the value of α, the sizes of the communities decrease and the number of possible candidates for the overlapping increases. We see that the percentage of replicated nodes is the same from $\alpha \geq 0.6$ to $\alpha \geq 0.8$ with f_{cc} and f_d. Nevertheless, the two methods do not replicate the candidate in the same way. f_{cc} function produces the same quality in term of communities than f_d but replicates more nodes for a high value of α. The highest modularity is obtained for $\alpha \geq 0.5$ with f_{cc} and f_d. It turns out the be the ground truth community partition.

The similarity matrix between the different covers found according to the four belonging functions is:

$$\sigma = \begin{array}{c} \\ f_d \\ f_{cc} \\ f_b \\ f_{cl} \end{array} \begin{array}{c} \begin{array}{cccc} f_d & f_{cc} & f_b & f_{cl} \end{array} \\ \begin{pmatrix} 1 & 0.93 & 0.09 & 0.1162 \\ 0.93 & 1 & 0.099 & 0.114 \\ 0.09 & 0.099 & 1 & 0.577 \\ 0.1162 & 0.114 & 0.577 & 1 \end{pmatrix} \end{array}$$

The overlapping nodes between f_d and f_{cc} are very similar but different from those of f_b and f_{cl}. The level of assignment of overlapping nodes to different communities is not the same. The similarity is high but the rate of assignment to the different communities is not the same for f_d and f_{cc}.

Football The football network is known to have 12 communities with high densities.

As shown in Table 4, when $\alpha \geq 0.5$ the first straddling nudges appear for the functions f_{cc} and f_d whereas the first ones overlapping communities appear from $\alpha \geq 0.2$ for the f_{cl} and f_b functions. The best results in terms of partitioning quality are given by f_{cc} and f_d. The higher the value of α,

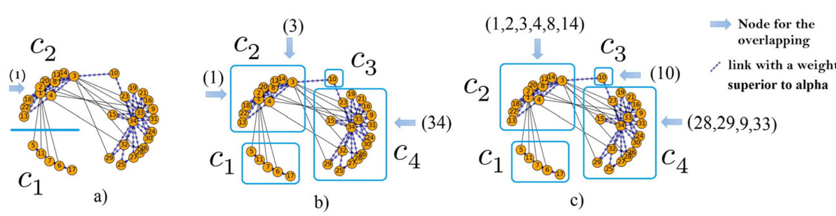

Fig. 10 Graph with different values of α using f_{cc}, **a** $\alpha \geq 0.6$, **b** $\alpha \geq 0.7$ **c** $\alpha \geq 0.8$

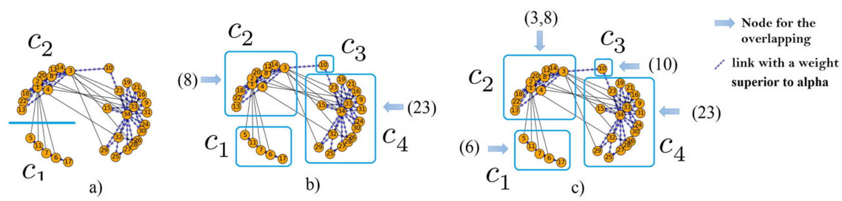

Fig. 11 Graph with different values of α using f_b, **a** $\alpha \geq 0.6$, **b** $\alpha \geq 0.7$ **c** $\alpha \geq 0.8$

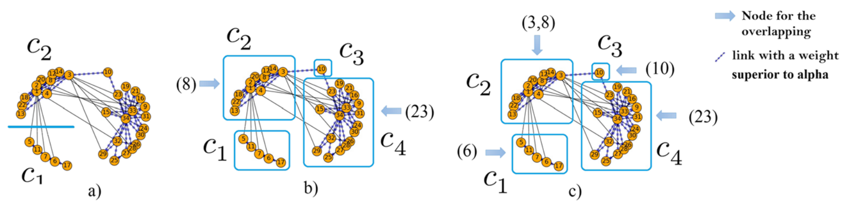

Fig. 12 Graph with different values of α using f_{cl}, **a** $\alpha \geq 0.6$, **b** $\alpha \geq 0.7$ **c** $\alpha \geq 0.8$

Table 2 Results with f_d, f_{cc}, f_{cl} and f_b on Zachary Karate Club

Results with f_d, f_{cc}, f_{cl} and f_b on Zachary Karate Club						
	Cand	EBC				
$\alpha \geq 0.6$	47.058%	17.95%				
$\alpha \geq 0.7$	41.17%	16.0%				
$\alpha \geq 0.8$	55.88%	26.92%				
$\alpha \geq 0.9$	55.88%	26.92%				
f_d	CandOv	Q_{Ov}^{Nic}	Ω	F_1	NMI	#
$\alpha \geq 0.6$	2.94% (1)	0.399	0.064	0.65	0.2365	2
$\alpha \geq 0.7$	8.8235% (3)	0.621	**0.711**	**0.86**	**0.518**	4
$\alpha \geq 0.8$	32.352% (11)	0.42	0.4923	0.75	0.3488	5
$\alpha \geq 0.9$	32.352% (11)	0.42	0.4923	0.75	0.3488	5
f_{cc}	CandOv	Q_{Ov}^{Nic}	Ω	F_1	NMI	#
$\alpha \geq 0.6$	2.941% (1)	0.3986	0.064	0.65	0.2365	2
$\alpha \geq 0.7$	8.823% (3)	0.6210	**0.711**	**0.86**	**0.518**	3
$\alpha \geq 0.8$	32.353% (11)	0.42	0.4923	0.75	0.3488	5
$\alpha \geq 0.9$	32.353% (11)	0.42	0.4923	0.75	0.3488	5
f_b	CandOv	Q_{Ov}^{Nic}	Ω	F_1	NMI	#
$\alpha \geq 0.6$	2.941% (1)	0.064	0.0645	0.65	0.2365	2
$\alpha \geq 0.7$	5.882% (2)	0.68	0.66	**0.86**	0.44	5
$\alpha \geq 0.8$	11.76% (4)	0.62	0.49	0.75	0.31	6
$\alpha \geq 0.9$	11.76% (4)	0.62	0.49	0.75	0.31	6
f_{cl}	CandOv	Q_{Ov}^{Nic}	Ω	F_1	NMI	#
$\alpha \geq 0.6$	2.94%b (1)	0.384	0.25	0.13	0.26	3
$\alpha \geq 0.7$	5.88% (2)	**0.73**	0.66	**0.86**	0.45	3
$\alpha \geq 0.8$	14.70% (5)	0.68	0.549	0.85	0.36	4
$\alpha \geq 0.9$	14.70% (5)	0.59	0.55	0.75	0.35	5

Cand: possible candidates, EBC: percentage of edges between communities, CandOv: Percentage of overlapping nodes, Q_{Ov}^{Nic}: Nicosia modularity, Ω: omega index, F_1: F_1-score, NMI: Normalised Mutual Information, #: communities number

Bold entries correspond to better results

Table 3 Results with f_d, f_{cc}, f_{cl} and f_b on Dolphins network

Results with f_d, f_{cc}, f_{cl} and f_b on Dolphins network						
	Cand	EBC				
$\alpha \geq 0.5$	51.61%	20.38%				
$\alpha \geq 0.6$	54.838%	24.050%				
$\alpha \geq 0.7$	64.51%	30.57%				
$\alpha \geq 0.8$	61.29%	29.30%				
$\alpha \geq 0.9$	77.41%	43.94%				
f_d	CandOv	Q_{Ov}^{Nic}	Ω	F_1	NMI	#
$\alpha \geq 0.5$	0.0%	**0.7959**	**1.0**	**1.0**	**1.0**	2
$\alpha \geq 0.6$	6.451% (4)	0.7502	0.6165	.8571	0.5936	4
$\alpha \geq 0.7$	8.0645% (5)	0.7144	0.4777	0.7499	0.457	5
$\alpha \geq 0.8$	19.355% (12)	0.6052	0.4777	0.6184	0.4421	8
$\alpha \geq 0.9$	25.81% (16)	0.5415	0.3549	0.5333	0.2456	12
f_{cc}	CandOv	Q_{Ov}^{Nic}	Ω	F_1	NMI	#
$\alpha \geq 0.5$	0.0%	**0.7959**	**1.0**	**1.0**	**1.0**	2
$\alpha \geq 0.6$	6.451% (4)	0.7502	0.6125	.8571	0.5936	4
$\alpha \geq 0.7$	8.064% (5)	0.7144	0.4294	0.7499	0.457	5
$\alpha \geq 0.8$	19.355% (12)	0.6062	0.4777	0.6184	0.4421	8
$\alpha \geq 0.9$	35.483% (22)	0.4412	0.5882	0.5489	0.2772	12
f_b	CandOv	Q_{Ov}^{Nic}	Ω	F_1	NMI	#
$\alpha \geq 0.5$	3.22% (2)	0.706	0.93	**1.0**	0.89	5
$\alpha \geq 0.6$	4.83% (3)	0.75	0.56	0.85	0.46	6
$\alpha \geq 0.7$	8.06% (5)	0.71	0.45	0.66	0.425	
$\alpha \geq 0.8$	9.67% (6)	0.644	0.42	0.54	0.37	10
$\alpha \geq 0.9$	19.35% (12)	0.496	0.33	0.5	0.21	12
f_{cl}	CandOv	Q_{Ov}^{Nic}	Ω	F_1	NMI	#
$\alpha \geq 0.5$	3.22% (2)	0.70	**1**	**1**	**1**	2
$\alpha \geq 0.6$	4.84% (3)	0.77	0.64	0.56	0.56	4
$\alpha \geq 0.7$	8.06% (5)	0.71	0.43	0.67	0.45	6
$\alpha \geq 0.8$	9.68% (6)	0.63	0.48	0.54	0.44	8
$\alpha \geq 0.9$	19.35% (12)	0.44	0.37	0.59	0.28	12

Cand: possible candidates, EBC: percentage of edges between communities, CandOv: Percentage of overlapping nodes, Q_{Ov}^{Nic}: Nicosia modularity, Ω: omega index, F_1: F_1-score, NMI: Normalised Mutual Information, #: communities number

Bold entries correspond to better results

the more the value of the quality recovery worsens for all functions. Functions f_{cl} and f_b react poorly to communities with high densities. The highest modularity concerns f_{cc} and f_d with $\alpha \geq 0.2$, giving 9 communities. Some small conferences, with less than 4 teams, are not detected.

The similarity matrix between the different covers found according to the four belonging functions is:

$$\sigma = \begin{array}{c} \\ f_d \\ f_{cc} \\ f_b \\ f_{cl} \end{array} \begin{array}{c} \begin{array}{cccc} f_d & f_{cc} & f_b & f_{cl} \end{array} \\ \begin{pmatrix} 1 & 1 & 0.011 & 0.03 \\ 1 & 1 & 0.026 & 0.019 \\ 0.011 & 0.026 & 1 & 0.836 \\ 0.03 & 0.019 & 0.836 & 1 \end{pmatrix} \end{array}$$

The same observations done on the precedent experiments still verified here: the overlapping nodes between f_d and f_{cc} are very similar (similarity of one) but different from those of f_b and f_{cl}. The rate of assignment of overlapping nodes to different communities is not the same.

Political books of Krebs This network of political books for the 2004 US presidential election was sold on the online sales site Amazon.com. This graph has three communities in the political sense, namely Democrats, Republicans and the center on the political chessboard (Table 5).

The results of the different methods are relatively similar. When α is relatively low ($\alpha \geq 0.4$ and $\alpha \geq 0.5$), the best

Table 4 Results with f_d, f_{cc}, f_{cl} and f_b on football clubs network

Results with f_d, f_{cc}, f_{cl} and f_b on Football Clubs network						
	Cand	EBC				
$\alpha \geq 0.2$	100.0%	29.53%				
$\alpha \geq 0.3$	100.0%	29.53%				
$\alpha \geq 0.4$	100.0%	30.01%				
$\alpha \geq 0.5$	100.0%	30.01%				
$\alpha \geq 0.6$	100.0%	30.83%				
$\alpha \geq 0.7$	100.0%	31.32%				
$\alpha \geq 0.8$	100.0%	31.32%				
f_d	CandOv	Q_{Ov}^{Nic}	Ω	F_1	NMI	#
$\alpha \geq 0.2$	0.0%	**0.722**	0.530	0.762	0.597	9
$\alpha \geq 0.3$	0.0%	0.708	0.681	0.810	0.639	10
$\alpha \geq 0.4$	0.0%	0.7	0 .865	0.854	**0.685**	11
$\alpha \geq 0.5$	0.87% (1)	0.699	0.851	0.854	0.682	11
$\alpha \geq 0.6$	1.74% (2)	0.690	**0.882**	**0.861**	0.666	12
$\alpha \geq 0.7$	8.69% (10)	0.629	0.825	0.819	0.629	13
$\alpha \geq 0.8$	8.69% (10)	0.629	0.825	0.819	0.629	13
f_{cc}	CandOv	Q_{Ov}^{Nic}	Ω	F_1	NMI	#
$\alpha \geq 0.3$	0.0%	0.708	0.681	0.81	0.639	10
$\alpha \geq 0.4$	0.0%	0.699	0.865	0.854	**0.685**	11
$\alpha \geq 0.5$	0.87% (1)	0.699	0.851	0.854	0.682	11
$\alpha \geq 0.6$	1.74% (2)	0.690	0.882	0.85	0.666	12
$\alpha \geq 0.7$	8.69% (10)	0.629	0.825	0.819	0.629	13
$\alpha \geq 0.8$	8.69% (10)	0.629	0.825	0.819	0.629	13
f_b	CandOv	Q_{Ov}^{Nic}	Ω	F_1	NMI	#
$\alpha \geq 0.2$	7.826% (9)	0.61	0.3	0.71	0.37	9
$\alpha \geq 0.3$	8.695% (10)	0.58	0.3538	0.76	0.39	10
$\alpha \geq 0.4$	9.565% (11)	0.565	0.37	0.76	0.39	11
$\alpha \geq 0.5$	9.565% (11)	0.565	0.38	0.77	0.3895	11
$\alpha \geq 0.6$	10.434% (12)	0.555	0.36	0.73	0.3649	12
$\alpha \geq 0.7$	11.304% (13)	0.55	0.36	0.7	0.35	12
$\alpha \geq 0.8$	11.304% (13)	0.55	0.36	0.7	0.35	13
f_{cl}	CandOv	Q_{Ov}^{Nic}	Ω	F_1	NMI	#
$\alpha \geq 0.2$	7.83% (9)	0.61	0.30	0.71	0.37	9
$\alpha \geq 0.3$	8.695% (10)	0.58	0.33	0.76	0.37	10
$\alpha \geq 0.4$	9.565% (11)	0.56	0.37	0.76	0.39	11
$\alpha \geq 0.5$	9.565% (11)	0.56	0.37	0.76	0.38	11
$\alpha \geq 0.6$	10.434% (12)	0.55	0.36	0.73	0.36	12
$\alpha \geq 0.7$	11.30% (13)	0.55	0.36	0.7	0.35	13
$\alpha \geq 0.8$	11.304% (13)	0.55	0.36	0.7	0.35	13

Cand: possible candidates, EBC: percentage of edges between communities, CandOv: Percentage of overlapping nodes, Q_{Ov}^{Nic}: Nicosia modularity, Ω: omega index, F_1: F_1-score, NMI: Normalised Mutual Information, #: communities number

Bold entries correspond to better results

results in terms of recovery are given. The assignment rate of the missing nodes is not the same between the different functions. Politically neutral books are more assigned to the Republican and Democrat communities when using f_b and f_{cl}, rather than the functions f_d and f_{cc}. Books that tell American history over several years as "Ghost wars" by Steve Coll (retracing the history of the CIA for the past fifty years) are assigned to the three communities

Table 5 Results with f_d, f_{cc}, f_{cl} and f_b on Political books network

Results with f_d, f_{cc}, f_{cl} and f_b on Political books network						
	Cand	EBC				
$\alpha \geq 0.4$	24.762%	5.215%				
$\alpha \geq 0.5$	26.67%	6.576%				
$\alpha \geq 0.6$	31.43%	7.71%				
$\alpha \geq 0.7$	32.38%	9.07%				
$\alpha \geq 0.8$	35.24%	9.98%				
f_d	CandOv	Q_{Ov}^{Nic}	Ω	F_1	NMI	#
$\alpha \geq 0.4$	0.0%	0.834	0.667	**0.8**	0.452	2
$\alpha \geq 0.5$	0.95% (1)	0.834	0.654	0.784	0.494	2
$\alpha \geq 0.6$	1.90% (2)	**0.845**	0.676	0.713	0.387	3
$\alpha \geq 0.7$	5.71% (6)	0.76	**0.686**	0.664	0.354	4
$\alpha \geq 0.8$	15.24% (16)	0.653	0.667	0.566	0.290	7
f_{cc}	CandOv	Q_{Ov}^{Nic}	Ω	F_1	NMI	#
$\alpha \geq 0.4$	0.0%	0.834	0.667	**0.8**	0.504	2
$\alpha \geq 0.5$	0.952% (1)	0.834	0.654	0.774	0.494	2
$\alpha \geq 0.6$	0.952% (1)	0.844	0.676	0.719	0.449	3
$\alpha \geq 0.7$	5.714% (6)	0.782	**0.687**	0.653	0.340	4
$\alpha \geq 0.8$	15.238% (16)	0.653	0.667	0.532	0.28	7
f_b	CandOv	Q_{Ov}^{Nic}	Ω	F_1	NMI	#
$\alpha \geq 0.4$	1.90% (2)	0.82	0.63	**0.8**	0.47	2
$\alpha \geq 0.5$	1.90% (2)	0.82	0.63	**0.8**	0.47	2
$\alpha \geq 0.6$	2.857% (3)	0.821	0.629	0.66	0.35	3
$\alpha \geq 0.7$	2.857% (3)	0.816	0.615	0.571	0.30	4
$\alpha \geq 0.8$	5.71% (6)	0.79	0.605	0.4	0.24	7
f_{cl}	CandOv	Q_{Ov}^{Nic}	Ω	F_1	NMI	#
$\alpha \geq 0.4$	0.952% (1)	0.83	0.65	**0.8**	0.49	2
$\alpha \geq 0.5$	0.952% (1)	0.83	0.65	**0.8**	0.49	2
$\alpha \geq 0.6$	2.857% (3)	0.826	0.63	0.66	0.36	3
$\alpha \geq 0.7$	2.857% (3)	0.824	0.62	0.57	0.31	4
$\alpha \geq 0.8$	5.714% (6)	0.798	0.59	0.4	0.24	7

Cand: possible candidates, EBC: percentage of edges between communities, CandOv: Percentage of overlapping nodes, Q_{Ov}^{Nic}: Nicosia modularity, Ω: omega index, F_1: F_1-score, NMI: Normalised Mutual Information, #: communities number

Bold entries correspond to better results

that are Republicans, Neutrals and Democrats. The highest modularity is obtained with f_d. the result consists of 3 communities, representing the 3 main political parties.

The similarity matrix between the different covers found according to the four belonging functions is:

$$\sigma = \begin{array}{c} \\ f_d \\ f_{cc} \\ f_b \\ f_{cl} \end{array} \begin{array}{c} \begin{array}{cccc} f_d & f_{cc} & f_b & f_{cl} \end{array} \\ \begin{pmatrix} 1 & 0.89 & 0.058 & 0.058 \\ 0.89 & 1 & 0.06 & 0.06 \\ 0.058 & 0.06 & 1 & 0.18 \\ 0.058 & 0.06 & 0.18 & 1 \end{pmatrix} \end{array}$$

The overlapping nodes are very similar between f_d and f_{cc} but different from those of f_{cl} and f_b in which the latter are either more prevalent on neutral books or isolated nodes when there are many communities. When nodes alone represent communities and are linked to other communities, the functions f_{cl} and f_b assign them to different communities.

Netscience The Netscience graph is characterised by a very weak density of 0.0021. The results in terms of recovery quality between the different functions are very similar with an overlapping modularity of 0.97 whatever function is used. Nevertheless, in average, there is more overlapping nodes using f_{cl} and f_b rather than f_d and f_{cc}. Regarding Table 6, it is for $\alpha \geq 0.4$ that the first overlapping nodes appear for f_{cl} and f_b, but not with f_d and f_{cc}. Observing the modularity values, results in terms of quality are both similar.

The similarity matrix between the different covers found according to the four belonging functions is:

$$\sigma = \begin{array}{c c} & \begin{array}{cccc} f_d & f_{cc} & f_b & f_{cl} \end{array} \\ \begin{array}{c} f_d \\ f_{cc} \\ f_b \\ f_{cl} \end{array} & \begin{pmatrix} 1 & 0.95 & 0.0645 & 0.078 \\ 0.95 & 1 & 0.076 & 0.085 \\ 0.0645 & 0.076 & 1 & 0.82 \\ 0.078 & 0.085 & 0.82 & 1 \end{pmatrix} \end{array}$$

The overlapping nodes are very similar between f_d and f_{cc} but different from f_b and f_{cl}. The rates of assignment for the overlapping communities are different regarding the used belonging function.

5.2 General observations

Functions based on density and clustering coefficient are computed on connected subgraphs while the betweenness and closeness centralities are computed on nodes the values of which are summed up through the communities.

One of the questions that can arise is about the choice of the belonging function. We saw that the results of the belonging functions depended on the topology of the communities. The more the communities have high densities (the greater the number of links within a community), the more the assignment rate to different communities for a candidate node overlap is important. The functions f_{cl} and f_b are more sensitive to community density than are the functions f_d and f_{cc} for assigning overlapping nodes to different communities. The interest of using the functions f_b and f_{cl} lays in the case of graphs having communities with low densities, where some overlaps are difficult to detect.

5.3 Comparative analysis

We compare our algorithm with the most used algorithms of overlapping communities detection, known to be the referent algorithms for their categories. As mentioned above. Categories (or classes) are described in Section 2. These algorithms are:

- CFinder [21]: a representative of the category *clique percolation*,
- Oslom [25]: representative algorithm of the class *local expansion and optimisation*
- COPRA ($\nu = 2$ and $\nu = 3$), ν being the number of communities to which nodes could belong, [15] and SLPA [65] are based on label propagation algorithm and are representative of the category *agent Based and dynamical algorithms*.

Table 6 Results with f_d, f_b, f_{cc} and f_{cl} on Netscience network

Results with f_d, f_b, f_{cc} and f_{cl} on Netscience network			
	Cand	EBC	
$\alpha \geq 0.2$	3.285%	2.3956%	
$\alpha \geq 0.3$	5.544%	4.1752%	
$\alpha \geq 0.4$	7.392%	5.794%	
$\alpha \geq 0.5$	9.24%	7.7344%	
$\alpha \geq 0.6$	14.1%	12.731%	
$\alpha \geq 0.7$	16.02%	15.332%	
$\alpha \geq 0.8$	18.07%	18.27%	
f_d	CandOv	Q_{Ov}^{Nic}	#
$\alpha \geq 0.2$	0.0%	**0.9768**	293
$\alpha \geq 0.3$	0.0%	0.972	297
$\alpha \geq 0.4$	0.616% (9)	0.948	308
$\alpha \geq 0.5$	0.684% (10)	0.94	315
$\alpha \geq 0.6$	2.396% (35)	0.886	342
$\alpha \geq 0.7$	4.517% (72)	0.845	360
$\alpha \geq 0.8$	6.365% (101)	0.817	371
f_{cc}	CandOv	Q_{Ov}^{Nic}	#
$\alpha \geq 0.2$	0.0%	**0.977**	293
$\alpha \geq 0.3$	0.0%	0.972	297
$\alpha \geq 0.4$	0.5475% (8)	0.949	308
$\alpha \geq 0.5$	0.6160% (9)	0.942	315
$\alpha \geq 0.6$	2.1640% (36)	0.886	342
$\alpha \geq 0.7$	5.3388% (85)	0.855	360
$\alpha \geq 0.8$	7.0499% (112)	0.813	371
f_b	CandOv	Q_{Ov}^{Nic}	#
$\alpha \geq 0.2$	0.616%	0.97	293
$\alpha \geq 0.3$	0.821%	0.96	297
$\alpha \geq 0.4$	1.30%	0.94	308
$\alpha \geq 0.5$	1.848%	0.92	315
$\alpha \geq 0.6$	3.08%	0.88	342
$\alpha \geq 0.7$	5.13%	0.826	371
$\alpha \geq 0.8$	5.13%	0.826	371
f_{cl}	CandOv	Q_{Ov}^{Nic}	#
$\alpha \geq 0.2$	0.684%	0.97	293
$\alpha \geq 0.3$	0.889%	0.96	297
$\alpha \geq 0.4$	1.368%	0.94	308
$\alpha \geq 0.5$	1.916%	0.92	315
$\alpha \geq 0.6$	3.148%	0.87	342
$\alpha \geq 0.7$	5.34%	0.82	371
$\alpha \geq 0.8$	5.34%	0.82	371

Cand: possible candidates, EBC: percentage of edges between communities, CandOv: Percentage of overlapping nodes, Q_{Ov}^{Nic}: Nicosia modularity, #: communities number

Bold entries correspond to better results

- CONGA [39]: an extension of the well known algorithm of Girvan and Newman's divisive clustering algorithm.

Table 7 Comparative analysis

Networks	Comparative analysis					
	F_1	Ω	NMI	Q_{Ov}^{Nic}	#	%
Zac #2						
CFinder	0.48	0.35	0.18	0.52	3	5.88%
OSLOM	**0.86**	**0.84**	**0.80**	**0.748**	2	2.94%
CONGA	0.65	0.113	0.274	0.441	2	2.94%
$COPRA_2$*	0.281	0.266	0.228	0.414	11.3	5.58%
$COPRA_3$*	0.684	0.359	0.347	0.452	6.4	12.64%
SLPA*	0.86	0.633	0.564	0.608	2.12	2.20%
CDLPOV f_d*	0.852	0.711	0.518	0.621	4	8.82%
CDLPOV f_{cc}*	0.852	0.711	0.518	0.621	4	8.82%
CDLPOV f_b*	0.857	0.65	0.44	0.68	5	2.94%
CDLPOV f_{cl}*	**0.86**	0.66	0.45	0.68	3	5.88%
Dol #2						
CFinder	0.57	0.35	0.26	0.66	4	3.72%
OSLOM	**1.0**	0.914	0.852	0.742	2	1.61%
CONGA	0.85	0.892	0.821	0.746	2	3.22%
$COPRA_2$*	0.933	0.788	0.751	0.693	10.8	0.52%
$COPRA_3$*	0.893	0.767	0.701	0.677	3.7	7.73%
SLPA*	0.56	0.754	0.632	0.742	3.44	2.00%
CDLPOV f_d*	**1.0**	**1.0**	**1.0**	**0.796**	2	0.0%
CDLPOV f_{cc}*	**1.0**	**1.0**	**1.0**	**0.796**	2	0.0%
CDLPOV f_b*	0.9	0.93	0.89	0.7	2	3.22%
CDLPOV f_{cl}*	1	1	1	0.7	2	3.22%
Foot #12						
CFinder	0.701	0.64	0.55	0.51	13	6.9%
OSLOM	0.814	0.704	0.55	0.847	2	1.90%
CONGA	0.823	0.321	0.423	0.451	11	60.0%
$COPRA_2$*	0.933	0.788	0.705	0.693	10.8	0.52%
$COPRA_3$*	0.944	0.747	0.712	0.668	11.2	2.52%
SLPA*	0.748	0.684	0.612	**0.715**	10.30	1.69%
CDLPOV f_d*	0.854	0.865	0.751	0.699	11	0.0%
CDLPOV f_{cc}*	0.854	0.865	0.685	0.699	11	0%
CDLPOV f_b*	0.86	0.37	0.39	0.565	11	9.56%
CDLPOV f_{cl}*	0.86	0.37	0.39	0.565	11	9.56%
Pol #4						
CFinder	**0.855**	**0.740**	**0.79**	**0.884**	4	(9)
OSLOM	**0.954**	**0.802**	**0.759**	0.696	12	0.0%
CONGA	0.688	0.651	0.49	0.779	4	4.16%
$COPRA_2$*	0.687	0.637	0.385	0.825	3	1.05%
$COPRA_3$*	0.702	0.649	0.416	0.827	2.8	6.47%
SLPA*	0.755	0.648	0.497	0.83	3.40	12.5%
CDLPOV f_d*	0.784	0.654	0.495	0.844	3	1.90%
CDLPOV f_{cc}*	0.5788	0.667	0.504	0.834	2	0%
CDLPOV f_b*	0.8	0.63	0.47	0.82	2	1.9%
CDLPOV f_{cl}*	0.8	0.65	0.49	0.83	2	0.952%

* algorithms based on the label propagation, F_1: F_1-score, Ω: omega index, NMI: Normalised Mutual Information, Q_{Ov}^{Nic}: Nicosia modularity and # the number of communities

Bold entries correspond to better results

We show the results obtained by our methods having the highest internal score (Q_{Ov}^{Nic} : Nicosia modularity) in Table 7 . Our proposed algorithms give relatively good and competitive results in term of quality but depend of the topology of the graph. For the dolphin graph, the different functions perform well and give a result close to an NMI of one. When graphs have high density communities, functions using clustering coefficient or density give better results than those using betweenness or closeness centrality, what we can observe for the football network. We have a better quality than COPRA, and a better stabilisation for Zachary, Dolphin and Political books. Even if label propagation based algorithms produce more communities, the CDLP with f_d and f_{cc} produces less communities than other label propagation approaches. We explain this fact by the use if the frequency matrix which stabilises the label propagations.

6 Conclusion and perspectives

We proposed a method to find overlapping communities from pre-computed disjoint communities obtained by using the *core detection label propagation* (CDLP) described in [7]. The algorithm selects candidates nodes for overlapping and uses *belonging functions* to decide the assignment or not of a candidate node to each of its neighbours communities. We proposed several belonging functions, all based on the topology of the communities. These belonging functions are either based on global measures which are the density and the clustering coefficient [8] (f_d and f_{cc}) or on average node measures which are the betweenness and the closeness centralities (f_b and f_{cl}). Belonging functions (f_d and f_{cc}) are computed on connected subgraphs, while the functions (f_b and f_{cl}) are based on the nodes.

The more the communities have strong densities, the higher the rate of overlap for a given candidate. For communities with high densities, f_d and f_{cc} perform better than f_b and f_{cl} in terms of quality (Modularity and NMI). The interest of using the functions f_b and f_{cl} lays in the case of graphs having communities with low densities, where some overlaps are difficult to detect.

We proposed and developed a new measure to compare the similarity on nodes that overlap several communities between two covers. This measure allowed us to observe experimentally on several benchmarks the strong similarity between f_d and f_{cc} and the differences between f_b and f_{cl}.

We compared our proposed methods with the more used overlapping community detection algorithms namely: CFinder [21], (the implementation of The clique percolation algorithm (CPM)) COPRA [15] ($\nu = 2$ and $\nu = 3$), ν being the number of communities to which nodes could belong, OSLOM [25], SLPA [65] and CONGA [39].

Experimental results showed that our proposed algorithms give relatively good and competitive results in term of quality but depend of the topology of the graph.

In future works, we propose to study the possibility to develop a parallel version of the different belonging functions knowing that a parallel and distributed version of the core label propagation using the Hadoop framework has been already developed [66].

References

1. Jain AK, Murty MN, Flynn PJ (1999) Data clustering: A review. ACM Comput Surv 31(3):264–323. [Online]. Available: https://doi.org/10.1145/331499.331504
2. Fortunato S (2010) Community detection in graphs. Phys Rep 486(3):75–174
3. Danon L, Duch J, Diaz-Guilera A, Arenas A (2005) Comparing community structure identification. [Online]. Available: https://doi.org/10.1088/1742-5468/2005/09/P09008
4. Fortunato S, Lancichinetti A (2009) Community detection algorithms: A comparative analysis: Invited presentation, extended abstract. In: Proceedings of the Fourth International ICST Conference on Performance Evaluation Methodologies and Tools, ser. VALUETOOLS '09. ICST, Brussels, Belgium, Belgium: ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), pp 27:1–27:2. [Online]. Available: http://dl.acm.org/citation.cfm?id=1698822.1698858
5. Yang J, Leskovec J (2012) Structure and overlaps of communities in networks. CoRR. arXiv: http://arxiv.org/abs/1205.6228
6. Lancichinetti A, Fortunato S, Kertész J (2009) Detecting the overlapping and hierarchical community structure in complex networks. New J Phys 11(3):033015
7. Attal J-P, Malek M (2015) A new label propagation with dams. In: 2015 IEEE/ACM International conference on advances in social networks analysis and mining (ASONAM). IEEE, pp 1292–1299
8. Attal J-P, Malek M, Zolghadri M (2016) Overlapping community detection using core label propagation and belonging function. In: International conference on neural information processing. Springer, pp 165–174
9. Freeman LC (1977) A set of measures of centrality based on betweenness. Sociometry 40:35–41
10. Bavelas A (1950) Communication patterns in task-oriented groups. J Acoust Soc Am 22(6):725–730
11. Kelley S (2009) The existence and discovery of overlapping communities in large-scale networks. Ph.D. dissertation, RENSSELAER POLYTECHNIC INSTITUTE
12. Lancichinetti A, Fortunato S, Kertész J (2009) Detecting the overlapping and hierarchical community structure in complex networks. New J Phys 11(3):033015
13. Lee C, Reid F, McDaid A, Hurley N (2010) Detecting highly overlapping community structure by greedy clique expansion. In: SNAKDD workshop, pp 4533–42
14. Reichardt J, Bornholdt S (2006) Statistical mechanics of community detection. Phys Rev E 74(1):016110
15. Gregory S (2010) Finding overlapping communities in networks by label propagation. New J Phys 12(10):103018
16. Wang J, Ren J, Li M, Wu F-X (2012) Identification of hierarchical and overlapping functional modules in ppi networks. IEEE Trans Nanobioscience 11(4):386–393

17. Sales-Pardo M, Guimera R, Moreira AA, Amaral LAN (2007) Extracting the hierarchical organization of complex systems. Proc Natl Acad Sci 104(39):15224–15229
18. Xie J, Kelley S, Szymanski BK (2013) Overlapping community detection in networks. The state-of-the-art and comparative study. ACM Comput Surv (CSUR) 45(4):43
19. Hajiabadi M, Zare H, Bobarshad H (2017) IEDC: an integrated approach for overlapping and non-overlapping community detection. Knowl-Based Syst 123:188–199. [Online]. Available: https://doi.org/10.1016/j.knosys.2017.02.018
20. Huang F, Li X, Zhang S, Zhang J, Chen J, Zhai Z (2017) Overlapping community detection for multimedia social networks. IEEE Trans Multimed 19(8):1881–1893. [Online]. Available: https://doi.org/10.1109/TMM.2017.2692650
21. Palla G, Derényi I, Farkas I, Vicsek T (2005) Uncovering the overlapping community structure of complex networks in nature and society. Nature 435(7043):814–818
22. Adamcsek B, Palla G, Farkas IJ, Derényi I, Vicsek T (2006) Cfinder: locating cliques and overlapping modules in biological networks. Bioinformatics 22(8):1021–1023
23. Shen H, Cheng X, Cai K, Hu M-B (2009) Detect overlapping and hierarchical community structure in networks. Phys A Stat Mech Appl 388(8):1706–1712
24. Ahn Y-Y, Bagrow J, Jø rgensen S (2010) Link communities reveal multiscale complexity in networks. Nature 466(7307):761–764
25. Lancichinetti A, Radicchi F, Ramasco JJ, Fortunato S (2011) Finding statistically significant communities in networks. PloS ONE 6(4):e18961
26. Baumes J, Goldberg M, Magdon-Ismail M, Merkle RC (2005) Efficient identification of overlapping communities. In: Kantor P, Muresan G, Roberts F, Zeng DD, Wang F-Y, Chen H (eds) Intelligence and security informatics. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 27–36
27. Whang JJ, Gleich DF, Dhillon IS (2016) Overlapping community detection using neighborhood-inflated seed expansion. IEEE Trans Knowl Data Eng 28(5):1272–1284. [Online]. Available: https://doi.org/10.1109/TKDE.2016.2518687
28. Ma X, Yang P, Guan S (2019) Overlapping community detection algorithm based on edge strength. IEEE Access 7:126642–126650. [Online]. Available: https://doi.org/10.1109/ACCESS.2019.2938783
29. Xu Y, Xu H, Zhang D, Zhang Y (2016) Finding overlapping community from social networks based on community forest model. Knowl-Based Syst 109:238–255. [Online]. Available: https://doi.org/10.1016/j.knosys.2016.07.007
30. Wang X, Liu G, Li J (2017) Overlapping community detection based on structural centrality in complex networks. IEEE Access 5:25258–25269. [Online]. Available: https://doi.org/10.1109/ACCESS.2017.2769484
31. Gregory S (2010) Fuzzy overlapping communities in networks, CoRR, vol. abs/1010.1523. [Online]. Available: arXiv:1010.1523
32. Zhang S, Wang R-S, Zhang X-S (2007) Identification of overlapping community structure in complex networks using fuzzy c-means clustering. Phys A Stat Mech Appl 374(1):483–490
33. Raghavan UN, Albert R, Kumara S (2007) Near linear time algorithm to detect community structures in large-scale networks. Phys Rev E 76(3):036106
34. Ren W, Yan G, Liao X, Xiao L (2009) Simple probabilistic algorithm for detecting community structure. Phys Rev E 79(3):036111
35. Wu Z-H, Lin Y-F, Gregory S, Wan H-Y, Tian S-F (2012) Balanced multi-label propagation for overlapping community detection in social networks. J Comput Sci Technol 27(3):468–479
36. Dai Q, Guo M, Liu Y, Liu X, Chen L (2013) Mlpa: Detecting overlapping communities by multi-label propagation approach. In: 2013 IEEE congress on evolutionary computation (CEC). IEEE, pp 681–688
37. Wu X, Zhang C (2015) Multi-label propagation for overlapping community detection based on connecting degree. In: Salah AA, Tonta Y, Salah AAA, Sugimoto CR, Al U (eds) Proceedings of the 15th international conference on scientometrics and informetrics, Istanbul, Turkey June 29 - July 3, 2015. ISSI Society
38. Nepusz T, Petróczi A, Négyessy L, Bazsó F (2008) Fuzzy communities and the concept of bridgeness in complex networks. Phys Rev E 77(1):016107
39. Gregory S (2007) An algorithm to find overlapping community structure in networks. In: Knowledge discovery in databases: PKDD 2007. Springer, pp 91–102
40. Girvan M, Newman ME (2002) Community structure in social and biological networks. Proc Natl Acad Sci 99(12):7821–7826
41. Rees BS, Gallagher KB (2010) Overlapping community detection by collective friendship group inference. In: 2010 International conference on advances in social networks analysis and mining (ASONAM). IEEE, pp 375–379
42. Kovács IA, Palotai R, Szalay MS, Csermely P (2010) Community landscapes: an integrative approach to determine overlapping network module hierarchy, identify key nodes and predict network dynamics. PloS ONE 5(9):e12528
43. Jin D, Gabrys B, Dang J (2015) Combined node and link partitions method for finding overlapping communities in complex networks. Sci Rep 5:8600
44. Wen X, Chen W-N, Lin Y, Gu T, Zhang H, Li Y, Yin Y, Zhang J (2017) A maximal clique based multiobjective evolutionary algorithm for overlapping community detection. IEEE Trans Evol Comput 21(3):363–377
45. Zhang L, Pan H, Su Y, Zhang X, Niu Y (2017) A mixed representation-based multiobjective evolutionary algorithm for overlapping community detection. IEEE Trans Cybern 47(9):2703–2716
46. Cheng J, Wu X, Zhou M, Gao S, Huang Z, Liu C (2019) A novel method for detecting new overlapping community in complex evolving networks. IEEE Trans Syst Man Cybern Syst 49(9):1832–1844. [Online]. Available: https://doi.org/10.1109/TSMC.2017.2779138
47. Tran TN, Wehrens R, Buydens LMC (2006) Knn-kernel density-based clustering for high-dimensional multivariate data. Comput Stat Data Anal 51(2):513–525. [Online]. Available: https://doi.org/10.1016/j.csda.2005.10.001
48. Newman ME, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69(2):026113
49. Kannan R, Vempala S, Vetta A (2004) On clusterings: Good, bad and spectral. J ACM (JACM) 51(3):497–515
50. Seifi M, Junier I, Rouquier J-B, Iskrov S, Guillaume J-L (2013) Stable community cores in complex networks. In: Complex networks. Springer, pp 87–98
51. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech Theory Exp 2008(10):P10008
52. Barabási A-L, Albert R (1999) Emergence of scaling in random networks. Science 286(5439):509–512
53. Newman ME (2003) The structure and function of complex networks. SIAM Rev 45(2):167–256
54. Clauset A, Shalizi CR, Newman ME (2009) Power-law distributions in empirical data. SIAM Rev 51(4):661–703
55. Watts D, Strogatz S (1998) Collective dynamics of small-world networks. Nature 393:440–442

56. Freeman LC (1978) Centrality in social networks conceptual clarification. Soc Networks 1(3):215–239
57. Nicosia V, Mangioni G, Carchiolo V, Malgeri M (2009) Extending the definition of modularity to directed graphs with overlapping communities. J Stat Mech Theory Exp 2009(03):P03024
58. Ana L, Jain AK (2003) Robust data clustering. In: Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on, vol. 2. IEEE, pp II–128
59. Hubert L, Arabie P (1985) Comparing partitions. J Classif 2(1):193–218
60. Zachary W (1977) An information flow model for conflict and fission in small groups. J Anthropol Res 33:452–473
61. Girvan M, Newman MEJ (2002) Community structure in social and biological networks. Proc Natl Acad Sci 99(12):7821–7826
62. Krebs V (2004) Books about us politics. unpublished, http://www.orgnet.com
63. Lusseau D, Schneider K, Boisseau OJ, Haase P, Slooten E, Dawson SM (2003) The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. Behav Ecol Sociobiol 54(4):396–405
64. Newman ME (2006) Finding community structure in networks using the eigenvectors of matrices. Phys Rev E 74(3):036104
65. Xie J, Szymanski BK, Liu X (2011) Slpa: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process. In: 2011 IEEE 11th International Conference on Data mining workshops (ICDMW). IEEE, pp 344–349
66. Attal J-P, Malek M, Zolghadri M (2019) Parallel and distributed core label propagation with graph coloring. Concurr Computat Pract Exp 31(2):e4355