Journal of Universal Computer Science, vol. 18, no. 4 (2012), 457-486
submitted: 30/9/11, accepted: 14/12/11, appeared: 28/2/12

Fuzziness and Overlapping Communities in Large-Scale Networks

Qinna Wang and Eric Fleury
(LIP ENS-LYON, D-NET INRIA
Université de Lyon
46 Allée d'Italie Lyon 69364 France
qinna.wang@ens-lyon.fr eric.fleury@inria.fr)

Abstract: Overlapping community detection is a popular topic in complex networks. As compared to disjoint community structure, overlapping community structure is more suitable to describe networks at a macroscopic level. Overlaps shared by communities play an important role in combining different communities. In this paper, two methods are proposed to detect overlapping community structure. One is called clique optimization, and the other is named fuzzy detection. Clique optimization aims at detecting granular overlaps. The clique optimization method is a fine grain scale approach. Each granular overlap is a node connected to distinct communities and it is highly connected to each community. Fuzzy detection is at a coarser grain scale and aims at identifying modular overlaps. Modular overlaps represent groups of nodes that have high community membership degrees with several communities. A modular overlap is itself a possible cluster/sub-community. Experimental studies in synthetic networks and real networks show that both methods provide good performances in detecting overlapping nodes but in different views. In addition, a new extension of modularity is introduced for measuring the quality of overlapping community structure.
Key Words: fuzzy community detection, overlapping community detection, community detection, modularity, large-scale networks
Category: L.3, L.6

1 Introduction

The empirical information of network datasets can be used to study several characteristics of networks, like small-world property, heavy-tailed degree distributions [Albert et al. 2007] and rumor spreading. These characteristics are closely related to the property of *community structure*. In the study of complex networks, a network is said to have *community structure* if the nodes of the network can be easily grouped into sets of nodes such that each set of nodes is densely connected internally, between which connections are sparse.

Modularity optimization is a popular approach to detect *partitions* of networks. A *partition* is the division of a network into disjoint communities, where each node belongs to one and only one community. Due to inhomogeneity of link distribution in real networks, the *modularity* [Girvan and Newman 2002], which compares differences between the number of links within communities and the expected number of links in the null model, can be used to measure the

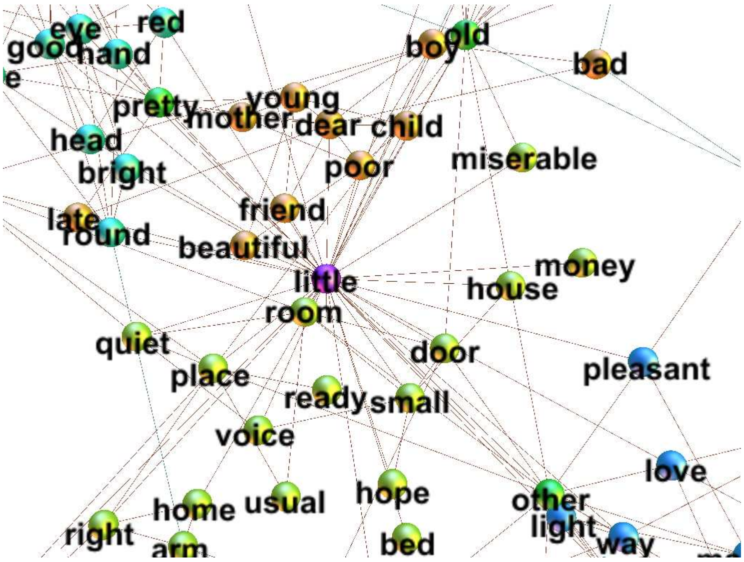


Figure 1: An example of overlapping communities in an adjacency network of common adjectives and nouns in the novel David Copperfield by *Charles Dickens* [Newman 2006]. The result is obtained by clique optimization [see Section 4], which shows communities through colors: orange denotes community "HUMAN", composed of "boy", "child", "friend", *etc.* , green represents community "OBJECT", comprised by "door", "house", "room", *etc.* , and light blue denotes community "HEAD", consisting of "hand", "head", and "eye", *etc.* . We observe an overlapping node "little", which can be used to describe these subjects.

quality of partitions. A good partition usually has the high modularity. Thus, modularity optimization is applied to a lot of datasets for capturing structural properties [Guimerà and Amaral 2005].

However, the modularity fails to measure the quality of *covers*. A *cover* is the division of a network into communities which are allowed to share common nodes. *Overlaps* are the common nodes shared by at least two communities, which play an important role in combining communities. Figure 1 shows an example: the overlapping node "little" can be used to describe humans, like "boy", "child", "friend", and can be used to describe objects, like "door", "house", "room", too.

Several algorithms for overlapping community detection are known in the literature. These methods include CPM [Palla et al. 2005], fitness-based algorithm [Lancichinetti et al. 2009] and OSLOM [Lancichinetti et al. 2010b]. These algorithms aim at detecting *local communities* without respect to the graph as a whole. The definition of *fuzzy community structure* is also used to detect overlapping communities, which considers a community as a group of nodes having high probability together with each other. For example, Reichardt et al. [Reichardt 2004] introduced the energy landscape survey method, and Sales Pardo et al. [Sales-Pardo et al. 2007] proposed the modularity-landscape survey method to construct a hierarchical tree. Both them detect fuzzy community structure by computing the probability that a pair of nodes belong to the same community.

In the following, we will propose a new extension of modularity for measuring the quality of overlapping community structure, which is derived from the Hamiltonian [Reichardt and Bornholdt 2006]. In addition, we introduce two complement methods to detect covers. We obtain overlapping community structure by adding these overlapping nodes to their related communities. Our first method is called clique optimization. Clique optimization aims at detecting *granular overlaps*. The clique optimization method is a fine grain scale approach. Each granular overlap is a node connected to distinct communities and it is highly connected to each community. Roughly speaking, a granular overlap is shared by several distinct communities while being intrinsically a member of each of them. The second method is named fuzzy detection. Fuzzy detection is at a coarser grain scale and aims at identifying *modular overlaps*. *Modular overlaps* represent groups of nodes that have high community membership degrees with several communities. A modular overlap is itself a possible cluster/subcommunity. As opposed to granular overlaps, modular overlaps imply the hierarchical organization of the graph: modular overlaps are sub-communities shared by several communities. The obtained results of the two methods are different. Since the two methods offer a different granularity scale (fine and coarse), they are complementary and meaningful in characterizing overlapping nodes.

This paper is organized as follows. Section 2 introduces the current work in cover detection. In [Section 3], we describe our new extension of modularity. In [Section 4] and [Section 5], we present our methods and show their performances by applying them to synthetic networks, respectively. We also compare their performances in analysing real networks in [Section 6]. Finally, we conclude our current work and the prospect for the future, in [Section 7].

2 Related work

2.1 Definition and notation

A complex network is modelled by a graph (network) which is used to describe the topology structure of a complex system. The nodes of the graph are individuals connected by edges which mimic their interactions.

Let us start with a graph $\mathcal{G} = (V, E)$ comprising $n = |V|$ nodes (or vertices) connected by $m = |E|$ links (or edges). The number of elements in V and E are denoted by n and m, respectively.

In the context of graph theory, an adjacency (or connectivity) matrix $\mathbf{A}$ is often used to describe a graph $\mathcal{G}$. Given a $n \times n$ matrix $\mathbf{A} = [A_{ij}]_{n \times n}$, its element A_{ij} is equal to 1 when the link e_{ij} exists, and zero otherwise.

A group of nodes having denser internal connections than external connections is called a *community.* Given a community $\mathcal{C}$ of a graph $\mathcal{G}$, we define the internal and external degree of node $v \in \mathcal{C}$, k_v^{int} and k_v^{ext}, as the number of edges connecting v to other nodes belonging to $\mathcal{C}$ or to the rest of the graph, respectively. If $k_v^{\text{ext}} = 0$, the node v has only neighbors within $\mathcal{C}$: assigning v to the current community $\mathcal{C}$ is likely to be a good choice. If $k_v^{\text{int}} = 0$ instead, the node is disjoint from $\mathcal{C}$ and it should better be assigned to a different community. Classically, we note $k_v = k_v^{\text{int}} + k_v^{\text{ext}}$ the degree of node v. The internal degree k^{int} of $\mathcal{C}$ is the sum of the internal degrees of its nodes. Likewise, the external degree k^{ext} of $\mathcal{C}$ is the sum of the external degrees of its nodes. The total degree $k_{\mathcal{C}}$ is the sum of the degrees of the nodes of $\mathcal{C}$. By definition: $k_{\mathcal{C}} = k_{\mathcal{C}}^{\text{int}} + k_{\mathcal{C}}^{\text{ext}}$.

A *partition* $\mathcal{P} = \{\mathcal{C}_1, \ldots, \mathcal{C}_k\}$ is a division of a graph into disjoint communities. For every pair of communities $\mathcal{C}_i$ and $\mathcal{C}_j$ in a partition $\mathcal{P}$, they have $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset$. A *cover* $\mathcal{S} = \{S_1, \ldots, S_k\}$ denotes a division of a graph into communities sharing nodes. Given a cover $\mathcal{S}$, someone may find that a pair of communities S_i and S_j share overlapping nodes such as $S_i \cap S_j \neq \emptyset$.

The partitions can be measured by the *quality function*, which assigns a score to the partition of a graph. In this way, we can rank partitions based on their score given by the quality function. Partitions with high scores are "good", so the partition with the largest score is by definition the best.

The widest accepted quality function is the modularity of Newman and Girvan [Newman 2004], which is defined as:

$$Q = \frac{1}{2m} \sum_{i \neq j} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(\sigma_i, \sigma_j) \tag{1}$$

where σ_i is the community to which node i belongs and $\delta(\sigma_i, \sigma_j)$ is the Kronecker delta symbol, which is equal to 1 if the pair of nodes i and j belong to the same community; otherwise it is equal to 0. The modularity is always smaller than one, and can be negative as well. For instance, the partition where each node represents a single community, is always negative. When taking the whole graph as a single community, the modularity is zero as the two terms in this case are equal.

2.2 Current work

We then present a class of algorithms for network clustering, which allow nodes belonging to more than one community.

The CPM (clique percolation method) [Palla et al. 2005] is one of early work for cover detection, which detects *k-clique communities*. A *k-clique community* is a series of *adjacent k-clique*. Two k-cliques are *adjacent* if and only they share $k-1$ nodes. However, this definition is too strict. It fails to resolve non-trivial communities, like WikiTalk [Lancichinetti et al. 2010a] which is a sparse network consisting of star-like communities.

Baumes et al. [Baumes et al. 2005] proposed a density metric for clustering nodes. In their method, nodes are added into clusters if and only if their fusion improves the cluster density. Under this condition, the results really depend on seeds for network clustering. The seed can be a random node or a disjoint community. As shown in their results, there is a huge difference in the number of communities based on different types of seeds.

Lancichinetti et al. has made many efforts in cover detection including fitness-based function [Lancichinetti et al. 2009] and OSLOM (Order Statistics Local Optimization Method) [Lancichinetti et al. 2010b]. The former is based on the local optimization of a k-fitness function, whose result is limited by the tunable parameter k, and the later uses the statistical significance [Lancichinetti and Radicchi 2009] of clusters with an expansive computational cost as it sweeps all nodes for each "worst" node. For the optimization, Lancichinetti et al. [Lancichinetti et al. 2010b] propose to detect significant communities based on a partition. They detect a community by adding nodes, between which the togetherness is high. This is one of popular techniques for overlapping community detection. There are similar endeavours like greedy clique expansion technique [Lee et al. 2010] and community strength-based overlapping community detection [Wang et al. 2009]. However, as they applied Lancichinetti et al. [Lancichinetti et al. 2009]'s k-fitness function, the results are limited by the tunable parameter k.

Some cover detection approaches are based on other basis. For example, Reichardt et al. [Reichardt and Bornholdt 2006] introduced the energy landscape survey method, and Sales Pardo et al. [Sales-Pardo et al. 2007] proposed the modularity-landscape survey method to construct a hierarchical tree. They aim at detecting fuzzy community structure, whose communities consist of nodes having high probability together with each other. As indicated in [Sales-Pardo et al. 2007], they are limited by scales of networks.

Evans et al. [Evans and Lambiotte 2009] proposed to construct a *line graph* (A *line graph* is constructed by using nodes to represent edges of the original graphs.) which transforms the problem of node clustering to the link clustering and allows nodes shared by several communities. The main drawback is that, in their results, overlapping communities always exist.

The problem of overlapping community detection remains.

3 A new extension of modularity

Modularity has been employed by a large number of community detection methods. However, it only evaluates the quality of partitions. Here, we introduce its extension for covers, which is combined with the energy model Hamiltonian for the spin system [Reichardt and Bornholdt 2006].

Let a community structure be represented by a spin configuration $\{\sigma\}$. Each spin state represents a community, and the number of spin states represents the number of communities of the graph. Thus, the quality of a community structure can be represented through the energy of spin glass. In [Reichardt and Bornholdt 2006], a simplified *Hamiltonian* is proposed to measure the quality of community structure, which is written in:

$$\mathcal{H}(\{\sigma\}) = -\sum_{i \neq j} (A_{ij} - \gamma p_{ij})\, \delta(\sigma_i, \sigma_j) \ , \tag{2}$$

where $(A_{ij} - \gamma p_{ij})$ represents a coupling between nodes i and j, σ_i, σ_j denote the spin states of nodes i and j, respectively. The Kronecker delta symbol $\delta(\sigma_i, \sigma_j)$ yields 1 if and only if $\sigma_i = \sigma_j$ and 0 otherwise.

Therefore, we can rewrite the modularity Q Eq. 1 as:

$$Q = -\frac{1}{m}\mathcal{H}(\{\sigma\}) \ , \tag{3}$$

with $\gamma = 1$ and $p_{ij} = \frac{k_i k_j}{2m}$.

Since a good quality function of community structure should reward the internal links and penalize the external links, the Hamiltonian Eq. 2 can be expressed in two ways. One describes the cohesion within the community, and the other shows the adhesion among different communities:

$$\mathcal{H}(\{\sigma\}) = -\sum_{s} (m_{ss} - \gamma [m_{ss}]_{p_{ij}}) = -\sum_{s} c_s \ , \tag{4}$$

and

$$\mathcal{H}(\{\sigma\}) = \sum_{s<r} (m_{sr} - \gamma [m_{sr}]_{p_{ij}}) = \sum_{s} a_{sr} \ . \tag{5}$$

For each community $\mathcal{C}_s$, m_{ss} represents the number of links within $\mathcal{C}_s$, m_{sr} represents the number of links between $\mathcal{C}_s$ and $\mathcal{C}_r$, $[m_s]_{p_{ij}}$ and $[m_{sr}]_{p_{ij}}$ are expected number of links with the link distribution p_{ij}, c_s denotes the cohesion of $\mathcal{C}_s$ and a_{sr} represents the adhesion between $\mathcal{C}_s$ and $\mathcal{C}_r$.

We can assume diverse expressions of $[\cdot]_{p_{ij}}$, which is an expectation under the link distribution p_{ij}. In case of [Fig. 2] for disjoint clusters n_1 and n_2, the choice should satisfy the following:

1. when n_s is a cluster belonging to the rest of the graph, $[m_{1s}]_{p_{ij}} + [m_{2s}]_{p_{ij}} = [m_{1+2,s}]_{p_{ij}}$;

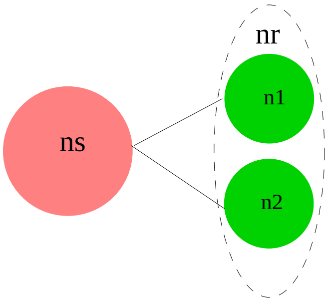
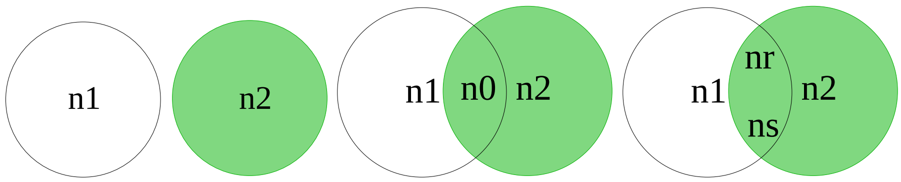


Figure 2: Example of $[\cdot]_{p_{ij}}$, where the union of clusters n_1 and n_2 is n_r such that $n_1 \cup n_2 = n_r$ and the cluster n_s belongs to the rest of the graph.



(a) $n_1 \cap n_2 = \emptyset$ (b) $n_{01} \cap n_{02} = n_0$ (c) $n_{rs1} \cap n_{rs2} = n_r \cup n_s$

Figure 3: Let us denote the union of the clusters n_0 and n_1 by n_{01}. Similarly, we denote the union of the clusters n_0 and n_2 by n_{02}, the union of the clusters n_r and n_s by n_{rs}, the union of the clusters n_1, n_r and n_s by n_{rs1} and the union of the clusters n_2, n_r and n_s by n_{rs2}. Three different subdivisions of the community n_3: (a) two disjoint sub-communities n_1, n_2 ; (b) two overlapping sub-communities n_{01}, n_{02} sharing a cluster n_0; and (c) two overlapping sub-communities n_{rs1}, n_{rs2} sharing two clusters n_r , n_s, where n_r, n_s are disjoint sub-communities of n_0 such as $n_r \cap n_s = \emptyset$ and $n_r \cup n_s = n_0$.

2. when n_r is an union cluster composed of n_1 and n_2, $[m_{rr}]_{p_{ij}} = [m_{11}]_{p_{ij}} + [m_{22}]_{p_{ij}} + [m_{12}]_{p_{ij}}$.

We show three different subdivisions of one community n_3 in [Fig. 3]. In the first subdivision [see Fig. 3(a)], community n_3 consists of n_1 and n_2 with empty intersection such as $n_1 \cup n_2 = n_3, n_1 \cap n_2 = \emptyset$. From Eq. 4 and Eq. 5, we can easily prove

$$c_3 = c_1 + c_2 + a_{12} \ , \tag{6}$$

where c_3 denotes the cohesion of n_3 that is the union of n_1 and n_2 with empty intersection, a_{12} denotes the adhesion between n_1 and n_2, c_1 and c_2 are the cohesions of sub-communities n_1 and n_2 respectively.

In the second subdivision [see Fig. 3(b)], there is an overlapping cluster n_0

between n_{01} and n_{02}. We write the cohesions for sub-communities n_{01} and n_{02} as:

$$\begin{cases} c^0_{01} = c^0_0 + c_1 + a^0_{01} \\ c^0_{02} = c^0_0 + c_2 + a^0_{02} \ , \end{cases}$$

where c^0_{01} and c^0_{02} denote the cohesions of the sub-communities n_{01} and n_{02} respectively, a^0_{01} and a^0_{02} denote the adhesions between n_0 and n_1, n_2. Here, n_0 is shared by n_{01} and n_{02}. In terms of the adhesion, we have

$$a^0_{01,02} = a^0_{01} + a^0_{02} + a_{12}$$

between n_{01} and n_{02}.

For the union of $n_3 = n_{01} \cup n_{02}$, we obtain

$$\begin{aligned} c_3 &= c_0 + c_1 + c_2 + a_{01} + a_{02} + a_{12} \\ &= 2c^0_0 + c_1 + c_2 + 2a^0_{01} + 2a^0_{02} + a_{12} \ . \end{aligned}$$

So we derive

$$c^0_0 = \frac{1}{2}c_0 \ , a^0_{01} = \frac{1}{2}a_{01} \text{ and } a^0_{02} = \frac{1}{2}a_{02} \ . \tag{7}$$

In the third subdivision [see Fig. 3(c)] such as $n_r \cup n_s = n_0$, we replace c_0 and c^0_0 by

$$\begin{cases} c_0 = c_r + c_s + a_{rs} \\ c^0_0 = c^r_r + c^s_s + a^{rs}_{rs} \ , \end{cases} \tag{8}$$

where c^r_r and c^s_s denote the cohesions of overlapping sub-communities n_r and n_s respectively. a^{rs}_{rs} denotes the adhesion between overlapping sub-communities n_r and n_s, which satisfies $a^{rs}_{rs} = \frac{1}{2}a_{rs}$ due to Eq. 7.

Therefore, we propose the contribution of a_{rs} for all communities $\{\mathcal{C}_1, \ldots, \mathcal{C}_k\}$ written in:

$$\sum_1^k \frac{1}{|d_r \cup d_s|} a_{rs} = \frac{|d_r \cap d_s|}{|d_r \cup d_s|} a_{rs} \ , \tag{9}$$

where d_r and d_s denote the community memberships of n_r and n_s, respectively.

For the relation between the Hamiltonian and the modularity Eq. 3, we write the quality of overlapping community structure in form of modularity:

$$Q_{ov} = \frac{1}{2m} \sum_{i \neq j} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \frac{|d_i \cap d_j|}{|d_i \cup d_j|} \ , \tag{10}$$

where d_i and d_j are memberships of nodes i and j, respectively. Then, for a pair of nodes i and j always belonging to the same community such as $d_i \cap d_j = d_i \cup d_j$, their contribution to the modularity is $\left(A_{ij} - \frac{k_i k_j}{2m} \right)$; for a pair of nodes i and j never belonging to the same community such as $d_i \cap d_j = \emptyset$, their contribution is 0; otherwise, their contribution is in range of $\left(0, \left(A_{ij} - \frac{k_i k_j}{2m} \right) \right)$. Furthermore, if the found community structure is a partition, its quality Q_{ov} is equal to the modularity Q Eq. 1.

This extension of modularity is able to measure the quality of overlapping community structure. However, we can not detect covers by optimizing it. Therefore, we propose two methods based on other basis. One is called clique optimization for detecting granular overlaps, and the other is named fuzzy detection aiming at identifying modular overlaps. Although granular overlaps and modular overlaps are used to denote overlapping nodes shared by several communities, they are different. Granular overlaps represent nodes that have high togetherness with distinct communities while modular overlaps denote sub-communities shared by several communities. Therefore, given a pair of communities, we may observe several modular overlaps shared by them, while there is only one group of granular overlapping nodes.

4 Clique optimization

The definition of community is not standard. The most commonly used one for overlapping community detection is that communities are clique-like objects. Given a clique, each member has connections with all other members. They are supposed to share common interests. The applications which detect clique-like communities like CPM [Palla et al. 2005], SCP [Kumpula et al. 2008] on social networks have good performance. Based on these observations, we propose to detect covers based on cliques.

4.1 Definition of granular overlaps

Given a partition, we often observe that cliques are cut by disjoint communities. For example, given a pair of communities $\mathcal{C}_i$ and $\mathcal{C}_j$, they may cut a clique K such as $(K \cap \mathcal{C}_i) \cup (K \cap \mathcal{C}_j) = K$, where $K \cap \mathcal{C}_i \neq \emptyset$ and $K \cap \mathcal{C}_j \neq \emptyset$. In our mind, a clique is an exclusive group of people who share common interests, views, purposes, patterns of behavior, *etc.* . Therefore, we define that a node is a possible *granular overlapping node* if it is involved into a clique cut by a partition.

CPM [Palla et al. 2005] is one popular method for cover detection. It is designed to uncover the community structure composed of *k-clique-communities.* A *k-clique-community* is the union of all k-cliques that can be reached from each other through a series of *adjacent k-cliques.* Two k-cliques are said to be *adjacent* if they share $k-1$ nodes. Extended from the definition of k-clique community, we define that a clique and a community are *adjacent* if they share $k-1$ nodes. We also define that, if a clique is k-adjacent to a disjoint community, all members of this clique can be assigned to this community. If a node can be assigned into more than one community, it is a *granular overlapping node.*

In the following, we give the definition of granular overlapping nodes in two senses:

Definition 1. A node v is a k-granular overlapping node shared by l communities $\mathcal{E} = \{\mathcal{C}_1, \ldots, \mathcal{C}_l\}$ in a strong sense if it belongs to a clique K adjacent to these communities, such as: $\forall \mathcal{C}_i \in \mathcal{E}$, $|K \cap \mathcal{C}_i| \geq k - 1$.

Definition 2. A node v is a k-granular overlapping node shared by l communities $\mathcal{E} = \{\mathcal{C}_1, \ldots, \mathcal{C}_l\}$ in a weak sense if it is involved in l' cliques $\mathcal{K} = \{K_1, \ldots, K_{l'}\}$ which are adjacent to them such as: $\forall \mathcal{C}_i \in \mathcal{E}$, $\exists K_j \in \mathcal{K}, |K_j \cap \mathcal{C}_i| \geq k - 1$.

Clearly an overlapping node in a strong sense is also an overlapping node in a weak sense, whereas the converse is not true.

4.2 Our algorithm of clique optimization

Algorithm 1 Clique optimization

Input: $\mathcal{G} = (V, E)$, k

Output: $\mathcal{S} = \{S_1, \ldots, S_{n_c}\}$ an overlapping community covering of V

Obtain a partition $\mathcal{P} = \{\mathcal{C}_1, \ldots, \mathcal{C}_{n_c}\}$ by running an efficient partition detection algorithm on the graph $\mathcal{G}$.
$\mathcal{S} \leftarrow \mathcal{P}$
// STEP 1: Find cliques which are $k-$adjacent to communities
for all Edges connecting one granular overlapping node candidate **do**
  Find a clique K_j, which is $k-$adjacent to at least one community
  Find all communities $\mathcal{E}_j = \{\mathcal{C}_1, \ldots, \mathcal{C}_\ell\}$ k-adjacent to K_j: $\forall \mathcal{C}_i \in \mathcal{E}_j$, $|K_j \cap \mathcal{C}_i| \geq k - 1$
  // STEP 2: Update overlapping communities
  for all k-adjacent communities $\mathcal{C}_i \in \mathcal{E}_j$ **do**
    Merge K_j to $\mathcal{C}_i$:$S_i \leftarrow \mathcal{S}_i \cup K_j$
  end for
end for
Return $\mathcal{S}$

Our clique optimization is proposed to detect k-granular overlapping nodes for cover detection. This algorithm consists of two phases: based on a partition, the first phase is to detect cliques which are k-adjacent to communities; the second phase is merging the above detected cliques into communities. The algorithm is sketched in Algo. 1. We describe it in details below.

After obtaining a partition by running an efficient partition detection algorithm (such as the Louvain algorithm) on the graph (line 1), we start our first phase. We define a node to be a *granular overlapping node candidate* if its external degree is at least $k-1$. In the first phase (line 3 – 9), we detect all cliques which are k-adjacent to communities. A simple resolution is based on edges connecting one granular overlapping node candidate to detect a clique which is k-adjacent to at least one community. Chosen a granular overlapping node candidate, we find a $k-1$ clique whose members belong to $\mathcal{N}$ ($\mathcal{N}$ is initialized by the neighbourhood of the chosen granular overlapping node candidate). Then, we obtain a clique k-adjacent to one community by adding the overlapping node candidate to the found $k-1$ clique.

Next, we merge this clique to communities in the second phase (line 6 – 8). For each clique which shares sets of $k-1$ nodes with one community, we merge them. Finally, we obtain a cover where granular overlapping nodes are shared by overlapping communities.

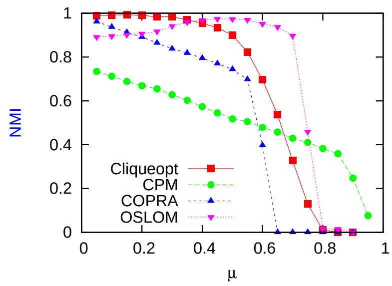
The worst-case complexity of clique optimization is in $\mathcal{O}(n^k k^2)$: there are $\mathcal{O}(n^k)$ subgraphs to check, each of which has $\mathcal{O}(k^2)$ edges, where n represents the number of nodes whose external degree is at least 1. Note that n is the size of the community given by the partition algorithm and one may expect that n is smaller than the total number of nodes in the graph. Our method is faster than CPM [Palla et al. 2005] or SCP [Kumpula et al. 2008], since it only detects cliques separated by community boundaries.

From the definitions given above, our clique optimization is defined for undirected and unweighted graphs. When analyzing an arbitrary system, one could decide that the directionality of the links could be ignored if it makes sense. If $u \to v$ means that the entity u is in interaction with the entity v, we may want to infer that $v \to u$ remains valid, yielding $u \leftrightarrow v$.

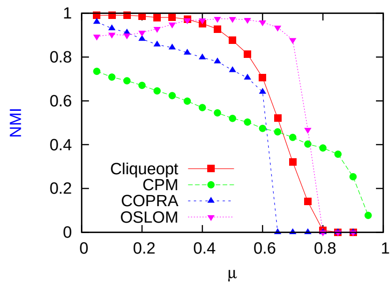
If connections are weighted, a threshold weight ω^* is used to prune weak links and keep those that are stronger than ω^*. Depending on the weight distribution, the threshold could be $\omega^* = \frac{1}{2m}\sum_{v=1}^{n} k_v$, where k_v is the weighted degree of node v. If we want to keep all links, ω^* is simply set to zero. If the threshold weight is increased, the number of edges is decreased and so is the number of overlapping nodes. Note that, if ω^* is increased, the granular overlapping nodes should have stronger links to their related communities.

4.3 Benchmark graphs

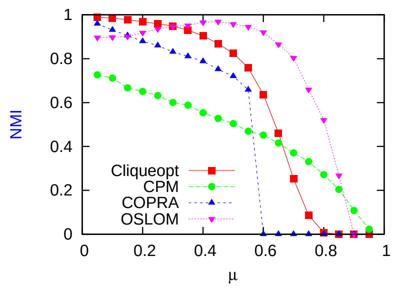
It is now possible to show performances of clique optimization. We have considered a set of artificial networks with the known community structure. We show their accuracy through the *normalized mutual information* (NMI) [Lancichinetti et al. 2009] by comparing to ground truth. The higher value of the variation of information is, the more similar two covers are. If two covers are identical, NMI is 1. The results obtained by our clique optimization on the following benchmark graphs are good and presented bellow.



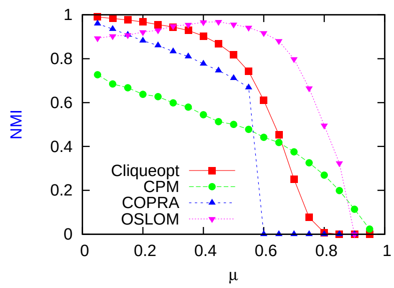
(a) N=20000,on=50



(b) N=20000,on=100



(c) N=10000,on=50



(d) N=10000,on=100

Figure 4: Tests of our clique optimization on computer generated networks with known community structure and comparison with CPM [Palla et al. 2005], CORPA [Gregory 2010] and OSLOM [Gregory 2010]. Here, x-axis denotes the varying mixing parameter μ and y-axis represents the average NMI of 50 samples by comparing the found community structure and the ground truth. Besides the number of nodes N, the number of overlapping nodes on and the tunable parameter μ, the other parameters are identical: average degree $k = 20$, maximum degree $maxk = 300$, minus exponent for the degree sequence $t1 = 2$, minus exponent for the community size distribution $t2 = 1$, minimum for community sizes $minc = 10$, maximum for community $maxc = 300$, and number of memberships of overlapping nodes $om = 2$.

In [Fig. 4], we present the comparison of NMI for clique optimization with the other cover detection algorithms including CPM [Palla et al. 2005], COPRA [Gregory 2010] and OSLOM [Lancichinetti et al. 2010b] through their applications to LFR benchmarks [Lancichinetti and Radicchi 2009]. LFR benchmarks are constructed corresponding to a series of parameters, including the number of nodes, average degree, the maximum degree, the number of overlapping nodes, the number of overlapping community memberships, and the mixing parameter. The *mixing parameter* μ is the ratio of external degree to the node degree. For each overlapping node i shared by ν_i communities, if it belongs to community ξ, its adjacent links to ξ satisfies: $k_i^{\xi} = k_i^{in}/\nu_i$. As we can see, clique optimization performs well such as $NMI \geq 0.9$ when $\mu < 0.5$ in [Fig. 4(a) and Fig. 4(b)]. It also has good performance in [Fig. 4(c) and Fig. 4(d)] when $\mu \leq 0.3$ and has lower NMI than OSLOM when $\mu > 0.3$. It can be understood since OSLOM detects *significant communities.* A *significant community* is a group of nodes having a larger density of internal connections than of external links. If a node can not improve any community's significance (the difference between the internal connection density and external connection density), it is defined to be an individual node and is not considered in the community structure.

5 Fuzzy detection

In this section, we will introduce another method for cover detection. It is named fuzzy detection, which is proposed for identifying modular overlaps. Modular overlaps are groups of nodes shared by communities. Different from granular overlaps, modular overlaps are related to the hierarchy organization. That is, modular overlaps are sub-communities shared by several communities.

5.1 Motivation

Our fuzzy detection is based on the Louvain algorithm [Blondel et al. 2008]. The Louvain algorithm is a partition detection algorithm and provides good partitions with high modularity. It consists of two phases that are iteratively repeated until no more positive gain of modularity. Initially, all nodes are assigned into a single community. Then, for each node whose move improves the modularity, will be removed from its current community to the neighbour community which yields the largest positive increase of modularity. The first phase repeatedly and sequentially sweeps all nodes until no further improvement of modularity can be gained. The second phase is building a new graph based on communities found in the first phase. Once the second phase is completed, the first phase is reapplied to the new network. The two phases are iteratively applied until no more change in community structure or maximum modularity is achieved. In the following, we use iteration to denote the combination of these two phases. The partition found by this algorithm is hierarchical organized, whose height of hierarchy is determined by the number of iterations. The Louvain algorithm is extremely fast and provides partitions having high modularity.

When running several times the Louvain algorithm on the same given network, we observe from a run to another that nodes may be grouped together with different community members in distinct partitions. Since the Louvain algorithm sweeps nodes in a non deterministic fashion (a random permutation of V), it naturally introduces instability which may be a weakness. It turns out that we can take benefit of this instability. By detecting nodes that jump from one community to another between distinct runs, we are in fact able to uncover nodes that have high community memberships with distinct communities. Such "*oscillating*" nodes can be considered as overlapping nodes. Therefore, we propose a fuzzy detection algorithm which detects groups of nodes having strong connection probability with several communities.

Algorithm 2 Louvain algorithm.

Input: $G = (V, E)$, l^* a level threshold
Output: $\mathcal{P}$ a partition
$l \leftarrow 0; G_0 \leftarrow G$
repeat
  $l \leftarrow l + 1$
  Initialize a partition $\mathcal{P}_l$ of $G_l(V_l, E_l)$
  // First phase: partition update
  repeat
    Nodes in a random permutation
    for all Nodes: $v \in V_l$ **do**
      Move from σ_v to one selected $\sigma_{v'}$ (v' is a neighbour of v)
    end for
  until no more change increases modularity
  // Second phase: Construct a new meta graph
  Replace each community by a node
  Replace connections between a pair of communities by one weighted edge
until $\mathcal{P}_l$ is not updated or $l = l^*$.
Return $\mathcal{P}$ corresponding to the roots of the hierarchical tree.

5.2 Our algorithm of fuzzy detection

To have the benefit of the potential Louvain algorithm instability [Aynaud 2011], we force the algorithm to use a random seed at each run. The random seed makes the nodes be swept in a random permutation during the modularity optimization. Thus, different runs may produces different partitions. By repeating Louvain algorithm, we are able to compute, a co-appearance matrix $\mathbf{P} = [p_{ij}]_{n\times n}$. For each pair of nodes (i, j), p_{ij} of $\mathbf{P}$ represents the probability for the pair nodes i and j to appear in the same community. Having $p_{ij} = 1$ implies that nodes i and

Algorithm 3 Fuzzy detection.

Input: $G = (V, E)$, α^*, β^*
Output: $\mathcal{S}$ an overlapping community covering of V
  // STEP 1: Detect robust clusters
$\mathbf{P}^0 \leftarrow 0$; $k \leftarrow 0$;$\text{modularity}_{\max} \leftarrow -\infty$
repeat
  $k \leftarrow k + 1$
  $\mathcal{P} \leftarrow$ Run the Louvain algorithm on G
  Update $\mathbf{P}^k$
  if modularity of $\mathcal{P}$ greater than $\text{modularity}_{\max}$ **then**
    Save the partition $\mathcal{P}$ in $\mathcal{P}_{\text{opt}}$ and update $\text{modularity}_{\max}$
  end if
until $\|\mathbf{P}^k - \mathbf{P}^{k-1}\| \leq \epsilon$
$\mathcal{P}_{\text{sc}} = \mathcal{P}_{\text{opt}}$
for all edge $e = (i, j)$ such that $p_{ij} < \alpha^*$ **do**
  Remove the external edge e from $\mathcal{P}_{\text{sc}}$
end for
  // STEP 2: Adjust the membership of robust clusters
Input: $G = (V, E)$, $\mathcal{P}_{\text{sc}}$, $\mathcal{S} \leftarrow \mathcal{P}_{opt}$
for all $\mathcal{C}_i \in \mathcal{P}_{opt}$ **do**
  Identify community core: $\widehat{c}_i = \arg\max_{c_j \subseteq \mathcal{C}_i} |c_j|$
end for
Compute $\mathbf{P}_{c_i, c_j}$
for all $c_j \in \mathcal{P}_{\text{sc}}$ and $c_j \notin \{\widehat{c}_1, \ldots, \}$ **do**
  if $p_{c_j, \widehat{c}_i} \geq \beta^*$ **then**
    $S_i \leftarrow S_i \cup c_j$
  end if
end for
Return $\mathcal{S}$

j are always in the same community while edges $e = (i, j)$ having a p_{ij} close to 0 implies that edge e connects two different communities. The underlying idea of fuzzy detection approach is thus to detect overlapping communities from a classical partition approach.

Detecting overlapping nodes also allows to detect more stable nodes that always belong together in the same community. In this algorithm, we use the notion of *community cores* to denote communities. Given a community, its *core* is a group of nodes offering high stability against random perturbation. To detect community cores, we're going to remove edges in order to keep only core nodes. First we remove all *external edges*, *i.e.* , all edges $e = (i, j)$, having a connection

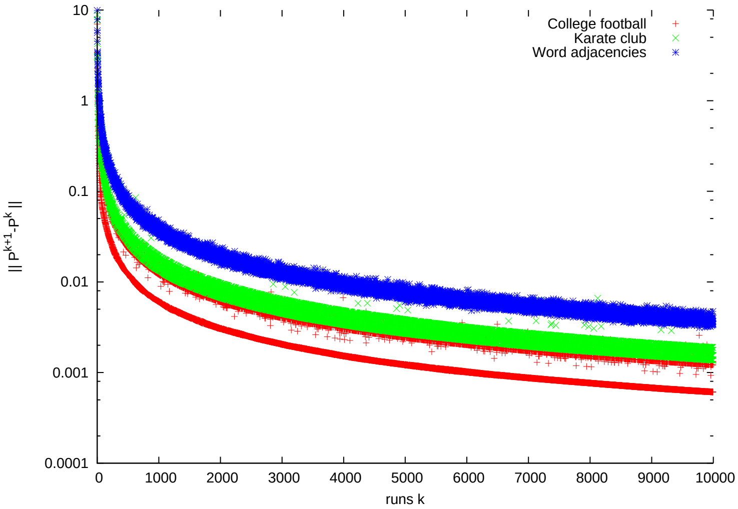


Figure 5: As the number of runs increases, the shape of the function value Eq. 11 gets closer and closer to 0. The figure shows results on College football [Girvan and Newman 2002], Karate club [Zachary 1977] and Word adjacencies [Newman 2006].

probability p_{ij} less than a threshold α^*. After this pruning phase, a set of disjoint robust cluster is obtained. A *robust cluster* is a group of nodes connected by edges having in-cluster probability larger than or equal to α^*. Note that a given community may have several robust clusters. We choose the community core corresponding to the robust cluster having the maximum size. The notion of external edges was used in [Gfeller et al. 2005] where authors add a random noise over the weight of the edges of the network (equally distributed between $[-\sigma, \sigma]$). Once community cores are identified, we continue iteratively, following the Louvain approach. Similarly, in our method, we replace the robust clusters by supernodes and connect them through the connection between robust clusters. In this case, the weight of the edge between the supernodes is the sum of the weights of the edges between the identified robust clusters. We run again the Louvain algorithm to compute the probability of robust clusters and community cores to appear in the same community. Finally, we add each robust cluster to the community if they have a high community membership degree such as their probability of appearing in the same community is high.

The global algorithm is shown in Algo. 3. First, (lines $2-9$) we compute the co-appearance matrix $\mathbf{P} = [p_{ij}]_{n\times n}$ by running the Louvain algorithm of Algo. 2 several times with a random seed. The number of runs is determined by the

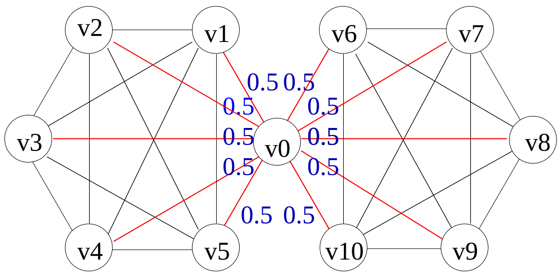


Figure 6: Illustration of our fuzzy detection on a toy graph which consists of two overlapping cliques. After removing all edges in low probability $p_{ij} = 50\%$ shown in red, robust clusters are obtained, concluding $\{v_1, v_2, v_3, v_4, v_5\}$, $\{v_6, v_7, v_8, v_9, v_{10}\}$, and a single v_0.

convergence criteria (line 9):

$$\|\mathbf{P}^{k+1} - \mathbf{P}^k\| = \sqrt{\frac{1}{m} \sum_{(i,j)\in E} (p_{ij}^{k+1} - p_{ij}^k)^2} < \varepsilon, \tag{11}$$

where $\mathbf{P}^k$ represents the result after k-th run and p_{ij}^k denotes the statistical probability of nodes i and j to belong to the same community after k-th runs (line 5) and ε is a small threshold. Figure 5 illustrates the convergence of the norm when running fuzzy detection algorithm. We observe that $\|\mathbf{P}^{k+1} - \mathbf{P}^k\|$ decreases as the number k of runs increases.

Then, we detect robust clusters $\{c_1, c_2, \ldots, c_s\} = \mathcal{P}_{\text{sc}}$ (lines $10 - 13$). Given a partition $\mathcal{P}_{opt}$ which has the maximum modularity among all computed partitions obtained during the first phase, the robust clusters are detected by removing all edges having a probability p_{ij} lower that a given threshold α^* (typically $\alpha^* = 0.9$). A simple illustration is given in [Fig. 6].

Finally in the second phase, we identify modular overlaps which have high community memberships with several communities. Given a community $\mathcal{C}_i \in \mathcal{P}_{\text{opt}}$, its core $\widehat{c}_i$ is the robust cluster $c_j \subseteq \mathcal{C}_i$ having the maximum size, such as:

$$\widehat{c}_i = \arg\max_{c_j \subseteq \mathcal{C}_i} |c_j| \tag{12}$$

We assign each robust cluster c_j to the community $\mathcal{C}_i$ if and only if their community membership $p_{c_j,\widehat{c}_i}$ is larger than a threshold β^* such as $p_{c_j,\widehat{c}_i} > \beta^*$ (typically $\beta^* = 0.1$). If one robust cluster is assigned to at least two communities, we call it a *modular overlap*. Given a modular overlap, its members are possible granular overlapping nodes. Only the granular overlapping nodes are required to have dense connection with related communities. The nodes shared by the same

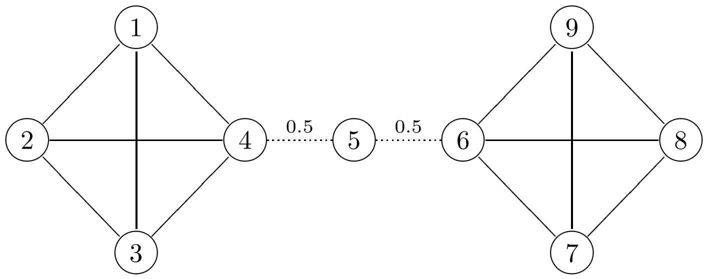


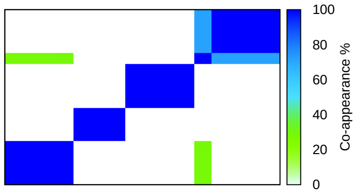
Figure 7: An example graph that contains a unstable node 5. Node 5 has a relatively high membership degrees with two communities ($p = 0.5$). However, it is connected to each community with only 1 link.

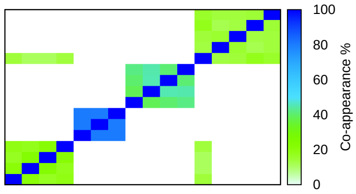
modular overlaps are not only required to have dense connection with related communities and also are required to have high internal modular degree (the number of links connected to other members within the robust cluster).

In cases where a community consists of several robuster clusters of comparable size, one may tune and increase the value of α^* in order to refine the core identification.

Since fuzzy detection is used to identify modular overlaps, which are subcommunities shared by several communities, we restrict the modular overlaps to have a size greater than 3. We can now introduce the notion of *unstable nodes*, which are nodes connecting communities with few links but are observed to have high co-appearance probability with several communities. Figure 7 illustrates such case. Due to unstable nodes, we do not use fuzzy detection to identify granular overlaps. Moreover, we may observe some modular overlaps that are not real overlapping nodes but are more like *unstable clusters*.

The running time of fuzzy detection mainly depends on the co-appearance matrix calculation. The complexity to find a partition by the Louvain algorithm is estimated by authors in [Blondel et al. 2008] to be in $\mathcal{O}(m)$, where m is the number of edges in the network (the worst complexity is much higher, but in practice, on real network, Louvain algorithm performs very well). Thus the computational complexity of fuzzy detection is in $\mathcal{O}(Km)$, where K is the number of runs of Louvain algorithm needed before reaching an acceptable convergence of $\mathbf{P}$. Once more, in practice, we take benefit of the efficient Louvain algorithm running time and our fuzzy detection is fast. We experiment storage limitation due to the matrices $\mathbf{P}^k$ and $\mathbf{P}^{k+1}$ more that time computing one.





(a) $l^* = 1$ (b) without l^*

Figure 8: The co-appearance matrix of artificial networks containing hierarchical structure. The color corresponds to the probability of nodes in the same community: the deep color represents the high probability; the color is white if the probability is 0%.

5.3 Benchmark graphs

In the following, we show performance of fuzzy detection in testing benchmark graphs with the known community structure in *hierarchical organization*.

A community structure can be hierarchically ordered when the graph offers several levels of organization/structure at different scales. In this case, the community structure is *hierarchically constructed* by small communities at each level, all nested within large communities at higher levels. As an example, one may consider in a social network the granularity of the living place (town), the working place (school) and refine it toward the graduate or class level.

We apply fuzzy detection to an artificial graph containing hierarchical structure [Lancichinetti et al. 2009] and a modular overlap. The benchmark graph consists of 512 nodes, which belong to 16 groups, arranged into 4 supergroups and one group is shared by two supergroups. Every node has an average of $k_1 = 30$ links with the nodes in the same micro-community, $k_2 = 13$ links with the nodes in the same macro-community but different micro-community. In addition, each node has $k_3 = 5$ links with the rest of the networks. As the modular overlap has macro-links with two communities, its nodes have the total degree $k = 61$ while the other nodes only have the total degree $k = 48$.

Figure 8(b) illustrates the co-appearance matrix by running the Louvain algorithm without fixing the level threshold l^* [Algo. 2], while Figure 8(a) provides the result by running the Louvain algorithm with $l^* = 1$. In both figures, the nodes are sorted in the same order corresponding to the robust clusters and the selected partition $\mathcal{P}_{\text{opt}}$. As the distinction among robust clusters is not clear in [Fig. 8(b)], we use [Fig. 8(a)] for the visualization, where we observe 4 communities, 32 robust clusters and one modular overlap. It shows the good performance of fuzzy detection in detecting modular overlaps.

Remark that, when running our fuzzy detection to identify modular overlaps, the community core is not a single robust cluster. As each community has four large robust clusters with comparable size. By increasing the value of α^*, we obtain a reasonable community core whose size is larger than the others within the same community.

6 Applications in real networks

6.1 Yeast protein complexes

As a further test, we consider the application to yeast protein complexes. The combined-AP/MS network [1] describes 9070 interactions among 1622 proteins. With a catalogue of protein complexes provided by CYC2008 [Pu et al. 2009], results are shown in [Tab. 1].

Method	NMI	Sensitivity	Specificity	Accuracy	Modularity Eq. 10
Clique Optimization	0.824323	0.514852	0.874587	0.6947195	0.772569
Fuzzy detection	0.702184	0.970297	0.290757	0.630527	0.866759
CPM	0.699512	0.287129	0.801471	0.5442995	0.816893
OSLOM [Lancichinetti et al. 2010b]	0.52039	0.257426	0.965677	0.6115515	0.662716
Copra [Gregory 2010]	0.517806	0.118812	0.967657	0.5432345	0.888672

Table 1: Results of different overlapping community detections on Yeast protein complexes, in views of NMI, sensitivity, specificity, accuracy and modularity.

We see that clique optimization identifies protein complexes with a high degree of success. By comparing to other overlapping detection techniques, it provides the highest NMI [Lancichinetti et al. 2009]. NMI measures the similarity between the results and the ground truth based on information theory. We also provide sensitivity, specificity, accuracy and modularity. *Sensitivity* relates to the ability to identify the real overlapping nodes, which is the proportion of real overlapping nodes among the found overlapping nodes. *Specificity* relates to the ability of identify non-overlapping nodes, which is the proportion of non-overlapping nodes among all found non-overlapping nodes. The *accuracy* is a "balanced accuracy", which is the sum of sensitivity and specificity with the equal importance. We use the accuracy to show the goodness in detecting overlapping nodes. We observe that clique optimization has the highest accuracy, too.

Compared to other methods, the advantage of our fuzzy detection in identifying granular overlaps, is not obvious. Its has the lower NMI and accuracy value

[1] Available at `http://interactome.dfci.harvard.edu/S_cerevisiae/`

than the clique optimization. Moreover, the low sensitivity of clique optimization is caused by our definition of k-granular overlapping nodes, *i.e.* , not all real overlapping nodes participate in k-cliques. In contrast, fuzzy detection provides results with a high sensitivity. Since fuzzy detection assigns nodes into communities without computing their connections. Simultaneously, clique optimization will not misclassify unstable nodes. Therefore, it has a higher specificity value than fuzzy detection. It suggests us to combine both methods to study overlapping community structure. We may obtain the complementary results.

6.2 Complex System Science

Next, we consider the applications of clique optimization and fuzzy detection to a real network called Complex System Science. It is a co-citation network, whose dataset is composed of articles extracted from the ISI Web of knowledge. Article were published between 2000 and 2009. The network is composed of 141 163 nodes and 19 603 888 links. The nodes correspond to articles containing a set of keywords relevant to the field of complex systems. The weight of the links between articles is calculated through their common references (bibliographic coupling [Kessler 1963]). A link exists between two articles if they share references, meaning that they cite common work which may implies that they are dealing with a same scientific object/domain. More precisely, given two articles (nodes) i and j, each one having a set of references R_i (respectively R_j), there exists a link $e = (i, j)$ between i and j if i and j share at least one reference and the weight is measured by: $w_{ij} = \dfrac{|R_i \cap R_j|}{\sqrt{|R_i|\,|R_j|}}$.

In [Fig. 9], we find 12 communities in scale above 100. These communities can be identified by research topics or theoretical fields through studies in topic keywords, see [Tab. 2]. We compute the frequency of topic keywords by aggregating the number of units (articles). For instance, if only one unite contains the topic keywords "Neurons", the corresponding frequency is 1. In the figure, the light green community is identified by neuroscience: biology psychology. This community contains high frequent keywords (NEURONS, PERFORMANCE, CENTRAL-NERVOUS-SYSTEM) very general in neuroscience while some high frequent keywords (BRAIN, LONG-TERM POTENTIATION, DISEASE) seem to emphasize the study in the field of biological psychology. To our knowledge, biological psychology or behavioral neuroscience is the study of the biological substrates of behavior and mental processes. Physiological psychologists use animal models, typically rats, to study the neural, genetic, and cellular mechanisms that underlie specific behaviors such as learning and memory and fear responses. Cognitive neuroscientists investigate the neural correlates of psychological processes in humans using neural imaging tools, and neuropsychologists conduct psychological assessments to determine, for instance, specific aspects and extent of cognitive

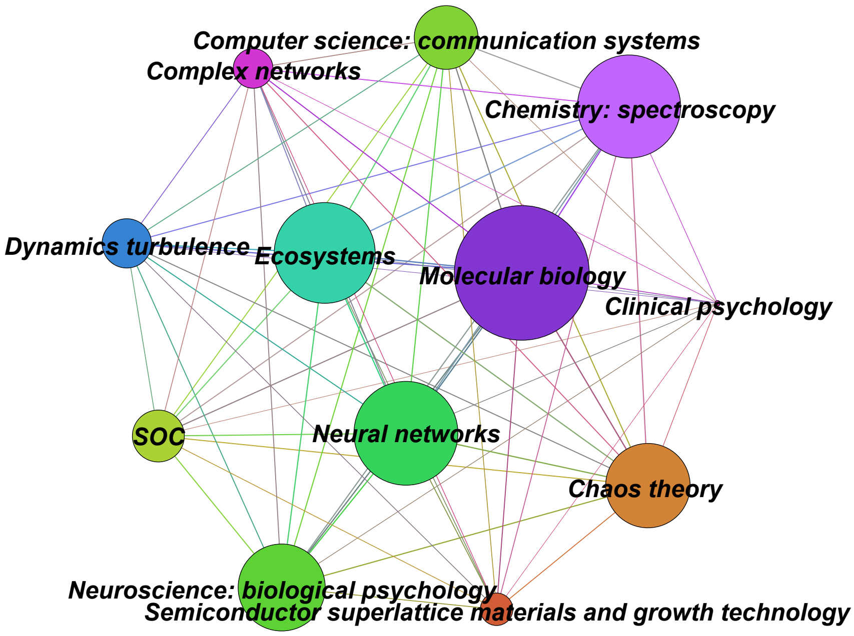


Figure 9: The community structure of Complex System Science, in which communities are identified by research topics or theoretical fields.

deficit caused by brain damage or disease.

Table 4 shows results of clique optimization in identifying granular overlaps. We see the applications of chaos theory in different disciplines including complex networks, nervous systems and ecosystems. We also observe the intermediation: visual cortex between neural networks and neuroscience: biological psychology. Visual cortex is one part of the visual systems, which receives visual information for processing images. These results are interesting in understanding the combination of different disciplines and applications..

In view of robust clusters [Fig. 10], these robust clusters can be considered as sub-specialities of the identified disciplines [Tab. 5]. For example, the community identified by neuroscience: biology psychology is composed of several clusters, which are also characterized by research topics or theoretical areas. Note that, the study in neuroplasticity supports the treatments of brain damage, long-term potentiation concerns learning and memory, pre-botzinger complex is essential for respiratory rhythm, and the activities in prefrontal cortex are considered to be orchestration of thoughts and actions in accordance with internal goals. All these topics and fields refer to the study in neuroscience and biological psychology. It reveals that fuzzy detection can extract communities in hierarchical organization.

Community	Highest Frequent Topic Keywords	High Frequent Topic Keywords
Neuroscience: Biological Psychology	Brain	Brain, Neurons, Long-Term Potentiation, Association, Expression, Performance, Disease, Model, Synaptic Plasticity, Activation, Complex, Children, Central-Nervous-System, Rat
Chaos Theory	Chaos	Chaos, Dynamics, Systems, Model, Stability, Complexity, Synchronization, Time-Series, Bifurcation, Self-Organization
Chemistry: Spectroscopy	Complexes	Complexes, Self-Organization, Crystal-Structure, Chemistry, Derivatives, Behavior, Films, Polymers, Systems, Phase-Transition, Spectroscopy, Dynamics, Thin-Films, Molecules, Nonlinear-Optical Properties
Complex Networks	Complex	Complex Networks, Dynamics, Small-World Networks, Model, Internet, Evolution, Systems, Organization, Topology, Scale-Free Networks, Metabolic Networks, Web, Graphs
Ecosystems	Ecology	Ecology, Systems, Model, Complexity, Evolution, Dynamics, Management, Growth, Behavior, Self-Organization, Patterns, Simulation, Biodiversity, Models
Molecular Biology	Expression	Expression, Complex, Gene-Expression, Protein, In-Vivo, Activation, Saccharomyces-Cerevisiae, Identification, Gene, Escherichia-Coli, Cells, In-Vitro, Binding, Crystal-Structure, Messenger-Rna, Phosphorylation, Proteins
Semiconductor Superlattice Materials And Growth Technology	Growth	Growth, Gaas, Islands, Molecular-Beam Epitaxy, Self-Organization, Quantum Dots, Surfaces, Films, Photoluminescence, Silicon, Nanostructures, Si(001)
Clinical Psychology	Management	Management, Therapy, Trauma, Experience, Hemorrhage, Surgery, Inhibitors, Optimization, Recombinant Factor Viia, Damage Control, Mortality, Cancer
Neural Networks	Neural Networs	Neural Networks, Model, Systems, Classification, Optimization, Algorithm, Identification, Design, Prediction, Self-Organizing Maps
Soc	Self-Organized Criticality	Self-Organized Criticality, Model, Dynamics, Econophysics, Evolution, Systems, Fluctuations, Behavior, Growth, Turbulence, Noise, Transport, Avalanches, Earthquakes, Patterns, Time-Series
Computer Science: Communication Systems	Systems	Systems, Design, Performance, Channels, Algorithm, Networks, Capacity, Ofdm, Stability, Optimization, Fading Channels, Algorithms, Model, Signals, Codes, Transmission
Dynamics Turbulence	Turbulence	Turbulence, Model, Flow, Simulation, Dynamics, Behavior, Large-Eddy Simulation, Complex Terrain, Plasticity, Flows, Boundary-Layer

Table 2: Results of communities in the partition. The shown high frequent topic keywords are sorted in descending order and each topic keyword is contained in at least 20 articles.

In terms of modular overlaps, our results are shown in [Tab. 3]. Except astronomy-ISM(Interstellar medium) which acts like a unstable cluster, the rest has a good agreement compared to the reality: discrete-event systems and multi-agents are very common for modelling and analysing general systems, computational complexity is a common property of complex systems, and genetic expression [Hugot et al. 2001, Limbergen et al. 2007] studies are often used to de-

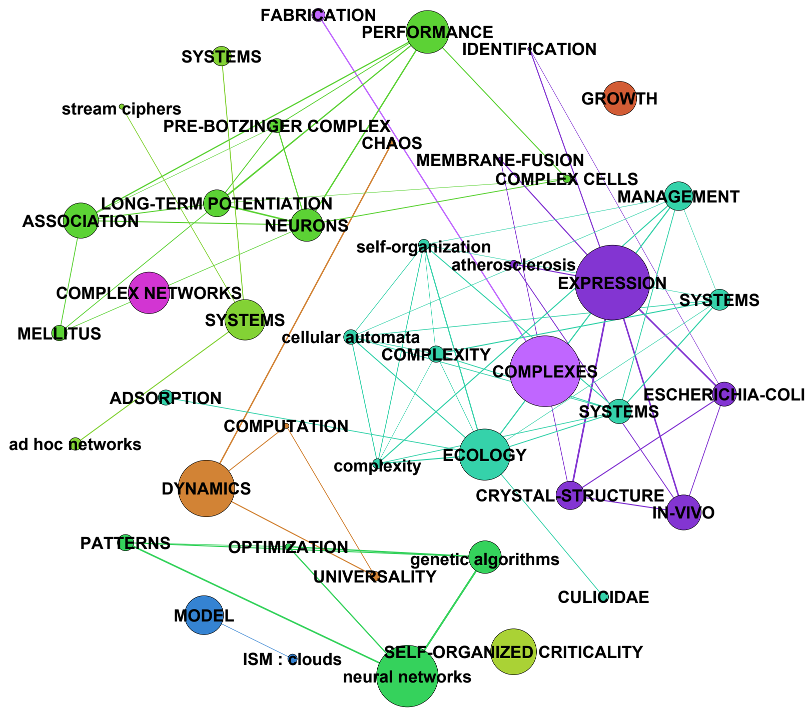


Figure 10: Results of fuzzy detection on Complex System Science. Robust clusters are marked by the highest frequent topic keywords. Their colours correspond to the relevant communities as shown in [Fig. 9].

termine whether a genetic variant is associated with a disease or trait.

Comparing the results of granular overlaps and modular overlaps, we see their difference. For instance, fuzzy detection considers three modular overlaps related to computer science: communication systems and ecosystems simultaneously, while clique optimization does not provide any result. We also observe their similarity. For example, both results use visual cortex to characterize the overlapping nodes shared by neural networks and neuroscience: biological psychology. It indicates that, for some cases, the two types of overlapping nodes can reach an agreement in characterizing overlaps.

Obviously, we can not compare the goodness between granular overlaps and modular overlaps in a definitive and quantitative way as they represent results based on different definitions. To the best of our knowledge, both definitions seem reasonable to use. Finally, we conclude that both methods: clique optimization and fuzzy detection, are useful to identify overlaps in complex networks.

Modular Overlaps	High Frequent Topic Keywords	Involving Communities
Genetic Association	**Association**, Susceptibility, Polymorphism, Linkage Disequilibrium, Disease, Major Histocompatibility Complex, Linkage, Complex Traits, Risk, Population	Molecular Biology, Neuroscience: Biological Psychology
Discrete-event Systems	**Systems**, Supervisory Control, Petri Nets, Complexity, Discrete-Event Systems, Verification, Design, Automata, Synchronization, Discrete Event Systems	Computer Science: Communication Systems, Ecosystems
Computational Complexity	**Complexity**, Algorithms, Computational Complexity, Algorithm, Networks, Optimization, Time, Systems, Search, Computational-Complexity	Computer Science: Communication Systems, Ecosystems
Astronomy-ISM (Interstellar Medium)	**Turbulence**, Ism: Clouds, Star-Formation, Stars: Formation, Molecular Clouds, Ism: Structure, Ism: Kinematics And Dynamics, Evolution, Radio Lines: Ism, Intergalactic Medium	Dynamics Turbulence, Clinical Psychology
Multi-Agent Systems	**Systems**, Multi-Agent Systems, Multiagent Systems, Design, Agents, Architecture, Multi-Agent System, Framework, Model, Intelligent Agents	Computer Science: Communication Systems, Ecosystems
Visual Cortex	**Complex Cells**, Lateral Geniculate-Nucleus, Cat Striate Cortex, Primary Visual-Cortex, Striate Cortex, Cortical-Neurons, Receptive-Fields, Contrast, Orientation Selectivity, Simple Cells	Neuroscience: Biological Psychology, Neural Networks

Table 3: Results of fuzzy detection: ten high frequent topic keywords contained by modular overlaps between pairs of communities. These high frequent topic keywords are contained in at least 20 articles and are shown in order of descending frequency. The highest frequent topic keywords are shown in bold font.

7 Conclusion and future work

In this paper, we propose a new extension of modularity for measuring the quality of overlapping community structure. And two different methods are introduced to identify overlapping nodes. One is called clique optimization for identifying granular overlaps, and the other is named fuzzy detection for detecting modular overlaps. Both methods have been tested successfully in synthetic graphs. Moreover, studies and analysis on large networks like the Complex System Science one give good results and useful insights on the structure of the network.

We believe that the elements presented in this paper can be of great help in the analysis of networks. On the one hand, the definition of granular overlaps and modular overlaps provide different insights in characterizing overlapping nodes for network analysis. On the other hand, the introduction of clique optimization and fuzzy detection could open the way for applications to large-scale systems. Several researches remain. We are currently studying underlying network organizations in both static and dynamic viewpoints. We are investigating the evolution of communities to mine more structural properties of complex networks.

	Complex Networks	Neural Networks	Semiconductor Superlattice Materials And Growth Technology	Ecosystems	Soc
Molecular Biology		**Saccharomyces-Cerevisiae**, Identification, Yeast, Complex Networks, Gene-Expression, Patterns, Database, Cell-Cycle, Organization, Self-Organizing Maps			**Complex Networks**, Organization, Dynamics, Model, Evolution, Metabolic Networks, Systems, Topology, Small-World Networks, Escherichia-Coli
Chaos Theory	**Synchronization**, Systems, Dynamics, Model, Complex Networks, Chaos, Self-Organized Criticality, Stability, Complexity, Econophysics	**Chaos**, Systems, Dynamics, Model, Complexity, Neural Networks, Time-Series, Synchronization, Stability, Networks		**Dynamics**, Chaos, Model, Systems, Complexity, Self-Organization, Stability, Evolution, Cellular Automata, Patterns	**Dynamics**, Systems, Self-Organized Criticality, Chaos, Time-Series, Complexity, Model, Complex Networks, Econophysics, Synchronization
Neuroscience: Biological Psychology	**Complex Networks**	**Neurons**, Model, Primary Visual-Cortex, Complex Cells, Visual-Cortex, Epistasis, Receptive-fields, Multifactor-Dimensionality Reduction, Cortex, Association			
Chemistry: Spectroscopy		**Dynamics**	**Self-Organization**, Superlattices, Nanoparticles, Clusters, Quantum Dots, Nanocrystals, Particles, Total-Energy Calculations, Self-Organized Growth		

Table 4: Results of clique optimization: ten high frequent topic keywords contained by granular overlaps between pairs of communities. The shown high frequent topic keywords are sorted in descending order and each topic keyword is contained in at least 20 articles. The highest frequent topic keywords are shown in bold font.

Community	Cluster	High Frequent Topic Keywords
Dynamics Turbulence	Flow Over Complex Terrain	**Turbulence**, Model, Flow, Simulation, Complex Terrain, Large-Eddy Simulation, Flows, Behavior, Boundary-Layer, Plasticity
	Astronomy-Ism (Interstellar Medium)	**Turbulence**, Ism : Clouds, Star-Formation, Stars : Formation, Ism : Structure, Molecular Clouds, Ism : Kinematics And Dynamics, Evolution, Radio Lines : Ism, Intergalactic Medium
Computer Science: Communication Systems	Telecommunication System	**Systems**, Performance, Channels, Synchronization, Fading Channels, Capacity, Ofdm, Equalization, Networks, Multiuser Detection
	Control Theory	**Systems**, Stability, Design, Robust Control, Optimization, Linear-Systems, Model-Predictive Control, Stabilization, H-Infinity Control, Model Predictive Control
	Wireless Network	**Ad Hoc Networks**, Sensor Networks, Wireless Sensor Networks, Self-Organization, Networks, Wireless Networks, Clustering
	Cryptography	**Stream Ciphers**, Cryptanalysis, Linear Complexity, Stream Cipher, Sequences
Molecular Biology	Expression	**Expression**, Complex, Gene-Expression, Protein, Saccharomyces-Cerevisiae, Gene, Activation, In-Vivo, Identification, In-Vitro
	Dendritic Cells	**Dendritic Cells**, In-Vivo, Expression, T-Cells, Infection, Complex, Mice, Activation, Major Histocompatibility Complex, Antigen
	Crystal structure Of Escherichia Coli	**Crystal-Structure** , Complex, Escherichia-Coli, Binding, Protein, Recognition, Mechanism, Proteins, Molecular-Dynamics, Complexes
	Gene Expression In Escherichia Coli	**Escherichia-Coli**, Gene-Expression, Systems, Expression, Model, Networks, Systems Biology, Protein, Transcription, Rhythms
	Atherosclerosis	**Atherosclerosis**, Inflammation, Expression, Disease, Myocardial-Infarction, In-Vivo, C-Reactive Protein, Smooth-Muscle-Cells, Activation, Low-Density-Lipoprotein
	Membrane Fusion And Exocytosis	**Membrane-Fusion**, Neurotransmitter Release, Exocytosis, Syntaxin, Snare, Complex, Protein, Snare Complex, Transmitter Release
	Proteomics	**Identification**, Proteomics, Mass-Spectrometry, Proteins, Peptides, Protein Identification
Chaos Theory	Chaotic Dynamics	**Chaos**, Dynamics, Systems, Complexity, Stability, Model, Time-Series, Synchronization, Nonlinear Dynamics, Bifurcation
	Quantum Chaos And Universality	**Universality**, Quantum Chaos, Systems, Chaos, States, Model, Random-Matrix Theory, Complex Systems, Fluctuations, Spectra
	Chaos In Population dynamics	**Chaos**, Stability, Dynamics, Population, Permanence, Models, Systems, Bifurcation, Predator-Prey System, Birth Pulses
	Neuroplasticity	**RAT**, Neurons, Plasticity, Hippocampus, Brain, Central-Nervous-System, Synaptic Plasticity, Long-Term

		Potentiation, Food-Intake, Memory
Neuroscience: Biological Psychology	Long-Term Potentiation	**Long-Term Potentiation**, Synaptic Plasticity, Plasticity, Hippocampus, Nmda Receptor, Glutamate Receptors, Expression, Neurons, In-Vivo, Hippocampal-Neurons
	Genetic Association	**Association**, Susceptibility, Polymorphism, Linkage Disequilibrium, Disease, Major Histocompatibility Complex, Linkage, Complex Traits, Risk, Population
	Pre-Botzinger Complex	**Pre-Botzinger Complex**, In-Vitro, Prebotzinger Complex, Brain-Stem, Respiratory Rhythm Generation, Rhythm Generation, Rat, Control Of Breathing, Neurons, Pacemaker Neurons
	Prefrontal Cortex	**Performance**, Attention, Fmri, Children, Prefrontal Cortex, Brain, Working-Memory, Cortex, Memory, Activation
	Diabetes Mellitus	**Mellitus**, Glycemic Control, Complications, Hypertension, Randomized Controlled-Trial, Diabetes, Therapy, Risk, Diabetes Mellitus, Management
Chemistry: Spectroscopy	Crystal Structure	**Complexes**, Self-Organization, Crystal-Structure, Derivatives, Chemistry, Polymers, Behavior, Films, Nonlinear-Optical Properties, Phase-Transition
	Anodic Alumina	**Fabrication**, Arrays, Films, Anodic Alumina, Anodization, Self-Organization, Growth, Self-Organized Formation, Hexagonal Pore Arrays, Titanium
Soc	Soc	**Self-Organized Criticality**, Model, Dynamics, Econophysics, Evolution, Systems, Fluctuations, Models, Behavior, Turbulence
Ecosystems	Innovation Management	**Management** , Innovation, Economics, Performance, Model, Complexity, Systems, Technology, Firm, Knowledge
	Discrete-Event Systems	**Systems**, Supervisory Control, Petri Nets, Complexity, Discrete-Event Systems, Verification, Design, Automata, Discrete Event Systems, Synchronization
	Computational Complexity	**Complexity**, Algorithms, Computational Complexity, Algorithm, Networks, Optimization, Time, Systems, Search, Computational-Complexity
	Ecosystems	**Ecology**, Dynamics, Evolution, Biodiversity, Patterns, Diversity, Growth, Model, Management, Conservation
	Absorption	**Adsorption**, Sorption, Speciation, Complexation, Humic Substances, Water, Natural-Waters, Kinetics, Ph, Copper
	Cellular Automaton	**Cellular Automata**, Systems, Simulation, Self-Organization, Model, Cellular-Automata, Flow, Cellular-Automaton Model, Traffic Flow, Dynamics
	Multi-agent Systems	**Systems**, Multi-Agent Systems, Multiagent Systems, Design, Agents, Architecture, Multi-Agent System, Framework, Model, Intelligent Agents
	Division Of Labor In Insect Societies	**Self-Organization**, Behavior, Division-Of-Labor, Hymenoptera, Ants, Colonies, Formicidae, Social Insects, Swarm Intelligence, Evolution
	Complex Adaptive Systems	**Complexity**, Self-Organization, Chaos, Emergence, Science, Complex Adaptive Systems, Complexity Theory

	Malaria	**Malaria**, Culicidae, Identification, Transmission, Complex, Diptera, Africa, Mosquitos, Anopheles-Gambiae Complex, Gambiae Complex
Neural Networks	Neural Networks	**Neural Networks**, Classification, Systems, Model, Self-Organizing Map, Neural Network, Algorithm, Identification, Artificial Neural Networks, Prediction
	Genetic Algorithm	**Optimization**, Genetic Algorithms, Genetic Algorithm, Design, Systems, Neural Networks, Model, Algorithm, Algorithms, Simulation
	Simulated Annealing	**Optimization** , Simulated Annealing, Algorithm, Model
	Gene Expression Patterns	**Patterns**, Self-Organizing Maps, Gene-Expression, Microarray, Identification, Gene Expression, Saccharomyces-Cerevisiae, Cancer, Expression, Classification
Complex Systems	Complex Systems	**Complex Networks**, Dynamics, Small-World Networks, Model, Internet, Networks, Evolution, Scale-Free Networks, Systems, Organization

Table 5: Results of fuzzy detection: ten high frequent topic keywords contained by robust clusters. These high frequent topic keywords are contained in at least 20 articles and are shown in order of descending frequency. The highest frequent topic keywords are shown in bold font.

References

[Albert et al. 2007] Albert, R., Barabasi, A. L., Asur, S., Parthasarathy, S., and Ucar, D. An event-based framework for characterizing the evolutionary behavior of interaction graphs Reviews of Modern Physics, 74 (2007), 1, 913–9211060.

[Aynaud 2011] Aynaud, T. *Détection de communautés dans les réseaux dynamiques.* PhD thesis, DOCTEUR DE L'UNIVERSITÉ PIERRE ET MARIE CURIE(2011).

[Baumes et al. 2005] Baumes, J., Goldberg, M., and Magdon-Ismail, M. Efficient identification of overlapping communities. Intelligence and Security Informatics, Proceedings, 3495 (2005), 27–36.

[Blondel et al. 2008] Blondel, V. D., Guillaume, J. L., Lambiotte, R., and Lefebvre, E. Fast unfolding of communities in large networks. Journal of Statistical Mechanics-Theory and Experiment, (2008).

[Evans and Lambiotte 2009] Evans, T. S. and Lambiotte, R. Line graphs, link partitions, and overlapping communities. Physical Review E, 80 (2009), 1.

[Gfeller et al. 2005] Gfeller, D., Chappelier, J.-C., and De Los Rios, P. Finding instabilities in the community structure of complex networks. Physical review. E, Statistical, nonlinear, and soft matter physics, 72 (2005), 5 Pt 2, 056135.

[Girvan and Newman 2002] Girvan, M. and Newman, M. E. J. Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA, 99, (2002), 7821–7826.

[Gregory 2010] Gregory, S. Finding overlapping communities in networks by label propagation. New Journal of Physics, 12 (2010), 10, 103018.

[Guimerà and Amaral 2005] Guimerà, R. and Amaral, L. A. N. Functional cartography of complex metabolic networks. Nature, 433 (2005), 7028, 895–900.

[Hugot et al. 2001] Hugot, J. P., Chamaillard, M., Zouali, H., Lesage, S., Cézard, J. P., Belaiche, J., Almer, S., Tysk, C., O'Morain, C. A., Gassull, M., Binder, V., Finkel, Y., Cortot, A., Modigliani, R., Laurent-Puig, P., Gower-Rousseau, C., Macry, J., Colombel, J. F., Sahbatou, M., and Thomas, G. Association of nod2 leucine-rich repeat variants with susceptibility to crohn's disease. Nature, 411 (2001), 6837, 599–603.

[Kessler 1963] Kessler, M. M. Bibliographic coupling between scientific papers. American Documentation, 14 (1963), 1, 10–25.

[Kumpula et al. 2008] Kumpula, J. M., Kivela, M., Kaski, K., and Saramaki, J. A sequential algorithm for fast clique percolation. PhysRevE.78.026109(2008).

[Lancichinetti et al. 2009] Lancichinetti, A., Fortunato, S., and Kertesz, J. Detecting the overlapping and hierarchical community structure in complex networks. New Journal of Physics, 11 (2009).

[Lancichinetti et al. 2010a] Lancichinetti, A., Kivela, M., Saramaki, J., and Fortunato, S. Characterizing the community structure of complex networks. PLoS One, 5(8),e11976(2010).

[Lancichinetti et al. 2010b] Lancichinetti, A., Radicchi, F., and Ramasco, José Javier Fortunato, S. Finding statistically significant communities in networks. PLoS ONE 6(4): e18961(2010b).

[Lancichinetti and Radicchi 2009] Lancichinetti, A. and Radicchi, Filippo Ramasco, J. J. Statistical significance of communities in networks. PhysRevE.81.046110(2009).

[Lee et al. 2010] Lee, C., Reid, F., McDaid, A., and Hurley, N. Detecting highly overlapping community structure by greedy clique expansion. ArXiv e-prints(2010).

[Limbergen et al. 2007] Limbergen, J. V., Russell, R. K., Nimmo, E. R., Torkvist, L., Lees, C. W., Drummond, H. E., Smith, L., Anderson, N. H., Gillett, P. M., McGrogan, P., Hassan, K., Weaver, L. T., Bisset, W. M., Mahdi, G., Arnott, I. D., Sjoqvist, U., Lordal, M., Farrington, S. M., Dunlop, M. G., Wilson, D. C., and Satsangi, J. Contribution of the nod1/card4 insertion/deletion polymorphism +32656 to inflammatory bowel disease in northern europe. Inflamm Bowel Dis, 13 (2007), 7, 882–889.

[Newman 2004] Newman, M. E. J. Fast algorithm for detecting community structure in networks. Phys Rev E Stat Nonlin Soft Matter Phys, 69 (2004), 6 Pt 2, 066133.

[Newman 2006] Newman, M. E. J. Finding community structure in networks using the eigenvectors of matrices. Phys. Rev. E, 74 (2006), 036104.

[Palla et al. 2005] Palla, G., Derenyi, I., Farkas, I., and Vicsek, T. Uncovering the overlapping community structure of complex networks in nature and society. Nature, 435 (2005), 814.

[Pu et al. 2009] Pu, S., Wong, J., Turner, B., Cho, E., and Wodak, S. J. Up-to-date catalogues of yeast protein complexes. Nucleic Acids Res, 37 (2009), 3, 825–831.

[Reichardt and Bornholdt 2006] Reichardt, J. and Bornholdt, S. Statistical mechanics of community detection. Phys. Rev. E, 74 (2006), 1, 016110.

[Reichardt 2004] Reichardt, Joerg Bornholdt, S. Detecting fuzzy community structures in complex networks with a potts model. PhysRevLett.93.218701(2004).

[Sales-Pardo et al. 2007] Sales-Pardo, M., Guimera, R., and Moreira, Andra A Amaral, L. A. N. Extracting the hierarchical organization of complex systems. Proc Natl Acad Sci U S A, 104 (2007), 39, 15224–15229.

[Wang et al. 2009] Wang, X. H., Jiao, L. C., and Wu, J. S. Adjusting from disjoint to overlapping community detection of complex networks. Physica a-Statistical Mechanics and Its Applications, 388 (2009), 24, 5045–5056.

[Zachary 1977] Zachary, W. W. An information flow model for conflict and fission in small groups. Journal of Anthropologica, 1 (1977), 33, 452–473.