

# Overlapping Community Structure and Modular Overlaps in Complex Networks

Qinna Wang and Eric Fleury

**Abstract** In order to find overlapping community structure of complex networks, many researchers make endeavours. Here, we first discuss some existing functions proposed for measuring the quality of overlapping community structure. Second, we propose a novel algorithm called fuzzy detection for overlapping community detection. Our new method benefits from an existing partition detection technique and aims at identifying modular overlaps. A modular overlap is a group of overlapping nodes. Therefore, the overlaps shared by several communities are possibly grouped into several different modular overlaps. The results in synthetic networks and real networks demonstrate that our method can uncover and characterize meaningful overlapping nodes.

**Keywords** Modularity · Co-citation network · Complex networks

## 1 Introduction

The empirical information of networks can be used to study structural characteristics, like heavy-tailed degree distributions [1], small-world property [3] and rumour spreading. These characteristics are 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.

Communities may thus overlap with each other. For example, people may share the same hobbies in social networks [28], some predator species have the same prey species in food webs [13] and different sciences are connected by their interdisciplinary domain in co-citation networks [20]. However, most of heuristic algorithms

---

Q. Wang (✉) · E. Fleury

DNET (ENS-Lyon/LIP Laboratoire de l'Informatique du Parallélisme/INRIA Grenoble Rhône-Alpes), Lyon, France  
e-mail: [qinna.wang@ens-lyon.fr](mailto:qinna.wang@ens-lyon.fr)

E. Fleury

e-mail: [eric.fleury@inria.fr](mailto:eric.fleury@inria.fr)

are proposed for partition detection, whose results are disjoint communities or partitions. A *partition* is a division of a graph into disjoint communities, such that each node belongs to a unique community. A division of a graph into overlapping (or fuzzy) communities is called a *cover*. We devote this paper to the detection of overlapping community structure.

In order to provide the exhaustive information about overlapping community structure of a graph, we introduce a novel quality function to measure the quality of the overlapping community structure. This quality function is derived from Reichardt and Bornholdt's work [25] and explains the quality of community structure through the energy of spin system.

Moreover, we propose a novel method called fuzzy detection for identifying overlapping nodes and detecting overlapping communities. It applies an existing and very efficient partition detection technique called Louvain algorithm [6]. When running the Louvain algorithm in a graph, we observe that some nodes are grouped together with different community members in distinct partitions. These oscillating nodes are possible overlapping nodes.

This paper is organized as following: we introduce related work in Sect. 2; next, we discuss the modified modularity for covers in Sect. 3; in Sect. 4, we describe our fuzzy detection in details, and applied to networks in Sect. 5 for which the community structure is already known from other studies, our method appears to give excellent agreement with the expected results; in Sect. 6, when applied to networks for which we do not have other information about communities, it gives promising results which may help us to understand better the interplay between network structure and function; finally, we give the conclusion and our future work in Sect. 7.

## 2 Related Work

### 2.1 Definition and Notation

Many real world problems (biological, social, web) can be effectively modeled as networks or graphs where nodes represent entities of interest and edges mimic the interactions or relationships among them. A graph  $G = (V, E)$  consists of two sets  $V$  and  $E$ , where  $V = \{v_1, v_2, \dots, v_n\}$  are the nodes (or vertices, or points) of the graph  $G$  and  $E \subseteq V \times V$  are its links (or edges, or lines). 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  $G$ . Specifically, the adjacency matrix of a finite graph  $G$  on  $n$  vertices is the  $n \times n$  matrix  $\mathbf{A} = [A_{ij}]_{n \times n}$ , where an entry  $A_{ij}$  of  $\mathbf{A}$  is equal to 1 if the link  $e_{ij} = (v_i, v_j) \in E$  exists, and zero otherwise.

A *partition* is a division of a graph into disjoint communities, such that each node belongs to a unique community. A division of a graph into overlapping (or fuzzy) communities is called a *cover*. We use  $\mathcal{P} = \{\mathcal{C}_1, \dots, \mathcal{C}_{n_c}\}$  to denote the partition, which is composed of  $n_c$  communities. In  $\mathcal{P}$ , the community to which the node  $v$

belongs to is denoted by  $\sigma_v$ . By definition we have  $V = \cup_1^{n_c} C_i$  and  $\forall i \neq j, C_i \cap C_j = \emptyset$ . We denote a cover composed of  $n_c$  communities by  $\mathcal{S} = \{S_1, \dots, S_{n_c}\}$ . In  $\mathcal{S}$ , we may find a pair of community  $S_i$  and  $S_j$  such that  $S_i \cap S_j \neq \emptyset$ .

Given a community  $\mathcal{C} \subseteq V$  of a graph  $G = (V, E)$ , we define the internal degree  $k_v^{\text{int}}$  (respectively the external degree  $k_v^{\text{ext}}$ ) of a node  $v \in \mathcal{C}$ , as the number of edges connecting  $v$  to other nodes belonging to  $\mathcal{C}$  (respectively to the rest of the graph). 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  $k^{\text{int}}$  of  $\mathcal{C}$  is the sum of the internal degrees of its nodes. Likewise, the external degree  $k^{\text{ext}}$  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:  $k_{\mathcal{C}} = k_{\mathcal{C}}^{\text{int}} + k_{\mathcal{C}}^{\text{ext}}$ .

## 2.2 Current Work

We then review existing methods for detecting overlapping community structure and discuss the shortcomings of these approaches.

Baumes et al. [4] 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 [14] and OSLOM (Order Statistics Local Optimization Method) [16]. 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 [15] of clusters with an expansive computational cost as it sweeps all nodes for each “worst” node. For the optimization, Lancichinetti et al. [16] 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 [17] and community strength-based overlapping community detection [29]. However, as they applied Lancichinetti et al. [14]’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. [25] introduced the energy landscape survey method, and Sales Pardo et al. [26] 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 [26], they are limited by scales of networks.

Evans et al. [7] 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 Modularity Extensions

Modularity has been employed by a large number of community detection methods. However, it only evaluates the quality of partitions. Here, we first introduce a novel extension for covers, which is combined with the energy model Hamiltonian for the spin system [25]. Second, we review some existing modularity extensions for covers and discuss the cases which these existing extensions may fail to capture. Studies show that our proposed modularity extension is able to avoid their shortcomings.

#### 3.1 A Novel Modularity

Many scientists deal with the problems in the area of computer science based on principles from statistical mechanics or analogies with physical models. When using spin models for clustering of multivariate data, the similarity measures are translated into coupling strengths and either dynamical properties such as spin-spin correlations are measured or energies are interpreted as quality functions. A ferromagnetic Potts model has been applied successfully by Blatt et al. [24]. Bengtsson and Roivainen [5] have used an antiferromagnetic Potts model with the number of clusters as input parameter and the assignment of spins in the ground state of the system defines the clustering solution. These works have motivated Reichardt and Bornholdt [25] to interpret the modularity of the community structure by an energy function of the spin glass with the spin states. The energy of the spin system is equivalent to the quality function of the clustering with the spins states being the community indices.

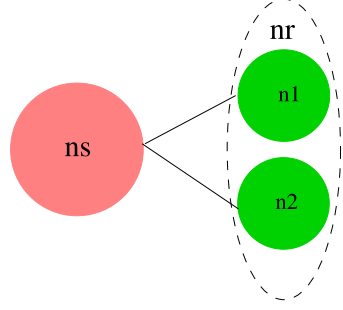
Let a community structure be represented by a spin configuration  $\{\sigma\}$  associated to each node  $u$  of a graph  $G$ . Each spin state represents a community, and the number of spin states represents the number of communities of the graph. The quality of a community structure can thus be represented through the energy of spin glass. In [25], a function of community structure is proposed, whose expression is written as:

$$\mathcal{H}(\{\sigma\}) = - \sum_{i \neq j} (A_{ij} - \gamma p_{ij}) \delta(\sigma_i, \sigma_j). \quad (1)$$

This function (Eq. 1) can be written in the following two ways:

$$\mathcal{H}(\{\sigma\}) = - \sum_s (m_{ss} - \gamma [m_{ss}]_{p_{ij}}) = - \sum_s c_s \quad (2)$$

**Fig. 1** 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



and

$$\mathcal{H}(\{\sigma\}) = \sum_{s < r} (m_{sr} - \gamma [m_{sr}]_{p_{ij}}) = \sum_s a_{sr}, \quad (3)$$

where for each community  $\mathcal{C}_s$ , we note  $m_{ss}$  the number of links within  $\mathcal{C}_s$ ,  $m_{sr}$  represents the number of links between a community  $\mathcal{C}_s$  and another community  $\mathcal{C}_r$ ,  $[m_{ss}]_{p_{ij}}$  and  $[m_{sr}]_{p_{ij}}$  are the expected number of links given a link distribution  $p_{ij}$ . The cohesion of  $\mathcal{C}_s$  is noted  $c_s$  and  $a_{sr}$  represents the adhesion between a community  $\mathcal{C}_s$  and another community  $\mathcal{C}_r$ .

We can assume diverse expressions of  $[\cdot]_{p_{ij}}$ , which is an expectation under the link distribution  $p_{ij}$ . In case of Fig. 1 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,  $[m_{1s}]_{p_{ij}} + [m_{2s}]_{p_{ij}} = [m_{1+2,s}]_{p_{ij}}$ ;
2. when  $n_r$  is an union cluster composed of  $n_1$  and  $n_2$ ,  $[m_{rr}]_{p_{ij}} = [m_{11}]_{p_{ij}} + [m_{22}]_{p_{ij}} + [m_{12}]_{p_{ij}}$ .

Similarly, we give a relation for the cohesion of a community  $n_3$  (the whole graph) and two sub-communities  $n_1$  and  $n_2$  with an empty intersection such as  $n_1 \cup n_2 = n_3$  and  $n_1 \cap n_2 = \emptyset$  (see Fig. 2(a)). From Eqs. 2 and 3, we can easily prove:

$$c_3 = c_1 + c_2 + a_{12} \quad (4)$$

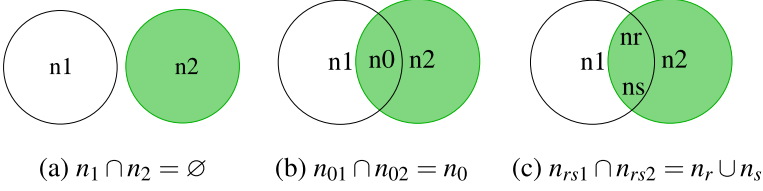
where  $c_3$  denotes the cohesion of  $n_3$  that is the union of  $n_1$  and  $n_2$  with an empty intersection,  $a_{12}$  denotes the adhesion between  $n_1$  and  $n_2$ ,  $c_1$  and  $c_2$  are the cohesions of sub-communities  $n_1$  and  $n_2$  respectively.

Furthermore, we can give the relations for the cohesion of  $n_3$  and two sub-communities  $n_1$  and  $n_2$  in other cases (see Fig. 2).

In the subdivision (see Fig. 2(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:

$$\begin{cases} c_{01}^0 = c_0^0 + c_1 + a_{01}^0, \\ c_{02}^0 = c_0^0 + c_2 + a_{02}^0, \end{cases}$$

where  $c_{01}^0$  and  $c_{02}^0$  denote the cohesion of the sub-communities  $n_{01}$  and  $n_{02}$  respectively,  $a_{01}^0$  and  $a_{02}^0$  denote the adhesion between  $n_0$  and  $n_1$ ,  $n_2$ . Here,  $n_0$  is shared by  $n_{01}$  and  $n_{02}$ .



**Fig. 2** 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  $n_r \cup n_s = n_0$

For the adhesion, we have:

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

between  $n_{01}$  and  $n_{02}$ .

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

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

So we derive

$$c_0^0 = \frac{1}{2}c_0, \quad a_{01}^0 = \frac{1}{2}a_{01} \quad \text{and} \quad a_{02}^0 = \frac{1}{2}a_{02}. \quad (5)$$

In the subdivision (see Fig. 2(c)) such as  $n_r \cup n_s = n_0$ , we replace  $c_0$  and  $c_0^0$  by

$$\begin{cases} c_0 = c_r + c_s + a_{rs}, \\ c_0^0 = c_r^r + c_s^s + a_{rs}^{rs}, \end{cases} \quad (6)$$

where  $c_r^r$  and  $c_s^s$  denote the cohesion 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. 5.

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

$$\sum_1^k \frac{1}{|d_r \cup d_s|} a_{rs} = \frac{|d_r \cap d_s|}{|d_r \cup d_s|} a_{rs}, \quad (7)$$

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

The widest used modularity [22] is given by:

$$Q = \frac{1}{2m} \sum_{i \neq j} \left( A_{ij} - \frac{k_i k_j}{2m} \right) \delta(\sigma_i, \sigma_j). \quad (8)$$

We rewrite the modularity  $Q$  Eq. 8 as:

$$Q = -\frac{1}{m} \mathcal{H}(\{\sigma\}). \quad (9)$$

Consequently, we can write the quality of an overlapping community structure in the form of the modularity function:

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

where  $d_i$  and  $d_j$  are memberships of nodes  $i$  and  $j$ , respectively. 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  $(A_{ij} - \frac{k_i k_j}{2m})$ . 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 within the range of  $[0, (A_{ij} - \frac{k_i k_j}{2m})]$ . Furthermore, if the found community structure is a strict partition, its quality  $Q_{ov}$  is equal to the initial modularity  $Q$  defined by Eq. 8.

### 3.2 Existing Modularity for Covers

There are other extensions of modularity designed to evaluate the quality of overlapping community structure. However, we are going to prove that they fail to satisfy above necessary constraints.

In the case Fig. 2(c), we assume that  $n_r$  is an overlapping node  $v_i$ . Similarly for  $n_s$ ,  $n_s$  is another overlapping node  $v_j$  which connects to  $v_i$ . The union of  $v_i$  and  $v_j$  is  $n_0$  such that  $n_0 = v_i \cup v_j$ . The overlapping communities  $n_{01}$  and  $n_{02}$  are denoted by  $\mathcal{C}_x$  and  $\mathcal{C}_y$  of a graph  $G_{\text{example}}$ , respectively.

Let  $O_v$  be the number of communities to which node  $v$  belongs. Shen et al. [27] have introduced an extended modularity:

$$Q_{\text{shen}} = \frac{1}{2m} \sum_{i=1}^{n_c} \sum_{v \in \mathcal{C}_i, w \in \mathcal{C}_j, v \neq w} \frac{1}{O_v O_w} \left( A_{vw} - \frac{k_v k_w}{2m} \right) \delta(\sigma_v, \sigma_w). \quad (11)$$

From Eq. 9, it is easy to obtain  $a_{01_{\text{shen}}}^0$  derived from  $Q_{\text{shen}}$  (Eq. 11):

$$a_{01_{\text{shen}}}^0 = \frac{1}{2} \sum_{v \in n_0, w \in \mathcal{C}_x \setminus n_0} \left( A_{vw} - \frac{k_v k_w}{2m} \right) + \frac{1}{2} \left( A_{v_i v_j} - \frac{k_{v_i} k_{v_j}}{2m} \right).$$

It fails to satisfy  $a_{01}^0 = \frac{1}{2} a_0$  (Eq. 5), where

$$a_{01_{\text{shen}}} = \sum_{v \in n_0, w \in \mathcal{C}_x \setminus n_0} \left( A_{vw} - \frac{k_v k_w}{2m} \right) + 2 \left( A_{v_i v_j} - \frac{k_{v_i} k_{v_j}}{2m} \right).$$

In other words, through the definition of  $Q_{\text{shen}}$ , we obtain different values of the quality in views of Figs. 2(b) and 2(c) although they represent the same cover.

In [21], Tamas Nepusz et al. have proposed a variant of modularity measure, which is defined by:

$$Q_{\text{fuzzy}} = \frac{1}{2m} \sum_{i,j} \left( A_{ij} - \frac{k_i k_j}{2m} \right) s_{ij}$$

where  $s_{ij} = \sum_{k=1}^{n_c} u_{ki} u_{kj}$ . The membership degree between node  $i$  and community  $k$ ,  $u_{ki}$  satisfies  $\sum_{k=1}^{n_c} u_{ik} = 1$ .

As we did previously, for node  $v_k \in n_0$  in  $G_{\text{example}}$ , under the assumption:  $u_{v_i \mathcal{C}_x} = u_{v_i \mathcal{C}_y} = u_{v_j \mathcal{C}_x} = u_{v_j \mathcal{C}_y} = \frac{1}{2}$ , it is easy to obtain

$$s_{v_k v_w} = \begin{cases} 0 & v_w \notin \mathcal{C}_x \cup \mathcal{C}_y, \\ 0.5 & v_w \in \mathcal{C}_x \cup \mathcal{C}_y, v_w \notin n_0, \\ 0.25 & v_k \neq v_w. \end{cases} \quad (12)$$

We obtain that

$$a_{01}^0_{\text{fuzzy}} = \frac{1}{2} \sum_{v \in n_0, w \in \mathcal{C}_x \setminus n_0} \left( A_{vw} - \frac{k_v k_w}{2m} \right) + \frac{1}{2} \left( A_{v_i v_j} - \frac{k_{v_i} k_{v_j}}{2m} \right).$$

It also does not satisfy  $a_{01}^0 = \frac{1}{2} a_0$  (Eq. 5) with  $a_{01}^0_{\text{fuzzy}} = a_{01}^0_{\text{shen}}$ .

By using the novel proposed modified modularity (Eq. 10), we obtain

$$a_{01}^0_{\text{ov}} = \frac{1}{2} \sum_{v \in n_0, w \in \mathcal{C}_x \setminus n_0} \left( A_{vw} - \frac{k_v k_w}{2m} \right) + \left( A_{v_i v_j} - \frac{k_{v_i} k_{v_j}}{2m} \right).$$

It satisfies  $a_{01}^0 = \frac{1}{2} a_0$  (Eq. 5), therefore we consider that our novel modified modularity is more reasonable to evaluate the quality of overlapping community structure. However, we can not detect covers by optimizing it since overlapping nodes may degenerate the modularity value. For example, in the case Fig. 2(b), the quality can be represented by

$$Q_{ov}^{\text{cover}} = -\frac{1}{m} \mathcal{H}(\{\sigma\}) = -\frac{1}{m} (c_0 + c_1 + c_2 + a_{01}^0 + a_{02}^0),$$

where  $a_{01}^0 = \frac{1}{2} a_{01}$  and  $a_{02}^0 = \frac{1}{2} a_{02}$ . And the quality of the partition is

$$Q_{ov}^{\text{partition}} = \begin{cases} -\frac{1}{m} (c_0 + c_1 + c_2 + a_{01}), & \text{when } \mathcal{P} = \{n_{01}, n_2\}, \\ -\frac{1}{m} (c_0 + c_1 + c_2 + a_{02}), & \text{when } \mathcal{P} = \{n_1, n_{02}\}. \end{cases}$$

We find  $Q_{ov}^{\text{cover}} = Q_{ov}^{\text{partition}}$  when  $a_{01} = a_{02}$ ; otherwise,  $Q_{ov}^{\text{cover}} < Q_{ov}^{\text{partition}}$  due to  $\min(a_{01}, a_{02}) < a_{01}^0 + a_{02}^0 = \frac{1}{2} a_{01} + \frac{1}{2} a_{02} < \max(a_{01}, a_{02})$ . Thus, even in a toy example where clearly there is a clear overlap (see Fig. 2(b)), if the number of links between  $n_0$  and  $n_1$  differs from the number of links between  $n_0$  and  $n_2$  the quality of the cover will be less than the quality of the partition once the difference between the number of links is greater than 0.

To overcome this optimization issue, we propose the method named fuzzy detection not based on modularity like function.

## 4 Our Method

In this section, we will introduce our method for cover detection named *fuzzy detection*. This novel cover detection heuristic aims at identifying modular overlaps.



Each modular overlap is a group of nodes shared by communities. More precisely, each modular overlap is a possible sub-community shared by several communities. For better understanding, we give two definitions of overlapping nodes: *granular overlaps* and *modular overlaps*. The traditional cover detection methods [4, 14, 16] aims at identifying *granular overlaps*, which are fine grain scale approaches. 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. As opposed to granular overlaps, modular overlaps imply the hierarchical organization of the graph: each modular overlap is a sub-community shared by several communities.

## 4.1 Motivation

Our fuzzy detection algorithm is based on the Louvain algorithm [6]. The Louvain algorithm is an efficient partition detection algorithm that provides good partitions with high modularity. It consists of two phases that are iteratively repeated until no more positive gain of modularity is obtained. Initially, all nodes are assigned into a single community. Then, for each node whose move improves the modularity, it will be removed from its current community to the neighbor community which offers the largest gain of modularity. The first phase repeatedly and sequentially sweeps all nodes until no further improvement of modularity can be gained. The second phase builds a new meta graph based on communities found in the first phase. It aggregates nodes of the same community and builds a new network whose nodes are the communities. 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, the hierarchy height is determined by the number of iterations. The Louvain algorithm is extremely fast and provides highly optimized partitions with 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 overlapping nodes. Therefore, we propose a fuzzy detection algorithm which detects groups of nodes having strong probability of appearing in several communities.

## 4.2 Fuzzy Detection Algorithm

To have the benefit of the potential Louvain algorithm instability [2], 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 produce 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  $j$  appearing in the same community. Having  $p_{ij} = 1$  implies that nodes  $i$  and  $j$  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 probability  $p_{ij}$  less than a threshold  $\alpha^*$ . After this pruning phase, a set of disjoint robust clusters is obtained. A *robust cluster* is a group of nodes connected by edges having in-cluster probability larger than or equal to  $\alpha^*$ . 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 [8] 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 Algorithm 2. First, (lines 2–9) we compute the co-appearance matrix  $\mathbf{P} = [p_{ij}]_{n \times n}$  by running the Louvain algorithm of Algorithm 1 several times with a random seed. The number of runs is determined by the convergence criteria (line 9):

$$\|\mathbf{P}^{k+1} - \mathbf{P}^k\| = \sqrt{\frac{1}{m} \sum_{(i,j) \in E} (p_{ij}^{k+1} - p_{ij}^k)^2} < \varepsilon, \quad (13)$$

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  $\varepsilon$  is a small threshold. Figure 3 illustrates the convergence of the norm when running

**Algorithm 1** Louvain algorithm**Require:**  $G = (V, E)$ ,  $l^*$  a level threshold**Ensure:**  $\mathcal{P}$  a partition

---

```

1:  $l \leftarrow 0$ ;  $G_0 \leftarrow G$ 
2: repeat
3:    $l \leftarrow l + 1$ 
4:   Initialize a partition  $\mathcal{P}_l$  of  $G_l(V_l, E_l)$ 
   // First phase: Partition update
5:   repeat
6:     Nodes in a random permutation
7:     for all Nodes:  $v \in V_l$  do
8:       Move from  $\sigma_v$  to one selected  $\sigma_{v'}$  ( $v'$  is a neighbor of  $v$ )
9:     end for
10:  until no more change increases modularity
  // Second phase: Construct a new meta graph
11:  Replace each community by a node
12:  Replace connections between a pair of communities by one weighted edge
13: until  $\mathcal{P}_l$  is not updated or  $l = l^*$ .
14: Return  $\mathcal{P}$  corresponding to the roots of the hierarchical tree.

```

---

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  $\{c_1, c_2, \dots, c_s\} = \mathcal{P}_{sc}$  (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 than a given threshold  $\alpha^*$  (typically  $\alpha^* = 0.9$ ). A simple illustration is given in Fig. 4.

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

$$\hat{c}_i = \arg \max_{c_j \subseteq \mathcal{C}_i} |c_j|. \quad (14)$$

We assign each robust cluster  $c_j$  to the community  $\mathcal{C}_i$  if and only if their community membership degree  $p_{c_j, \hat{c}_i}$  is larger than a threshold  $\beta^*$  such as  $p_{c_j, \hat{c}_i} \geq \beta^*$  (typically  $\beta^* = 0.1$ ). If one robust cluster is assigned to at least two communities, we call it a *modular overlap*.

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

Since fuzzy detection is used to identify modular overlaps, which are sub-communities 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-

**Algorithm 2** Fuzzy detection**Require:**  $G = (V, E)$ ,  $\alpha^*$ ,  $\beta^*$ **Ensure:**  $\mathcal{S}$  an overlapping community covering of  $V$ *// STEP 1: Detect robust clusters*

```

1:  $\mathbf{P}^0 \leftarrow 0$ ;  $k \leftarrow 0$ ; modularitymax  $\leftarrow -\infty$ 
2: repeat
3:    $k \leftarrow k + 1$ 
4:    $\mathcal{P} \leftarrow$  Run the Louvain algorithm on  $G$ 
5:   Update  $\mathbf{P}^k$ 
6:   if modularity of  $\mathcal{P}$  greater than modularitymax then
7:     Save the partition  $\mathcal{P}$  in  $\mathcal{P}_{\text{opt}}$  and update modularitymax
8:   end if
9: until  $\|\mathbf{P}^k - \mathbf{P}^{k-1}\| \leq \epsilon$ 
10:  $\mathcal{P}_{\text{sc}} = \mathcal{P}_{\text{opt}}$ 
11: for all edge  $e = (i, j)$  such that  $p_{ij} < \alpha^*$  do
12:   Remove the external edge  $e$  from  $\mathcal{P}_{\text{sc}}$ 
13: end for

```

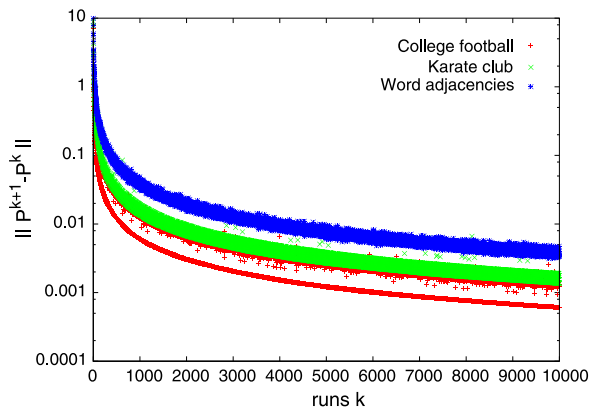
*// STEP 2: Adjust the membership of robust clusters***Require:**  $G = (V, E)$ ,  $\mathcal{P}_{\text{sc}}$ ,  $\mathcal{S} \leftarrow \mathcal{P}_{\text{opt}}$ 

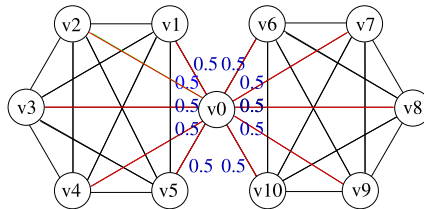
```

14: for all  $\mathcal{C}_i \in \mathcal{P}_{\text{opt}}$  do
15:   Identify community core:  $\hat{c}_i = \arg \max_{c_j \subseteq \mathcal{C}_i} |c_j|$ 
16: end for
17: Compute  $\mathbf{P}_{c_i, c_j}$ 
18: for all  $c_j \in \mathcal{P}_{\text{sc}}$  and  $c_j \notin \{\hat{c}_1, \dots, \}$  do
19:   if  $p_{c_j, \hat{c}_i} \geq \beta^*$  then
20:      $S_i \leftarrow S_i \cup c_j$ 
21:   end if
22: end for
23: Return  $\mathcal{S}$ 

```

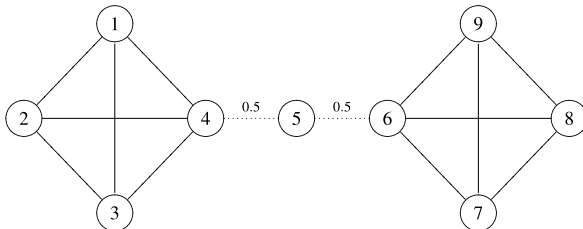
**Fig. 3** As the number of runs increases, the shape of the function value Eq. 13 gets closer and closer to 0. The figure shows results on College football [9], Karate club [30] and Word adjacencies [23]





**Fig. 4** Illustration of our fuzzy detection on a toy graph which consists of two overlapping cliques. After removing all edges in low probability  $p_{ij} = 50\%$  (which connect to the node  $v_0$ ), 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  $v_0$

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



appearance probability with several communities. Figure 5 illustrates such case. Due to unstable nodes, we only use fuzzy detection to identify modular overlaps.

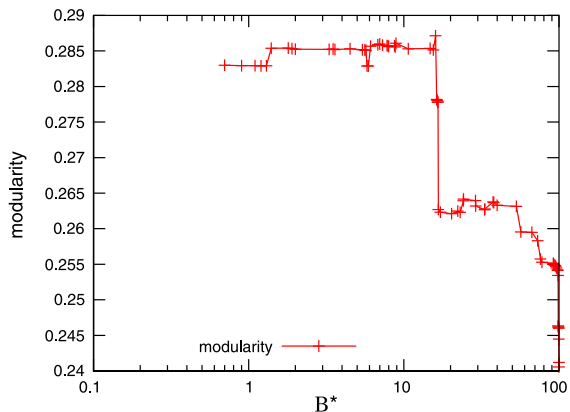
The running time of fuzzy detection mainly depends on the co-appearance matrix calculation. The complexity to find a partition by the Louvain algorithm is estimated by authors in [6] to be in  $\mathcal{O}(m)$ , where  $m$  is the number of edges in the network (the worst complexity is much higher, but 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 than time computing one.

### 4.3 Discussion

Our fuzzy detection has applied  $\beta^*$  to determine community memberships. If the threshold  $\beta^*$  increased, the number of modular overlaps decreased; otherwise, more robust clusters are identified as modular overlaps. The criterion we used to fix the optimal  $\beta^*$  value should be based on finding a community structure having the good quality. In the following, we apply our method to a real network and study the modularity by increasing the value of  $\beta^*$ .

Wikipedia is a free encyclopedia written collaboratively by volunteers around the world. A small part of Wikipedia contributors are administrators, who are users with

**Fig. 6** Performance of fuzzy detection in testing Wikipedia vote network, where the value of the modularity corresponds to the community structure obtained by the relevant  $\beta^*$ . The critical point which corresponds to the maximum modularity is observed



access to additional technical features that aid in maintenance. In order for a user to become an administrator a Request for adminship (RfA) is issued and the Wikipedia community via a public discussion or a vote decides who to promote to adminship. Using the dump of Wikipedia page edit history, 2,794 elections with 103,663 total votes and 7,066 users participating in the elections (either casting a vote or being voted on) are extracted. About half of the votes in the dataset are by existing admins, while the other half comes from ordinary Wikipedia users.<sup>1</sup>

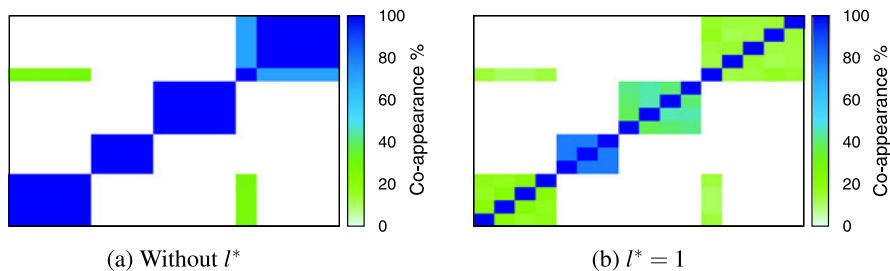
By applying our method to the Wikipedia vote network, we show the modularity by increasing the value of  $\beta^*$ . We observe the critical point:  $\beta^* = 18\%$  in Fig. 6, which corresponds to the maximum modularity Eq. 10. In practice, we use the value corresponding to the critical point to set  $\beta^*$  which is approximate 10%. Note that we do not set a high value upon  $\beta^*$  since the obtained membership degree is obtained by modularity optimization. Such that the membership degree  $p_{c_j, \hat{c}_i}$  value must be very high if the robust cluster  $c_j$  obtains the highest modularity gain with the community  $\mathcal{C}_i$  than others. (Even if the modularity gain variance between  $\mathcal{C}_i$  and another community is very slight.)

## 5 Tests of the Method

In the following, we test the performances of fuzzy detection. We have considered a set of synthetic networks and a real network for which the community structure is known. The results show that our fuzzy detection algorithm extracts communities while preserving the *hierarchical organization* and also providing overlaps.

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

<sup>1</sup><http://snap.stanford.edu/data/wiki-Vote.html>.



**Fig. 7** 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 %

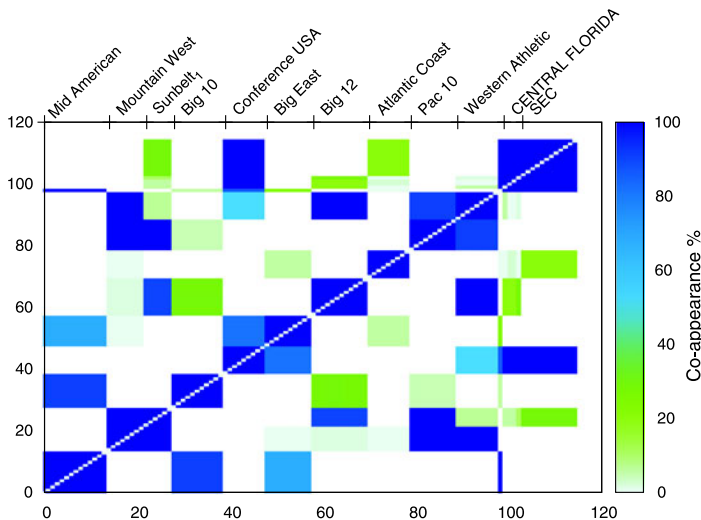
within large communities at higher levels. As an example, one may consider in a social network the granularity of the living place (town), the working place (school) and refine it toward the graduate or class level.

## 5.1 Synthetic Graphs Containing Hierarchical Structure

First, we apply the fuzzy detection algorithm to an artificial graph containing hierarchical structure [14] and a modular overlap.

The result is shown in Fig. 7. We observe that fuzzy detection extracts communities in hierarchical organization. The graph is composed 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 nodes in the same micro-community,  $k_2 = 13$  links with 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 overlaps has macro-links with two communities, its nodes have a total degree  $k = 61$  while the other nodes only have a total degree  $k = 48$ . This process constructs two hierarchical levels: one consisting of 16 small groups, and the other one composed of 4 supergroups. Figure 7(a) illustrates the co-appearance matrix by running the Louvain algorithm without fixing the level threshold  $l^*$  (see Algorithm 1), while Fig. 7(b) 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. 7(a), we use Fig. 7(b) for the visualization. We observe 4 communities and 16 robust clusters, where one robust cluster is shared by two communities. The result agrees with the ground truth.

Remark that, when running our fuzzy detection to identify modular overlaps, we may need to increase the value of  $\alpha^*$  to obtain a reasonable community core whose size is larger than the others within the same community. It occurs when one community contains several large robust clusters having comparable size.



**Fig. 8** The co-appearance matrix of college football network by running our fuzzy detection. We order the nodes corresponding to their conferences and mark the conference indices. 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 %

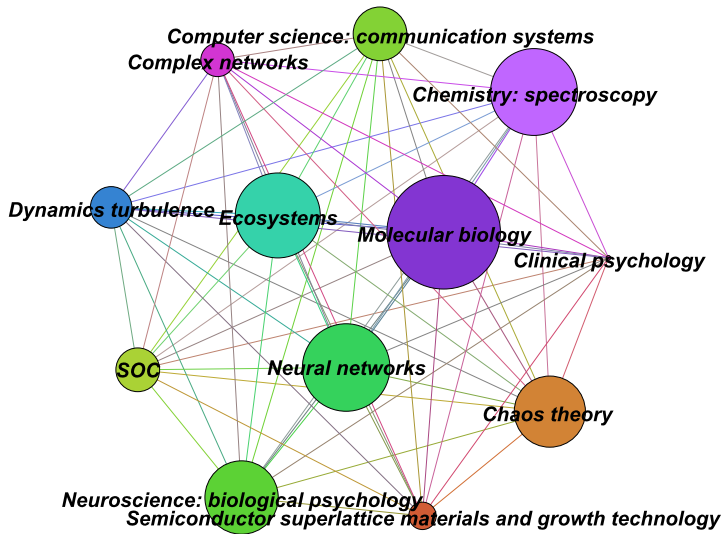
## 5.2 College Football Network

We also run the fuzzy detection algorithm to real networks. A famous real but small and tractable network is the *US college football* [9]. This network records the schedule of Division I games for the 2000 season: 115 nodes represent teams (identified by their college names) and 613 edges represent regular season games between the two teams they connect. What makes this network interesting [9] is that it incorporates a known community structure. The teams are divided into “conferences” containing around 8 to 12 teams each. Games are more frequent between members of the same conference than between members of different conferences, with teams playing an average of about 7 intra-conference games and 4 inter-conference games fraction of vertices classified correctly in the 2000 season. Inter-conference play is not uniformly distributed; teams that are geographically close to one another but belong to different conferences are more likely to play one another than teams separated by large geographic distances.

In Fig. 8, we illustrate the results: the community “Mountain West Sunbelt” is split into “Mountain West” and “Sunbelt<sub>1</sub>”, the community “Sunbelt SEC” has a possible subdivision into “Sunbelt<sub>2</sub>”<sup>2</sup> and “SEC”, and a node “CentralFlorida” is split from the community “Pac 10”. Among them, only “Sunbelt<sub>1</sub>” is identified

<sup>2</sup>We do not mark “Sunbelt<sub>2</sub>” due to the visualization, since its position is too close to “CentralFlorida” in the figure.



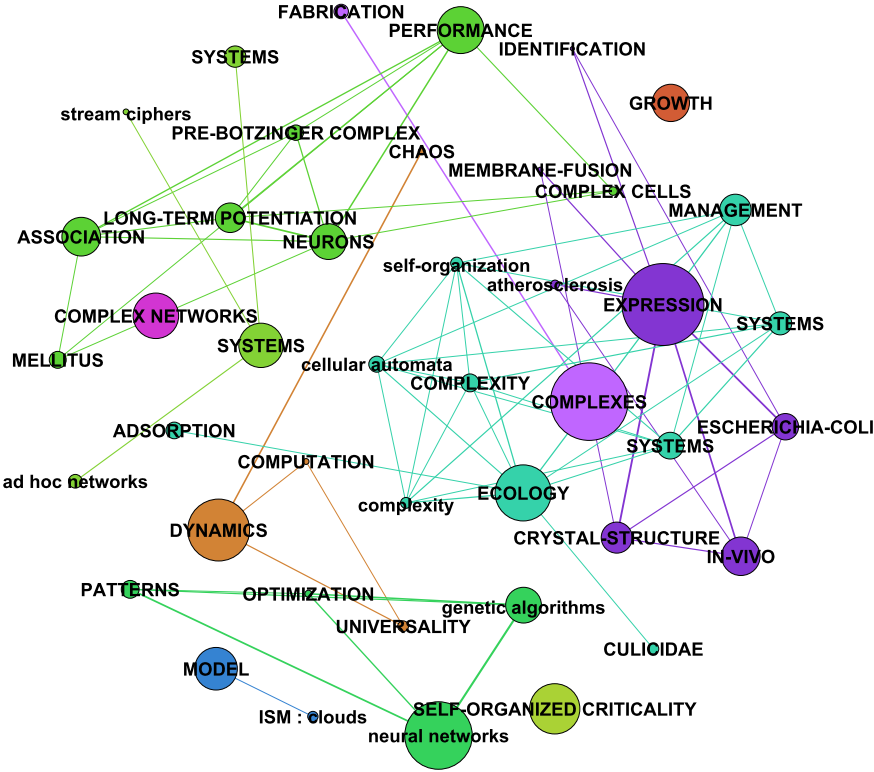


**Fig. 9** The community structure of Complex System Science, in which communities are identified by complex systems fields

as a modular overlaps. “CentralFlorida” has high membership degree with different communities, too. But it is a granular overlapping node rather than a modular overlap. In reality, the team “CentralFlorida” did not belong to any conference, and the teams in the “Sunbelt” conference played nearly as many games against Western Athletic teams as they did within their own conference. Therefore, we consider fuzzy detection has a good performance in detecting modular overlaps for this real network.

## 6 Application to a Real Network: Complex System Science

In this section we consider the application of 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 [12]). 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:  $w_{ij} = \frac{|R_i \cap R_j|}{\sqrt{|R_i| |R_j|}}$ .



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

For the visualization, we only show clusters which contain at least 100 nodes.<sup>3</sup> The partition of the graph is shown in Fig. 9. Each community corresponds to a unique color. Our obtained robust clusters are shown in Fig. 10. The color of each robust cluster corresponds to the relevant community in the partition shown in Fig. 9. Only robust clusters belonging to the same community in the partition share the same color.

Figure 9 shows 12 communities (fields or disciplines). Through studies in topic keywords,<sup>4</sup> see Table 1, we observe nearly all important fields of complex systems such as: complex networks, neural networks, self-organization criticality, dynamical systems (chaos theory, dynamics turbulence) and so on [10]. It shows that the community structure of this network reveals the complex systems fields. For more

<sup>3</sup>In [18], the community which has size roughly 100 nodes is good.

<sup>4</sup>We compute the frequency of topic keywords by aggregating the number of units (article), i.e., if only one unite contains the topic keywords “Neurons”, the corresponding frequency is 1.

**Table 1** 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

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 Networks	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 Networks	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 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

Modular overlaps	High frequent topic keywords	Involving communities
Genetic Association	<i>Association</i> , Susceptibility, Polymorphism, Linkage Disequilibrium, Disease, Major Histocompatibility Complex, Linkage, Complex Traits, Risk, Population	Molecular Biology, Neuroscience: Biological Psychology
Discrete-event Systems	<i>Systems</i> , Supervisory Control, Petri Nets, Complexity, Discrete-Event Systems, Verification, Design, Automata, Synchronization, Discrete Event Systems	Computer Science: Communication Systems, Ecosystems
Computational Complexity	<i>Complexity</i> , Algorithms, Computational Complexity, Algorithm, Networks, Optimization, Time, Systems, Search, Computational-Complexity	Computer Science: Communication Systems, Ecosystems
Astronomy-ISM (Interstellar Medium)	<i>Turbulence</i> , 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	<i>Systems</i> , Multi-Agent Systems, Multiagent Systems, Design, Agents, Architecture, Multi-Agent System, Framework, Model, Intelligent Agents	Computer Science: Communication Systems, Ecosystems
Visual Cortex	<i>Complex Cells</i> , 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

details, we analyze robust clusters, which can be considered as sub-communities (subfields or subdisciplines). The result is depicted on Fig. 10, whose description is listed in Table 3. It is no surprise to observe the connection between subfields and fields. 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 subfields refer to the study in neuroscience and biological psychology. It reveals that fuzzy detection extracts communities in hierarchical organization.

In terms of modular overlaps, our results are shown in Table 2. 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 modeling and analyzing general systems, computational com-

**Table 3** 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

Community	Cluster	High frequent topic keywords
Dynamics Turbulence	Flow Over Complex Terrain	<i>Turbulence</i> , Model, Flow, Simulation, Complex Terrain, Large-Eddy Simulation, Flows, Behavior, Boundary-Layer, Plasticity
	Astronomy-Ism (Interstellar Medium)	<i>Turbulence</i> , 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	<i>Systems</i> , Performance, Channels, Synchronization, Fading Channels, Capacity, Ofdm, Equalization, Networks, Multiuser Detection
	Control Theory	<i>Systems</i> , Stability, Design, Robust Control, Optimization, Linear-Systems, Model-Predictive Control, Stabilization, H-Infinity Control, Model Predictive Control
	Wireless Network	<i>Ad Hoc Networks</i> , Sensor Networks, Wireless Sensor Networks, Self-Organization, Networks, Wireless Networks, Clustering
	Cryptography	<i>Stream Ciphers</i> , Cryptanalysis, Linear Complexity, Stream Cipher, Sequences
Molecular Biology	Expression	<i>Expression</i> , Complex, Gene-Expression, Protein, Saccharomyces-Cerevisiae, Gene, Activation, In-Vivo, Identification, In-Vitro
	Dendritic Cells	<i>Dendritic Cells</i> , In-Vivo, Expression, T-Cells, Infection, Complex, Mice, Activation, Major Histocompatibility Complex, Antigen
	Crystal structure of Escherichia Coli	<i>Crystal-Structure</i> , Complex, Escherichia-Coli, Binding, Protein, Recognition, Mechanism, Proteins, Molecular-Dynamics, Complexes
	Gene Expression In Escherichia Coli	<i>Escherichia-Coli</i> , Gene-Expression, Systems, Expression, Model, Networks, Systems Biology, Protein, Transcription, Rhythms
	Atherosclerosis	<i>Atherosclerosis</i> , Inflammation, Expression, Disease, Myocardial-Infarction, In-Vivo, C-Reactive Protein, Smooth-Muscle-Cells, Activation, Low-Density-Lipoprotein

**Table 3** (Continued)

Community	Cluster	High frequent topic keywords
Molecular Biology	Membrane Fusion And Exocytosis	<i>Membrane-Fusion</i> , Neurotransmitter Release, Exocytosis, Syntaxin, Snare, Complex, Protein, Snare Complex, Transmitter Release
	Proteomics	<i>Identification</i> , Proteomics, Mass-Spectrometry, Proteins, Peptides, Protein Identification
Chaos Theory	Chaotic Dynamics	<i>Chaos</i> , Dynamics, Systems, Complexity, Stability, Model, Time-Series, Synchronization, Nonlinear Dynamics, Bifurcation
	Quantum Chaos And Universality	<i>Universality</i> , Quantum Chaos, Systems, Chaos, States, Model, Random- Matrix Theory, Complex Systems, Fluctuations, Spectra
	Chaos In Population dynamics	<i>Chaos</i> , Stability, Dynamics, Population, Permanence, Models, Systems, Bifurcation, Predator-Prey System, Birth Pulses
Neuroscience: Biological Psychology	Neuroplasticity	<i>RAT</i> , Neurons, Plasticity, Hippocampus, Brain, Central-Nervous-System, Synaptic Plasticity, Long-Term Potentiation, Food-Intake, Memory
	Long-Term Potentiation	<i>Long-Term Potentiation</i> , Synaptic Plasticity, Plasticity, Hippocampus, Nmda Receptor, Glutamate Receptors, Expression, Neurons, In-Vivo, Hippocampal-Neurons
	Genetic Association	<i>Association</i> , Susceptibility, Polymorphism, Linkage Disequilibrium, Disease, Major Histocompatibility Complex, Linkage, Complex Traits, Risk, Population
	Pre-Botzinger Complex	<i>Pre-Botzinger Complex</i> , In-Vitro, Pre-Botzinger Complex, Brain-Stem, Respiratory Rhythm Generation, Rhythm Generation, Rat, Control of Breathing, Neurons, Pacemaker Neurons
	Prefrontal Cortex	<i>Performance</i> , Attention, Fmri, Children, Prefrontal Cortex, Brain, Working-Memory, Cortex, Memory, Activation
	Diabetes Mellitus	<i>Mellitus</i> , Glycemic Control, Complications, Hypertension, Randomized Controlled-Trial, Diabetes, Therapy, Risk, Diabetes Mellitus, Management

**Table 3** (Continued)

Community	Cluster	High frequent topic keywords
Chemistry: Spectroscopy	Crystal Structure	<i>Complexes</i> , Self-Organization, Crystal-Structure, Derivatives, Chemistry, Polymers, Behavior, Films, Nonlinear-Optical Properties, Phase-Transition
	Anodic Alumina	<i>Fabrication</i> , Arrays, Films, Anodic Alumina, Anodization, Self-Organization, Growth, Self-Organized Formation, Hexagonal Pore Arrays, Titanium
Soc	Soc	<i>Self-Organized Criticality</i> , Model, Dynamics, Econophysics, Evolution, Systems, Fluctuations, Models, Behavior, Turbulence
Ecosystems	Innovation Management	<i>Management</i> , Innovation, Economics, Performance, Model, Complexity, Systems, Technology, Firm, Knowledge
	Discrete-Event Systems	<i>Systems</i> , Supervisory Control, Petri Nets, Complexity, Discrete-Event Systems, Verification, Design, Automata, Discrete Event Systems, Synchronization
	Computational Complexity	<i>Complexity</i> , Algorithms, Computational Complexity, Algorithm, Networks, Optimization, Time, Systems, Search, Computational-Complexity
	Ecosystems	<i>Ecology</i> , Dynamics, Evolution, Biodiversity, Patterns, Diversity, Growth, Model, Management, Conservation
	Absorption	<i>Adsorption</i> , Sorption, Speciation, Complexation, Humic Substances, Water, Natural-Waters, Kinetics, Ph, Copper
	Cellular Automaton	<i>Cellular Automata</i> , Systems, Simulation, Self-Organization, Model, Cellular-Automata, Flow, Cellular-Automaton Model, Traffic Flow, Dynamics
	Multi-agent Systems	<i>Systems</i> , Multi-Agent Systems, Multiagent Systems, Design, Agents, Architecture, Multi-Agent System, Framework, Model, Intelligent Agents
	Division of Labor in Insect Societies	<i>Self-Organization</i> , Behavior, Division-Of-Labor, Hymenoptera, Ants, Colonies, Formicidae, Social Insects, Swarm Intelligence, Evolution
	Complex Adaptive Systems	<i>Complexity</i> , Self-Organization, Chaos, Emergence, Science, Complex Adaptive Systems, Complexity Theory

**Table 3** (Continued)

Community	Cluster	High frequent topic keywords
Ecosystems	Malaria	<i>Malaria</i> , Culicidae, Identification, Transmission, Complex, Diptera, Africa, Mosquitos, Anopheles-Gambiae Complex, Gambiae Complex
Neural Networks	Neural Networks	<i>Neural Networks</i> , Classification, Systems, Model, Self-Organizing Map, Neural Network, Algorithm, Identification, Artificial Neural Networks, Prediction
	Genetic Algorithm	<i>Optimization</i> , Genetic Algorithms, Genetic Algorithm, Design, Systems, Neural Networks, Model, Algorithm, Algorithms, Simulation
	Simulated Annealing	<i>Optimization</i> , Simulated Annealing, Algorithm, Model
	Gene Expression Patterns	<i>Patterns</i> , Self-Organizing Maps, Gene-Expression, Microarray, Identification, Gene Expression, Saccharomyces-Cerevisiae, Cancer, Expression, Classification
Complex Systems	Complex Systems	<i>Complex Networks</i> , Dynamics, Small-World Networks, Model, Internet, Networks, Evolution, Scale-Free Networks, Systems, Organization

plexity is a common property of complex systems, and genetic expression [11, 19] studies are often used to determine whether a genetic variant is associated with a disease or trait. Visual cortex is one part of visual systems, which receives visual information for processing images. These results can be validated from the trivial. This also suggests that the interdisciplinarity is important in studies of complex systems.

## 7 Conclusion

In this paper, we introduce a new extension of modularity for covers and a new method for overlapping community detection. Our definition of modularity is derived from Reichardt and Bornholdt's work [25] and explains the quality of community structure through the energy of spin system. The proposed fuzzy detection benefits from the Louvain algorithm and detects modular overlaps. Modular overlaps are groups of nodes shared by several communities. We have tested our fuzzy detection on synthetic networks and observed its good performances by comparing



to the ground truth. Its application to a real network also hints that our algorithm provides insights in characterizing overlapping nodes.

We hope that our idea and method will provide useful information in the analysis of other types of networks. Possible further applications to dynamic networks will be done for studying effects of overlaps in community changes. We hope to see such applications in the future.

## References

1. Albert R, Barabasi A-L (2002) Statistical mechanics of complex networks. *Rev Mod Phys* 74(1):47–97
2. Aynaud T (2011) Détection de communautés dans les réseaux dynamiques. PhD thesis, Docteur de L'université Pierre et Marie Curie
3. Barmpoutis RM, Murray D (2010) Networks with the smallest average distance and the largest average clustering. [arXiv:1007.4031](https://arxiv.org/abs/1007.4031) [q-bio.MN]
4. Baumes J, Goldberg M, Magdon-Ismaïl M (2005) Efficient identification of overlapping communities. In: *Intelligence and security informatics, proceedings*, vol 3495, pp 27–36
5. Bengtsson M, Roivainen P (1995) Using the potts glass for solving the clustering problem. *Int J Neural Syst* 6(2):119–132
6. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. *J Stat Mech Theory Exp* 2008(10):P10008. doi:[10.1088/1742-5468/2008/10/p10008](https://doi.org/10.1088/1742-5468/2008/10/p10008)
7. Evans TS, Lambiotte R (2009) Line graphs, link partitions, and overlapping communities. *Phys Rev E* 80(1):016105
8. Gfeller D, Chappelier J-C, De Los Rios P (2005) Finding instabilities in the community structure of complex networks. *Phys Rev E, Stat Nonlinear Soft Matter Phys* 72(5):056135
9. Girvan M, Newman MEJ (2002) Community structure in social and biological networks. *Proc Natl Acad Sci USA* 99:7821–7826
10. Grauwin S, Beslon G, Fleury E, Franceschelli S, Robardet C, Rouquier J-B, Jensen P (2012) Complex systems science: dreams of universality, interdisciplinarity reality. *J Am Soc Inf Sci Technol* 63(7):1327–1338
11. Hugot JP, Chamaillard M, Zouali H, Lesage S, Cézard JP, Belaiche J, Almer S, Tysk C, O'Morain CA, Gassull M, Binder V, Finkel Y, Cortot A, Modigliani R, Laurent-Puig P, Gower-Rousseau C, Macry J, Colombel JF, Sahbatou M, Thomas G (2001) Association of nod2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* 411(6837):599–603
12. Kessler MM (1963) Bibliographic coupling between scientific papers. *Am Doc* 14(1):10–25
13. Krause AE, Frank KA, Mason DM, Ulanowicz RE, Taylor WW (2003) Compartments revealed in food-web structure. *Nature* 426(6964):282–285
14. Lancichinetti A, Fortunato S, Kertesz J (2009) Detecting the overlapping and hierarchical community structure in complex networks. *New J Phys* 11:033015
15. Lancichinetti A, Radicchi F, Ramasco JJ, Fortunato S (2010) Finding statistically significant communities in networks. *PLoS ONE* 6(4):e18961. doi:[10.1371/journal.pone.0018961](https://doi.org/10.1371/journal.pone.0018961)
16. Lancichinetti A, Radicchi F, Ramasco JJ (2009) Statistical significance of communities in networks. *Phys Rev E* 81:046110. [arXiv:0907.3708](https://arxiv.org/abs/0907.3708) [physics.soc-ph]
17. Lee C, Reid F, McDaid A, Hurley N (2010) Detecting highly overlapping community structure by greedy clique expansion. In: *Proceedings of the 4th SNA-KDD workshop*. [arXiv:1002.1827](https://arxiv.org/abs/1002.1827) [physics.data-an]
18. Leskovec J, Lang KJ, Dasgupta A, Mahoney MW (2009) Community structure in large networks: natural cluster sizes and the absence of large well-defined clusters. *Internet Math* 6(1):29–123. [arXiv:0810.1355](https://arxiv.org/abs/0810.1355)

19. Limbergen JV, Russell RK, Nimmo ER, Torkvist L, Lees CW, Drummond HE, Smith L, Anderson NH, Gillett PM, McGrogan P, Hassan K, Weaver LT, Bisset WM, Mahdi G, Arnott ID, Sjoqvist U, Lordal M, Farrington SM, Dunlop MG, Wilson DC, Satsangi J (2007) Contribution of the *nod1/card4* insertion/deletion polymorphism +32656 to inflammatory bowel disease in northern Europe. *Inflamm Bowel Dis* 13(7):882–889
20. Michon F, Tummers M (2009) The dynamic interest in topics within the biomedical scientific community. *PLoS ONE* 4(8):e6544–08
21. Nepusz T, Petroczi A, Negyessy L, Bazso F (2008) Fuzzy communities and the concept of bridgeness in complex networks. *Phys Rev E, Stat Nonlinear Soft Matter Phys* 77(1):016107
22. Newman MEJ (2004) Analysis of weighted networks. *Phys Rev E* 70:056131
23. Newman MEJ (2006) Finding community structure in networks using the eigenvectors of matrices. *Phys Rev E* 74:036104
24. Pu S, Wong J, Turner B, Cho E, Wodak SJ (2009) Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Res* 37(3):825–831
25. Reichardt J, Bornholdt S (2006) Statistical mechanics of community detection. *Phys Rev E* 74(1):016110
26. Sales-Pardo M, Guimera R, Moreira A, Amaral L (2007) Extracting the hierarchical organization of complex systems. *Proc Natl Acad Sci USA* 104(39):15224–15229
27. Shen HW, Cheng XQ, Guo JF (2009) Quantifying and identifying the overlapping community structure in networks. *J Stat Mech Theory Exp* P07042. doi:[10.1088/1742-5468/2009/07/P07042](https://doi.org/10.1088/1742-5468/2009/07/P07042)
28. Traud A, Kelsic E, Mucha P, Porter M (2009) Community structure in online collegiate social networks. *J Stat Mech Theory Exp* 2009:P07042
29. Wang XH, Jiao LC, Wu JS (2009) Adjusting from disjoint to overlapping community detection of complex networks. *Phys A, Stat Mech Appl* 388(24):5045–5056
30. Zachary WW (1977) An information flow model for conflict and fission in small groups. *J Anthropol* 1(33):452–473

Mining Social Networks and Security Informatics

Özyer, T.; Erdem, Z.; Rokne, J.; Khoury, S. (Eds.)

2013, VI, 283 p. 92 illus., Hardcover

ISBN: 978-94-007-6358-6