Evidential communities for complex networks

Kuang Zhou[1,2], Arnaud Martin[2], and Quan Pan[1]

[1] School of Automation, Northwestern Polytechnical University,
Xi'an, Shaanxi 710072, PR China
[2] IRISA, University of Rennes 1, Rue E. Branly, 22300 Lannion, France
kzhoumath@163.com, Arnaud.Martin@univ-rennes1.fr, quanpan@nwpu.edu.cn

Abstract. Community detection is of great importance for understanding graph structure in social networks. The communities in real-world networks are often overlapped, *i.e.* some nodes may be a member of multiple clusters. How to uncover the overlapping communities/clusters in a complex network is a general problem in data mining of network data sets. In this paper, a novel algorithm to identify overlapping communities in complex networks by a combination of an evidential modularity function, a spectral mapping method and evidential c-means clustering is devised. Experimental results indicate that this detection approach can take advantage of the theory of belief functions, and preforms good both at detecting community structure and determining the appropriate number of clusters. Moreover, the credal partition obtained by the proposed method could give us a deeper insight into the graph structure.

Keywords: Evidential modularity; Evidential c-means; Overlapping communities; Credal partition:

1 Introduction

In order to have a better understanding of organizations and functions in the real networked system, the community structure, or the clustering in the graph is a primary feature that should be taken into consideration [3]. As a result, community detection, which can extract specific structures from complex networks, has attracted considerable attention crossing many areas from physics, biology, and economics to sociology [1], where systems are often represented as graphs.

Generally, a community in a network is a subgraph whose nodes are densely connected within itself but sparsely connected with the rest of the network [17]. Many of the community detection approaches are in the frame of probability theory, that is to say, one actor in the network can belong to only one community of the graph [9, 4]. However, in real-world networks, each node can fully or partially belong to more than one associated community, and thus communities often overlap to some extent [11, 15]. For instance, in collaboration networks, a researcher may be active in many areas but with different levels of commitment, and in social networks, an actor usually has connections to several social groups like family, friends, and colleagues. In biological networks, a node might have multiple functions [11].

In the last decades, for identifying such clusters that are not necessarily disjoint, there is growing interest in overlapping community detection algorithms. Zhang et al. [17] devised a novel algorithm to identify overlapping communities in complex networks based on fuzzy c-means (FCM). Nepusz et al. [8] created an optimization algorithm for determining the optimal fuzzy membership degrees, and a new fuzzified variant of the modularity function is introduced to determine the number of communities. Havens et al. [5, 6] discussed a new formulation of a fuzzy validity index and pointed out this modularity measure performs better compared with the existing ones.

As can be seen, most of methods for uncovering the overlapping community structure are based on the idea of fuzzy partition, which subsumes crisp partition, resulting in greater expressive power of fuzzy community detection compared with hard ones. Whereas credal partition [2], which is even more general and allows in some cases to gain deeper insight into the structure of the data, it has not been applied to community detection.

In this paper, an algorithm for detecting overlapping community structure is proposed based on credal partition. An evidential modular function is introduced to determine the optimal number of communities. Spectral relaxation and evidential c-means are conducted to obtain the basic belief assignment (bba) of each nodes in the network. The experiments on two well-studied networks show that meaningful partitions of the graph could be obtained by the proposed detection approach and it indeed could provide us more informative information of the graph structure than the existing methods.

2 Background

2.1 Modularity-based community detection

Let $G(V, E, W)$ be an undirected network, V is the set of n nodes, E is the set of m edges, and W is a $n \times n$ edge weight matrix with elements $w_{ij}, i, j = 1, 2, \cdots, n$. The objective of the hard (crisp) community detection is to divide graph G into c clusters, denoted by

$$\Omega = \{\omega_1, \omega_2, \cdots, \omega_c\}, \tag{1}$$

and each node should belong to one and only one of the detected communities [8]. Parameter c can be given in advanced or determined by the detection method itself.

The modularity, which measures the quality of a partition of a graph, was first introduced by Newman and Girvan [10]. This validity index measures how good a specific community structure is by calculating the difference between the actual edge density intra-clusters in the obtained partition and the expected one under some null models, such as random graph. One of the most popular form of modularity is given by [3]. Given a partition with c group shown in Eq. (1), and let $\|W\| = \sum_{i,j=1}^{n} w_{ij}$, $k_i = \sum_{j=1}^{n} w_{ij}$, its modularity can be defined as:

$$Q_h = \frac{1}{\|W\|} \sum_{k=1}^{c} \sum_{i,j=1}^{n} (w_{ij} - \frac{k_i k_j}{\|W\|}) \delta_{ik} \delta_{jk}, \tag{2}$$

where δ_{ik} is one if vertex i belongs to the k_{th} community, 0 otherwise.

The communities of graph G can be detected by modularity optimization, like spectral clustering algorithm [13], which aims at finding the optimal partition with the maximum modularity value [3].

2.2 Belief function theory and evidential *c*-means

The credal partition, a general extension of the crisp and fuzzy ones in the theoretical framework of belief function theory, has been introduced in [2, 7]. Suppose the discernment frame of the clusters is Ω as in Eq. (1). Partial knowledge regarding the actual cluster node n_i belongs to can be represented by a basis belief assignment defined as a function m from the power set of Ω to $[0, 1]$, verifying $\sum_{A\subseteq\Omega} m(A) = 1$. Every $A \in 2^{\Omega}$ such that $m(A) > 0$ is called a focal element. The credibility and plausibility functions are defined in Eq. (3) and Eq. (4).

$$Bel(A) = \sum_{\emptyset\neq B\subseteq A} m(B), \forall A \subseteq \Omega, \tag{3}$$

$$Pl(A) = \sum_{B\cap A\neq\emptyset} m(B), \forall A \subseteq \Omega. \tag{4}$$

Each quantity $Bel(A)$ represents the degree to which the evidence supports A, while $Pl(A)$ can be interpreted as an upper bound on the degree of support that could be assigned to A if more specific information is available [12]. The function $pl : \Omega \to [0, 1]$ such that $pl(\omega) = Pl(\{\omega\})$ is called the contour function associated to m.

The bbas in the credal level can be expressed in the form of probabilities by pignistic transformation [2], which is defined as

$$BetP(\omega_i) = \sum_{\omega_i\in A\subseteq\Omega} \frac{m(A)}{|A|(1 - m(\emptyset))}, \tag{5}$$

where $|A|$ is the number of elements of Ω in A.

Evidential c-means (ECM) [7] is a direct generalization of FCM. The optimal credal partition is obtained by minimizing the following objective function:

$$J_{\text{ECM}} = \sum_{i=1}^{n} \sum_{A_j\subseteq\Omega, A_j\neq\emptyset} |A_j|^{\alpha} m_i(A_j)^{\beta} d_{ij}^2 + \sum_{i=1}^{n} \delta^2 m_i(\emptyset)^{\beta}, \tag{6}$$

constrained on

$$\sum_{A_j\subseteq\Omega, A_j\neq\emptyset} m_i(A_j) + m_i(\emptyset) = 1, \tag{7}$$

where $m_i(A_j)$ is the bba of n_i given to the nonempty set A_j, while $m_i(\emptyset)$ is the bba of n_i assigned to the emptyset. The value d_{ij} denotes the distance between n_i and the barycenter associated to A_j, and $|\cdot|$ is the cardinal of the set. Parameters α, β, δ are adjustable and can be determined based on the requirement.

3 Evidential community detection

Before presenting the credal partition of a graph $G(V, E, W)$, the hard and fuzzy partitions are firstly recalled. The crisp partition can be represented by a matrix $U^h = (u_{ik})_{n\times c}$, where $u_{ik}^h = 1$ if the i_{th} node n_i belongs to the k_{th} cluster ω_i in the partition, and $u_{ik}^h = 0$ otherwise. From the property of this partition, it clearly should satisfy that $\sum_{k=1}^{c} u_{ik}^h = 1, i = 1, 2, \cdots, n$. The generalization of the hard partition, following that a node may belong to more communities than one but with different degrees, can be described by the fuzzy partition matrix $U^f = (u_{ik})_{n\times c}$, where u_{ik}^f is not restricted in $\{0, 1\}$ but can attain any real value from the interval $[0, 1]$. The value u_{ik}^f could be interpreted as a degree of membership of n_i to community ω_k.

The credal partition of G, which refers to the framework of belief function theory, can be represented by a n-tuple: $M = (\boldsymbol{m}_1, \boldsymbol{m}_2, \cdots, \boldsymbol{m}_n)$. Each $\boldsymbol{m}_i = \{m_{i1}, m_{i2}, \cdots, m_{i2^c}\}$ is a bba in a 2^c-dimensional space, where c is the cardinality of the given discernment frame of communities $\Omega = \{\omega_1, \omega_2, \cdots, \omega_c\}$ as before, and ω_i denotes the i_{th} detected community. Note that Ω is the discernment frame in the framework of belief function theory.

3.1 The evidential modular function

Similar to the fuzzy modularity by Nepusz et al. [8] and by Havens et al. [5], here we introduce an evidential modularity:

$$Q_e = \frac{1}{\|W\|} \sum_{k=1}^{c} \sum_{i,j=1}^{n} (w_{ij} - \frac{k_i k_j}{\|W\|}) pl_{ik} pl_{jk}, \tag{8}$$

where $\boldsymbol{pl}_i = \{pl_{i1}, pl_{i2}, \cdots, pl_{ic}\}$ is the contour function associated to m_i, which describes the upper value of our belief to the proposition that the i_{th} node belongs to the k_{th} community.

Let $\boldsymbol{k} = (k_1, k_2, \cdots, k_n)^T$, $B = W - \boldsymbol{k}^T\boldsymbol{k}/\|W\|$, and $PL = (pl_{ik})_{n\times c}$, then Eq. (8) can be rewritten as:

$$Q_e = \frac{\text{trace}(PL\ B\ PL^T)}{\|W\|}. \tag{9}$$

Q_e is a directly extension of the crisp modularity function (2). When the credal partition degrades into the hard one, Q_e is equal to Q_h.

3.2 Spectral mapping

White and Smyth [13] showed that optimizing the modularity measure Q can be reformulated as a spectral relaxation problem and proposed spectral clustering algorithms that seek to maximize Q. By eigendecomposing a related matrix, these methods can map graph data points into Euclidean space, the clustering problem on which space is of equivalence to that on the original graph.

Let $A = (a_{ij})_{n\times n}$ be the adjacent matrix of the graph G. The adjacency matrix for a weighted graph is given by the matrix whose element a_{ij} represents the weight w_{ij} connecting nodes i and j. The degree matrix $D = (d_{ii})$ is the diagonal matrix whose elements are the degrees of the nodes of G, *i.e.* $d_{ii} = \sum_{j=1}^{n} a_{ij}$. The eigenvectors of the transition matrix $\mathcal{M} = D^{-1}A$ are used.

Verma and Meila [14] and Zhang et al. [17] suggested to use the eigenvectors of a generalised eigensystem $Ax = \lambda Dx$, and pointed out that it is mathematically equivalent and numerically more stable than computing the eigenvectors of matrix $\mathcal{M}$ [14]. To partition the nodes of the graph into c communities, the top $c-1$ eigenvectors of the above eigensystem are used to map the graph data into points in the Euclidean space, where the traditional clustering methods, such as c-means (CM), FCM and ECM can be evoked.

3.3 Evidential community detection scheme

Let C be the upper bound of the number of communities. The evidential community detection scheme is displayed as follows:

S.1 Spectral mapping:
For $2 \leq c \leq C$, Find the top c generalized eigenvectors $E_c = [e_1, e_2, \cdots, e_c]$ of the eigensystem $Ax = \lambda Dx$, where A and D are the adjacent and the degree matrix respectively.

S.2 Evidential c-means:
For each value of c ($2 \leq c \leq C$), let $E_c = [e_2, \cdots, e_c]$. Use ECM to partition the n samples (each row of E_c is a sample data on the $c-1$ dimensional Euclidean space) into c classes. And we can get a credal partition M for the graph.

S.3 Choosing the number of communities:
Find the suitable number of clusters and the corresponding evidential partition scheme by maximizing the evidential modular function Q_e.

In the algorithm, C can be determined by the original graph. It is an empirical range of the community number of the network. If c is given, we can get a credal partition using the proposed method and then the evidential modularity can be derived. The modularity is a function of c and it should peak around the optimum value of c for the given network. As in ECM, the number of parameters to be optimized is exponential in the number of communities and linear in the number of nodes. When the number of communities is large, we can reduce the complexity by considering only a subclass of bbas with a limited number of focal sets [7].

4 Experimental results

To evaluate the proposed method in this paper, two real-world networks are discussed in this section. A comparison for the detected communities by credal, hard and fuzzy partitions is also illustrated to show the advantages of evidential community structure over others.

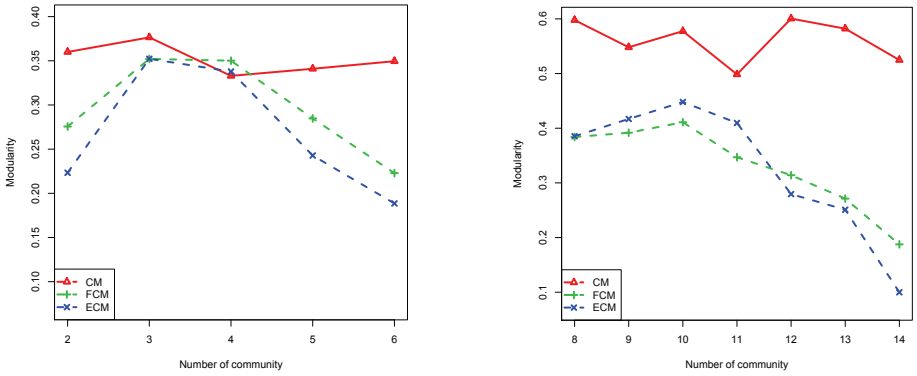
4.1 Zachary's Karate Club

The Zachary's Karate Club [16] is an undirected graph which consists of 34 vertices and 78 edges, describing the friendship between members of the club observed by Zachary in his two-year study. This club is visually divided into two parts, due to an incipient conflict between the president and instructor (see Fig. 2-a).

The modularity peaks around $c = 2$ or $c = 3$ as shown in Fig. 1-a. Let $c = 3$, the detected communities by CM, FCM and ECM are displayed in Fig. 2. As it can be seen, a small community separated from ω_1 is detected by all the approaches. The result by FCM shown here is got by partitioning nodes to the cluster with the highest membership. Zhang et al. [17] suggested to use a threshold λ to covert the fuzzy membership into the final community structure. For node i, let the fuzzy assignment to its communities be $\mu_{ij}, j = 1, 2, \cdots, c$. Node i is regarded as a member of multiple communities ω_k with $\mu_{ik} > \lambda$. But there is no criterion for determining the appropriate λ. However, in ECM we can directly get the imprecise classes indicating our uncertainty on the actual cluster of some nodes by hard credal partitions [7].

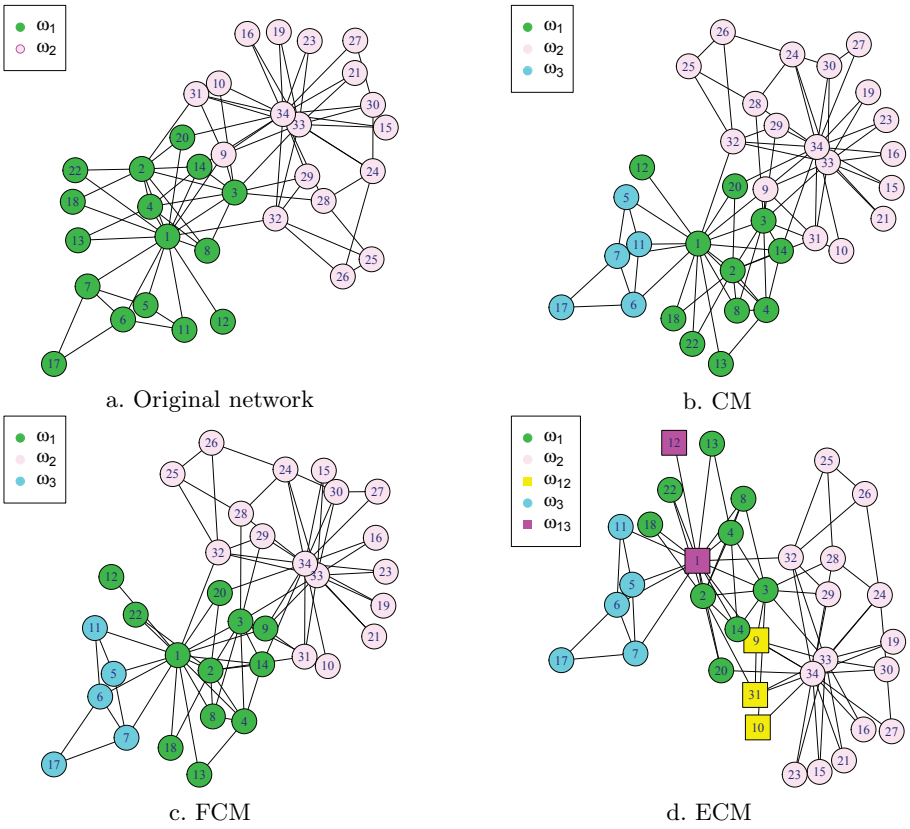
As we can see in Fig. 2-c, for ECM, node 1,9,10,12,31 belong to two clusters at the same time. This is coincident with the conclusion in [17] apart from the fact that a significant high membership value is given to ω_1 for node 12 by FCM. Actually, the case that node 12 is clustered into $\omega_{12} \triangleq \{\omega_1, \omega_2\}$ seems reasonable when the special behavior of this node is considered. The person 12 has no contact with others except the instructor (node 1). Therefore, the most probable class of node 12 should be the same as that of node 1. It is counterintuitive if the person 12 is partitioned into either ω_1 or ω_2, as it has no relation with any member in these two communities at all. The credal partition can reflect the fact that ω_1 and ω_2 is indistinguishable to node 12, while the fuzzy method could not. Furthermore, the mass belief assigned to imprecise classes reflects our degree of uncertainty on the clusters of the included nodes. As illustrated in Fig. 3-b, the mass given to imprecise clusters for node 1 is larger than that to the other four nodes. This reflects our uncertain on node 1's community is largest. As node 1 is the instructor of the club, this fact seems reasonable.

Actually, the concept of credal partitions suggests different ways of summarizing data. For example, the data can be analysed in the form of fuzzy partition thanks to the pignistic probability transformation shown in Eq. (5). It is shown in Fig. 3-a pignistic probabilities play the same role as fuzzy membership. A crisp partition can then be easily obtained by partitioning each node to the community with the highest pignistic probability. In this sense, the proposed method could be regarded as a general model of hard and fuzzy community detection approaches.



a. Karate Club network b. American football network

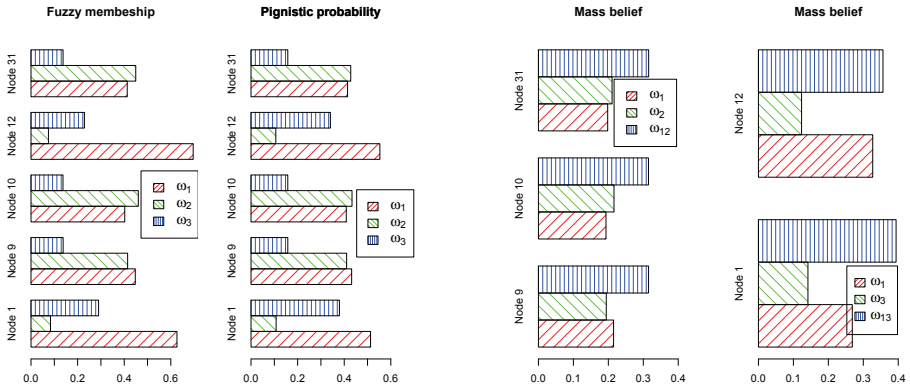
Fig. 1. Modularity values with community numbers.



a. Original network b. CM

c. FCM d. ECM

Fig. 2. Karate Club.



a. Probability membership b. Mass belief

Fig. 3. Clustering results of Karate club network.

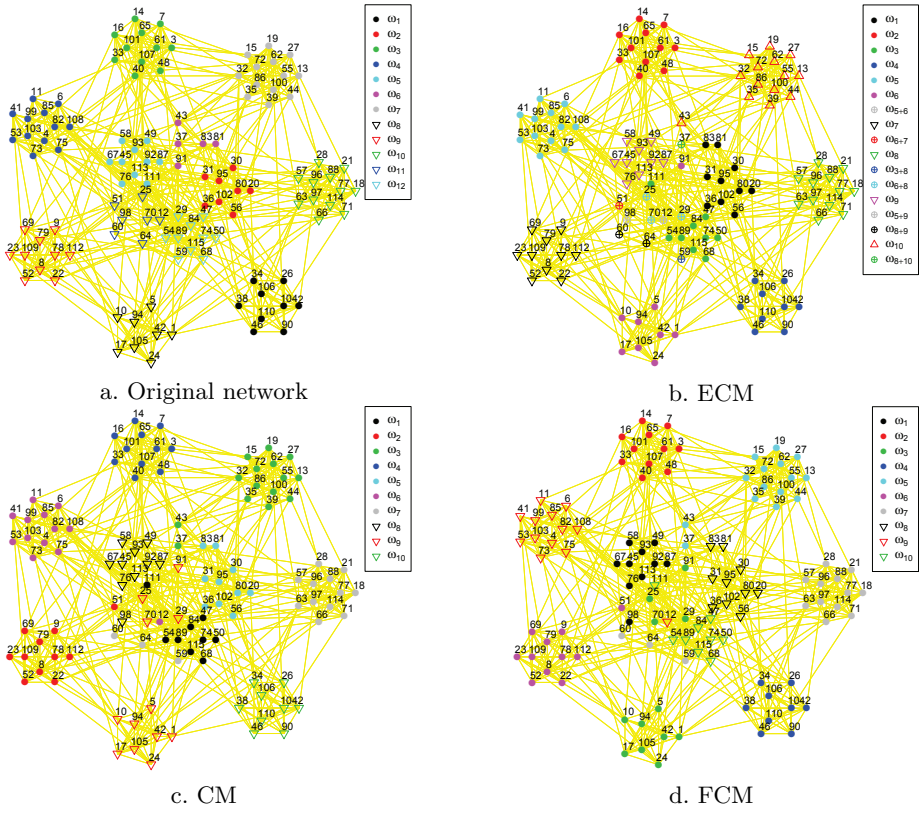
4.2 American football network

The network we investigate in this experiment is the world of American college football games between Division IA colleges during regular season Fall 2000 [4]. The vertices in the network represent 115 teams, while the links denote 613 regular-season games between the two teams they connect. The teams are divided into 12 conferences containing around 8-12 teams each and generally games are more frequent between members from the same conference than between those from different conferences.

In ECM, the number of parameters to be optimized is exponential in the number of clusters [7]. For the number of class larger than 10, calculations are not tractable. But we can consider only a subclass with a limited number of focal sets [7]. In this example, we constrain the focal sets to be composed of at most two classes (except Ω). Fig. 1-b shows how the modularity varies with the number of communities. For credal partitions, the peak is at $c = 10$. This is consensus with the original network (shown in Fig. 4-a) composed of 10 large communities (more than 8 members) and 2 small communities (8 members or less than 8 members). Set $c = 10$ in ECM, we can find all the ten large communities, eight of which are exactly detected. For the nodes in small communities, ECM partitions most of them into imprecise classes. As there are more than 10 communities in this network, we use ω_{i+j} to denote the imprecise communities instead of ω_{ij} in the figures related to this experiment to obviate misunderstanding.

For hard partitions, nodes in small communities are simply partitioned into their "closest" detected cluster, which will certainly result in a loss of accuracy for the final results. Credal partitions make cautious decisions by clustering nodes which we are uncertain into imprecise communities. The introduced imprecise clusters can avoid the risk to group a node into a specific class without strong belief. In other words, a data pair can be clustered into the same specific group

only when we are quite confident and thus the misclassification rate will be reduced.



a. Original network b. ECM

c. CM d. FCM

Fig. 4. American football network.

5 Conclusion

In this paper, a new community detection approach combing the evidential modularity, spectral mapping and evidential c-means is presented to identify the overlapping graph structure in complex networks. Although many overlapping community-detection algorithms have been developed before, most of them are based on fuzzy partitions. Credal partitions, in the frame of belief function theory, have many advantages compared with fuzzy ones and enable us to have a better insight into the data structure. As shown in the experimental results for two networks in the real world, credal partitions can reflect our degree of uncertain more intuitively. Actually, the credal partition is an extension of both hard and fuzzy ones, thus there is no doubt that more rich information of the graph structure could be available from the detected structure by the method

proposed here. We expect that the evidential clustering approaches will be employed with promising results in the detection of overlapping communities in complex networks with practical significance.

References

1. Costa, L.d.F., Oliveira Jr, O.N., Travieso, G., Rodrigues, F.A., Villas Boas, P.R., Antiqueira, L., Viana, M.P., Correa Rocha, L.E.: Analyzing and modeling real-world phenomena with complex networks: a survey of applications. Advances in Physics 60(3), 329–412 (2011)
2. Denœux, T., Masson, M.H.: Evclus: evidential clustering of proximity data. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on 34(1), 95–109 (2004)
3. Fortunato, S.: Community detection in graphs. Physics Reports 486(3), 75–174 (2010)
4. Girvan, M., Newman, M.E.: Community structure in social and biological networks. Proceedings of the National Academy of Sciences 99(12), 7821–7826 (2002)
5. Havens, T., Bezdek, J., Leckie, C., Ramamohanarao, K., Palaniswami, M.: A soft modularity function for detecting fuzzy communities in social networks. Fuzzy Systems, IEEE Transactions on 21(6), 1170–1175 (2013)
6. Havens, T.C., Bezdek, J.C., Leckie, C., Chan, J., Liu, W., Bailey, J., Ramamohanarao, K., Palaniswami, M.: Clustering and visualization of fuzzy communities in social networks. In: Fuzzy Systems (FUZZ), 2013 IEEE International Conference on. pp. 1–7. IEEE (2013)
7. Masson, M.H., Denoeux, T.: Ecm: An evidential version of the fuzzy c-means algorithm. Pattern Recognition 41(4), 1384–1397 (2008)
8. Nepusz, T., Petróczi, A., Négyessy, L., Bazsó, F.: Fuzzy communities and the concept of bridgeness in complex networks. Physical Review E 77(1), 016107 (2008)
9. Newman, M.E.: Fast algorithm for detecting community structure in networks. Physical review E 69(6), 066133 (2004)
10. Newman, M.E., Girvan, M.: Finding and evaluating community structure in networks. Physical review E 69(2), 026113 (2004)
11. Palla, G., Derényi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. Nature 435(7043), 814–818 (2005)
12. Smets, P., Kennes, R.: The transferable belief model. Artificial Intelligence 66(2), 191 – 234 (1994)
13. Smyth, S., White, S.: A spectral clustering approach to finding communities in graphs. In: Proceedings of the 5th SIAM International Conference on Data Mining. pp. 76–84 (2005)
14. Verma, D., Meila, M.: A comparison of spectral clustering algorithms. Tech. rep., UW CSE (2003)
15. Wang, X., Jiao, L., Wu, J.: Adjusting from disjoint to overlapping community detection of complex networks. Physica A: Statistical Mechanics and its Applications 388(24), 5045–5056 (2009)
16. Zachary, W.W.: An information flow model for conflict and fission in small groups. Journal of anthropological research pp. 452–473 (1977)
17. Zhang, S., Wang, R.S., Zhang, X.S.: Identification of overlapping community structure in complex networks using fuzzy c-means clustering. Physica A: Statistical Mechanics and its Applications 374(1), 483–490 (2007)