Physica A 490 (2018) 24–37

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

A framework for detecting communities of unbalanced sizes in networks

Krista Rizman Žalik *, Borut Žalik

University of Maribor, Faculty of Electrical Engineering and Computer Science, Slovenia

HIGHLIGHTS

- We propose a local unsupervised network cluster framework that works well also for communities of different densities and/or sizes.
- The proposed community detection algorithm integrates more different measures that define good communities.
- The proposed algorithm is a fast local expansion algorithm for uncovering communities of different sizes and densities.
- It reveals rich information on input networks.

ARTICLE INFO

Article history:
Received 25 January 2017
Received in revised form 6 July 2017
Available online 7 August 2017

Keywords:
Complex networks
Social networks
Network analysis
Community detection

ABSTRACT

Community detection in large networks has been a focus of recent research in many of fields, including biology, physics, social sciences, and computer science. Most community detection methods partition the entire network into communities, groups of nodes that have many connections within communities and few connections between them and do not identify different roles that nodes can have in communities. We propose a community detection model that integrates more different measures that can fast identify communities of different sizes and densities. We use node degree centrality, strong similarity with one node from community, maximal similarity of node to community, compactness of communities and separation between communities. Each measure has its own strength and weakness. Thus, combining different measures can benefit from the strengths of each one and eliminate encountered problems of using an individual measure. We present a fast local expansion algorithm for uncovering communities of different sizes and densities and reveals rich information on input networks. Experimental results show that the proposed algorithm is better or as effective as the other community detection algorithms for both real-world and synthetic networks while it requires less time.

1. Introduction

Many huge complex systems such as biological and chemical networks or large social and communication systems take the form of networks, sets of nodes joined together in pairs by edges [1]. An important problem in network analysis is discovering communities that are also named groups, modules, partitions or communities [2]. Network systems consist of several communities that have much fewer connections to the rest of the network than inside the same community. Networks can be represented by graphs and communities are subgraphs. Community detection problem is to identify a set of meaningful dense subgraphs in a given graph.

* Corresponding author.
E-mail addresses: krista.zalik@um.si (K.R. Žalik), borut.zalik@um.si (B. Žalik).

http://dx.doi.org/10.1016/j.physa.2017.07.028
0378-4371/© 2017 Elsevier B.V. All rights reserved.

A graph is mathematical representation of a network. A graph $G(V, E)$ has a set of nodes $V = (v_1, v_2, \ldots, v_n)$ that models objects and a set of edges $E = (e_1, e_2, \ldots, e_m)$ which connect pairs of nodes that interact in complex systems or they have just enough similar attributes, or they are connected as in citation networks or in world wide web pages. Individual two nodes can be connected with an edge that describes a certain degree of similarity between them. Communities are subgraphs defined as groups of densely interconnected nodes that are only sparsely connected to the rest of the network. Community detection methods find the partition $P = C_1 \bigcup C_2 \bigcup \cdot \bigcup C_k$ where $C_1, C_2, \ldots C_k$ are set of nodes called clusters or communities that can be disjoint or can overlap. Graph partition is NP-hard and different community detection methods have been proposed that seek to identify natural groups of related nodes in networks [3].

The community detection problem is increasing in importance, because of increased interest in studying many real complex systems such as social networks or World Wide Web structures [4]. Identifying and knowing communities can have practical importance. Community members have more common properties among themselves than with non-members of a community and identification of community structure can help in analyzing of functionality of networks and may discover not only similarity structure that is hidden, but also internal organization and how systems work [5].

In this paper, we use more measures in the process of forming communities. In contrast to many existing methods, a key feature is the ability to detect also small and/or sparse communities. One community is extracted at a time, while the remainder of the network has no influence on the forming of a community. So the extraction process allows the remainder of network to contain any number of weakly or tightly connected communities.

We propose a community detection algorithm that bases on the integration of more different measures that can fast identify communities of different sizes and densities. We use node degree centrality, strong similarity with one node from community, maximal similarity of node to community and separation measures. Each measure has its own strength and weakness. Thus, combining different measures can benefit from the strengths of each one and eliminate encountered problems of using an individual measure.

The rest of the paper is structured as follows. The next section contains a review of related work. Section 3 defines community detection model and Section 4 describes the proposed framework and outlines the proposed algorithm. Section 5 presents experimental results and compares the proposed methods using synthetic and real-world social networks. The final section contains conclusions.

2. Community detection methods

The recent research has given various methods for solving the community detection problem from different aspects. Density-based methods extract a large number of subgraphs with enough high density [6,7]. These methods require an input parameter—some density threshold. The other community detection methods define some conditions that have to be satisfied by communities. One such condition are cliques, subgraphs in which each pair of nodes has to be connected and k-cliques, subgraphs where each node is connected with k nodes from subgraph employed by Cfinder method with time complexity $O(\exp(n))$ [8].

Partition based methods partition the network to minimize interaction between parts. They evaluate the resulted partitions for a specified heuristic. They are divisive or agglomerative. Girvan and Newman [9] proposed hierarchical divisive heuristic in which edge with the largest betweenness is iteratively removed. The proposed divisive algorithm detects all edges that connect nodes from different communities and removes them. At the beginning, community is the whole input dataset. Each time a community is split into two, connected components correspond to two new communities. The betweenness centrality algorithm has time complexity $O(n^2m)$, where n is the number of nodes and m is the number of edges. Recently more betweenness centrality measures have been proposed [10]. Instead of betweenness centrality a random walk based algorithm for community detection has been proposed [11]. The communities are detected from maps of random walks referenced as Infomaps. Nodes central to a community are more often visited in a random walk. Time complexity of Infomap is $O(m)$ where m is the number of edges.

Agglomerative methods group nodes into communities iteratively, starting from communities consisting from only one node. The main drawback of many agglomerative methods is that they require the number of partitions as an input parameter before the start of community identification process. But it is difficult to know the number of communities in most networks a priori. Many methods avoid this problem by using some quality measure such as modularity [12]. The second drawback of many agglomerative community detection methods is that they identify balanced communities. Most often used quality measure modularity has a resolution limit while it cannot detect communities with sizes smaller than a threshold, which depends on the network size [13]. It has been shown that any two interconnected modules are merged if the number of edges inside each of them is not greater than $\sqrt{\frac{|E|}{2}}$, where $|E|$ is the number of edges in the whole network. The third drawback is that quality measures require the complete information about the whole network. Since obtaining complete information about networks is unrealistic nowadays, when networks are becoming larger and larger, there is a growing emphasis on local community detection [14–17].

Spectral clustering models [18] can be used for community detection methods, which apply spectral analysis to obtain the cut minimization. Label propagation models mainly use the neighbor information of each node to determine its label and do not need any prior knowledge of community structure. One of the fastest algorithms proposed to date is the label propagation algorithm LPA was proposed by Raghavan [19]. Each label of the neighbor having the most common neighbors propagate. Label propagation community detection provides fast uncovering of communities, but it uncovers different communities that depend on label propagation in the case when multiple neighbor labels are equally frequent. In addition, the convergence speed and clustering effectiveness of the algorithm are very sensitive to the update order of label information. Some improvements of label propagation to solve this problem have also been proposed [20].

Another important feature of complex networks is the hierarchical community structure. Lancichinetti et al. [21] proposed a hierarchical and overlapping community detection method LFM based on a local fitness measure, which generates multiple communities to show hierarchies of the network by randomizing the starting nodes and varying a resolution parameter.

Some proposed community detection methods are based on different specified conditions that have to be satisfied by communities. One such condition are cliques, in which each pair of nodes must be connected by an edge. This criteria is too difficult to compute for large networks. Radicchi et al. [22] proposed two well known criteria: community in weak sense and community in strong sense and a divisive hierarchical method where edges are iteratively removed using edge coefficient with time complexity $O(n^2)$. They defined a subgraph of nodes of network to form a community in a strong sense if the number of neighbors of each node within this subgraph is larger than the number of neighbors outside this subgraph. And a set of nodes forms a community in a weak sense when the number of communities' neighbors within the community is greater than the number of neighbors outside this community.

The most simple agglomerative algorithm exploiting the strong community definition can start forming a new community at the node with the greatest degree. Then all neighbor nodes are added to the community and those nodes with connections inside community smaller than toward other communities are removed from community. This is very fast algorithm but it does not identify some small communities. It identifies two communities in the dolphin social network (described in Section 5.3), but it identifies only one community in Zackary karate club (Section 5.2). One possibility to improve this simple approach is to use and optimize an objective function. An recent example of such approach that gives also rich information on networks together with outliers and overlapping nodes is in Ref. [23]. We improve this simple approach with adding additional measures used for expansion.

3. The proposed community detection model

3.1. Community detection problem

Each complex network can be represented by an undirected graph and graph concepts can be used to understand complex network systems [24]. $G = (V, E)$ is an undirected graph with a set of nodes $V = (v_1, v_2, \ldots, v_n)$ that models objects and a set of m edges connecting pairs of nodes that are similar. Edges connect pairs of nodes. A graph $G(V, E)$ can be represented by $n * n$ adjacency matrix $A = \left[A_{ij}\right]$ where $A_{ij} = 1$, if there exist a link between node i and node j and $A_{ij} = 0$ otherwise. There are many possible divisions of nodes into more subgraphs. The community detection algorithms partition a graph into k subgraphs called communities, such that the quality of partition is the greatest. Reliable algorithms are supposed to identify good partitions with the greatest quality. Communities can be described by $k * k$ matrix $C = \left[C_{ij}\right]$ where $C_{ij} = 1$, if node i is assigned to community j and $C_{ij} = 0$ otherwise. But there is no common agreed quality definition. In order to distinguish between good and bad partitions, it would be useful to require that partitions satisfy a set of basic properties.

3.1.1. Basic definitions

In the graph $G = (V, E)$, each node i is a direct neighbor of node j if i and j are connected by an edge $e_{ij} \in E$. The degree deg_i of node i is the number of edges that connect node with its direct neighbor nodes:

$$deg_i = \sum_{j \in G} A_{i,j}. \tag{1}$$

Communities are groups of nodes similar to each other. One of the most common node similarity measure is structural similarity in which, node similarity is estimated from the local connections and local density structure of a network. Two nodes are more similar if they are connected by an edge than two unconnected nodes. Two connected nodes are more similar if they share more common direct neighbors.

Similarity between each two nodes is described in matrix S with elements $S_{i,j}$ which are the number of common direct neighbors of nodes i and j:

$$S_{i,j} = A_{i,j} \cdot A_i^T \cdot A_j. \tag{2}$$

Similarity is symmetric: $S_{i,j} = S_{j,i}$. Beside degree of node we can define similarity degree of node. Similarity degree of node i is the sum of similarities to all direct neighbor nodes:

$$simdeg_i = \sum_{j \in G} S_{i,j}. \tag{3}$$

The node with the greatest similarity degree is chosen as a starting node of new community.

We can define also similarity degree of node to another node. Similarity degree $simdeg_{ij}$ of node i to another node j is the fraction of similarity and node degree:

$$simdeg_{i,j} = \frac{S_{i,j}}{deg_i}. \tag{4}$$

If similarity degree of a node i from any other community of partition except l to another node j from community l is enough high, we can say that the node i is a possible member of the community l.

All nodes (i), that do not belong to community l with the similarity degree $simdeg_{i,j}$ with one node j from community l greater then given thresholds α and $simdeg_{j,i}$ greater than β, form a set of possible nodes PV_l of community l (see Eq. (10)). The value 0.5 for both thresholds gives good result. The best values we determined experimentally are $\alpha = 0.4$ and $\beta = 0.2$ and we call them normal resolution parameters.

$$PV_l = \left\{ i \mid simdeg_{i,j} > \alpha \; AND \; simdeg_{j,i} > \beta, i \notin l, j \in l \right\}. \tag{5}$$

For each community l we can define extended community l', which consists of all nodes assigned to community l and all possible nodes of community l.

3.1.2. Compatibility of a node to a community

Similarity of node $i \in G$ to community l is the sum of similarities of a node i to all nodes from community l.

$$simC_{i,l} = S_i^T \cdot C_l. \tag{6}$$

And similarity to all other communities than l is denoted with $simC_{i,\bar{l}}$. Sum of similarities of all nodes of community l is $simC_l$.

$$simC_l = \sum_{i \in C_l} simC_{i,l}. \tag{7}$$

A node has different similarities to different communities forming partition P and one greatest similarity value.

$$simC_i^{\max} = \max(simC_{i,l'}), \quad l' \in P. \tag{8}$$

Greater are values of similarity of each node to a community greater are values of cohesion of nodes in all communities. Cohesion refers to the degree to which the elements of a community belong together. Cohesion of nodes is higher near the center of cluster. We call all nodes in the center core nodes and other nodes of community are peripheral.

The node i assigned to community l is said to be core node, if similarity to other core nodes of the community is greater than some ratio of similarity degree and degree of center (Eq. (12)). Center is the node of community with the greatest similarity degree.

$$Core = \left\{ i \mid simC_{i,l} > \gamma * \frac{simdeg_{center}}{deg_{center}}; \; simC_{i,\bar{l}} < simC_{i,l}; \, i, center \in l \right\}. \tag{9}$$

All nodes that are not core nodes are peripheral. In all figures, core nodes of the same community are colored with the same color, while peripheral nodes are white (see Fig. 4). Each node has community label and then follows node label. All nodes of the same community have the same community label.

We define a new measure called compatibility. Nodes that belong to community to which they have the greatest similarity are called compatible.

A node i is compatible to community l when:

$$simC_{i,l} = simC_i^{\max}. \tag{10}$$

The node i is always compatible with community l' when similarity to community l' is greater than half of similarity degree of node i :

$$simC_{i,l'} \geq \frac{simdeg_i}{2}. \tag{11}$$

A node with maximal similarity degree is chosen as a community center. And then a fast expansion algorithm can add each neighbor node to growing community using the upper definition compatibility of a node to community. We call this single neighbor addition or merging. More neighbor nodes can be added in each iteration which makes the community detection process faster. This merging will not guarantee the increase the modularity value. We obtain a set of connected communities, we call them preliminary communities. Some pairs of these connected preliminary communities can be merged together that will result in an increase in the modularity value. Consequently, we use single merging and merging of preliminary communities.

3.1.3. Coupling of communities and stability of community

Coupling is the degree to which two communities are similar to each other. The good community is well separated from the rest of the network—from its neighborhood. Good partition contains communities with low coupling values.

Coupling of two communities l and u is the sum of similarities between all nodes of both communities:

$$coupling_{l,u} = \sum_{i \in l, j \in u} S_{i,j}. \tag{12}$$

Lower coupling values denote more separable communities, and so higher quality communities and partitions. We can find maximal coupling value for each community l. For a given partition P the minimal separability of community l corresponds to maximal value of coupling of community l to any other community in P . We denote it $coupling_l^{\max}$.

$$coupling_l^{\max} = max(coupling_{l,u}), \quad u \in P. \tag{13}$$

We define a measure called *stability*. It is based on the idea that community is a set of objects with strong connections among them and few interactions with the nodes outside of the community.

For a given partition P the community l is stable ($stability_l > 0$) when maximal coupling value is smaller than half of the internal similarity of community.

$$stability_l = \frac{X_l}{simdeg_l} \tag{14}$$

$$X_l = \begin{cases} 0 & \text{if } simC_l - 2 * coupling_l^{\max} \le 0 \\ simC_l - 2 * coupling_l^{\max} & \text{if } simC_l - 2 * coupling_l^{\max} > 0. \end{cases} \tag{15}$$

Stability of partition $stability_P$ is a measure of the strength of division of a network into communities. It can be used as optimization criteria and as the quality function QM:

$$QM = stability_P = \sum_{l=1}^{k} stability_l. \tag{16}$$

Maximal value of the quality function QM means dense connection between the nodes within communities and sparse connections between nodes of different communities.

We obtain some connected preliminary communities by expansion algorithm using maximal similarity degree for choosing a node as a new community center and then using definition compatibility of node to community. Each two preliminary communities C_i and C_j can be merged into new community C_l so that the merge maximizes the increase of optimization function $\triangle QM$:

$$QM = \sum_{l=1}^{k} \frac{e_{ll} - 2 * coupling_l^{\max}}{a_l}; \quad e_{ll} = simC_l;\ a_l = \sum_{i \in l} simdeg_i \tag{17}$$

$$\triangle QM = Q(C_l) - (Q(i) + Q(j)) \tag{18}$$

$$\triangle QM = \frac{e_{ll} - 2 * coupling_l^{\max}}{a_l} - \left(\frac{e_{ii} - 2 * coupling_i^{\max}}{a_i} + \frac{e_{jj} - 2 * coupling_j^{\max}}{a_j} \right) \tag{19}$$

Similarity of nodes in community l is sum of similarities of nodes in both communities i and j and the sum of similarity of the edges between both communities E_{ij}. Because the community l contains all nodes from the community i and j, a_l is the sum of a_i and a_j.

$$e_{ll} = e_{ii} + e_{jj} + E_{ij}; \qquad a_l = a_i + a_j \tag{20}$$

4. The community identification framework

4.1. Basic idea

One community can be detected at a time. The community detection process starts with the detection of a community around the node of the densest area. Nodes are sorted according to similarity degree and put into set of free nodes. The node with the greatest similarity degree (Eq. (5)) is taken from set of free nodes and chosen as a starting node. The neighbor nodes of the forming community are added to community if they are connected strongly enough to the community (Eq. (10)) extended with all possible nodes of the community (Eq. (6)) . Then the next free node with the highest degree is chosen and the second community is started to be created. Detected communities are called preliminary communities. They can be unstable and they have to be merged in the final stable communities. Each community is merged with this direct neighbor community that cause the maximal and positive increase of quality function that maximizes stability.

4.2. Algorithm

The proposed algorithm uses multiple measures in expanding process and we named it multiple measure expansion algorithm—MME. It consists from three phases: an initialization phase, a community discovering step and a finalization phase. In the initialization step we calculate similarity (Eq. (3)) of each two direct neighbor nodes i and j. Then the similarity degree $simdeg_i$ for each node i is calculated (Eq. (4)). At the beginning of community detection process all nodes are free and all they are placed in the set F and sorted according to the calculated similarity degrees in descending order.

The second phase is the preliminary community discovery step. The first node is taken from the set of free nodes F and used as a new community center.

All steps of the proposed algorithm are illustrated using the Zackary karate club network (described in Section 5.2). It is a real dataset consisting of 34 nodes. The node with the highest similarity degree in Zackary karate club network is the node with a label 1 (shown in Fig. 1).

We then create a set N from all neighbors of nodes which have already been assigned to community. In our case we add to the set N all neighbors of node 1. All neighbor nodes that are connected enough strongly with at least one node from a community are identified. They are called possible candidates PV. All neighbor nodes from N that are connected with nodes in community and with possible community candidates PV more strongly than to the rest network are added to a community and removed from the set F. All nodes that are connected to the center and other core nodes more than threshold are identified as core nodes (Eq. (9)). The community with starting node 1 of the karate dataset is shown in Fig. 2.

Then the next community is started to be formed with the first taken node from the set F as a center for new community (see Fig. 2 and for next steps Fig. 3). Forming of new communities is stopped when the set of free nodes F is empty.

Then the finalization step begins. Based on the internal similarity and maximal coupling of each community, we are able to decide which communities are stable and significant (Eq. (16)), and which are not. Unstable communities are merged with their nearest neighbors. Each community is merged with this direct neighbor community that gives the maximal and positive increase of the quality function.

In the karate club preliminary community with only one node 29 is unstable. Joining community 29 and 34 increases the quality function QM (Eq. (18)) for 0.6, while joining community 29 with 32 increases the quality function QM for 0.4. Preliminary community 29 is merged with preliminary community labeled 34 (see Fig. 4). All joins of remaining preliminary communities gives negative increase of modularity function. Joining of new community 34 to community 23 increases the quality function QM for 0.2 and all other joins of community 34 with other communities gives negative increase of quality function QM. So the resulted partition contains communities 34, 1, and 7. But using Newman modularity community 29 is joined with community labeled 32 and so 4 communities are discovered for karate club friendship network. This partition has the modularity 0.41 that is the highest modularity among compared methods in Table 1.

The running time of the proposed community detection algorithm MME is not mainly consumed in selecting one neighboring node that maximally increases the quality function or with larger fitness value during the process of forming local communities such as the Clauset method or LFM. Over one iteration, more nodes can be added to the community depending on the intrinsic edge structure of the graph. This requires MME algorithm to perform less iterations of steps 2 and 3 than other local community detection methods such as the Clauset method or LFM. Table 1 shows the number of iterations for the datasets used in experiments together with the number of nodes and edges. It can be observed that algorithm MME needs less iterations than LFM and Clauset's method, while they all use internal and external edges for calculation of quality function.

4.2.1. Time complexity analysis

Time complexity analysis of the proposed algorithms for all three steps follows.

1. Initialization step: The number of common direct neighbors is counted and the similarity value assigned to each edge. Using the adjacency matrix data structure, checking neighboring of nodes can be done in $O(1)$ time and thus finding the number of common friends between nodes i and j connected with an $edge(i, j)$ can be done in $O(d_i)$ time where d_i is the number of direct neighbors of node i. d_i defines density of network and is weakly correlated with number of all nodes n. Because each node i has d_i neighbor edges, finding the weight of all neighbor edges of node i, has the time complexity of $O(d_i^2)$. Then we calculate similarity degree for all nodes with time complexity of $O(nd_i^2)$. We then put all similarity degrees of nodes in a set and then sort them based on their similarity degrees in descending order. Running time of this stage is mainly concerned with the sorting time of similarity degrees of all n nodes that can be performed with time complexity $O(n \ \log(n))$ using merge sort.

2. Preliminary community detection step: In this step we find preliminary communities. Node neighbors are identified in $O(1)$ time. For each direct neighbor node i the similarity to extended community with possible nodes C_l' can be calculated in $O(d_i)$ time. Each node compatible with community C_l' is assigned to community C_l in $O(1)$ time. We consider this single merging. Running time of this stage is $O(n * d_n * d_i)$ where d_n is the number of direct community neighbor nodes and it depends on the number of direct neighbors of nodes on the border of community.

What are the required number of iterations that the MME needs to find all the preliminary communities as they are. In each iteration more nodes can be added to community so the number of iteration is less than $n/2$ (see Table 2) and time complexity is $O(n)$. We identified all preliminary communities.

3. Preliminary community merging step: preliminary communities are merged into larger communities and thereby maximizing the objective function. In each iteration more preliminary communities can be merged. Assume that we get from

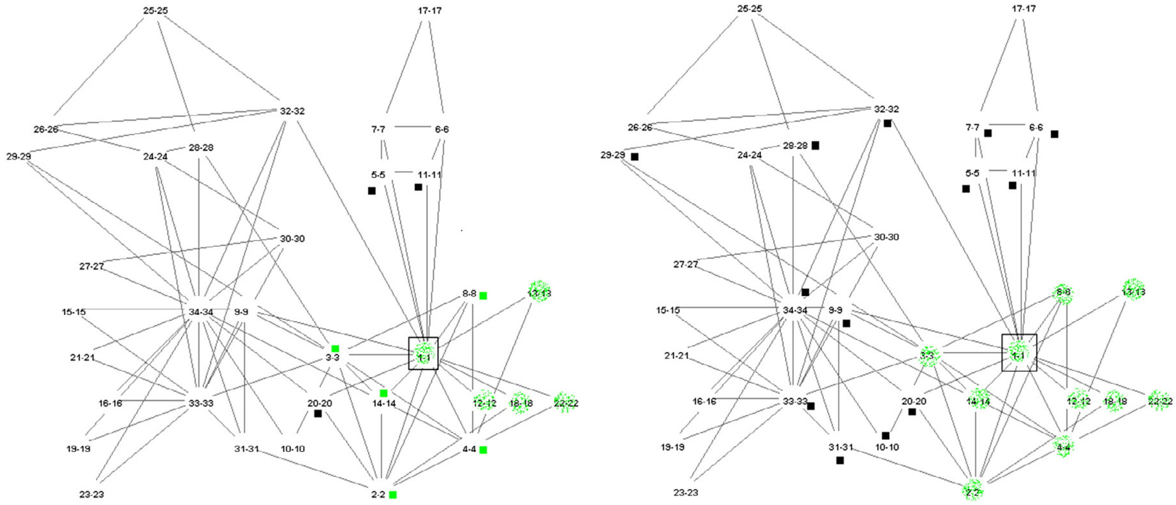

Fig. 1. The friendship network for Zackary's karate club for parameters $\alpha = 0.4$, $\beta = 0.2$, $\gamma = 3$: the first and the second step (from left to right) of building the community with the center node 1 (in a black square). Nodes added to community are shown in green circuits. All possible nodes are marked with green point and all other neighbors are marked with a black point. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

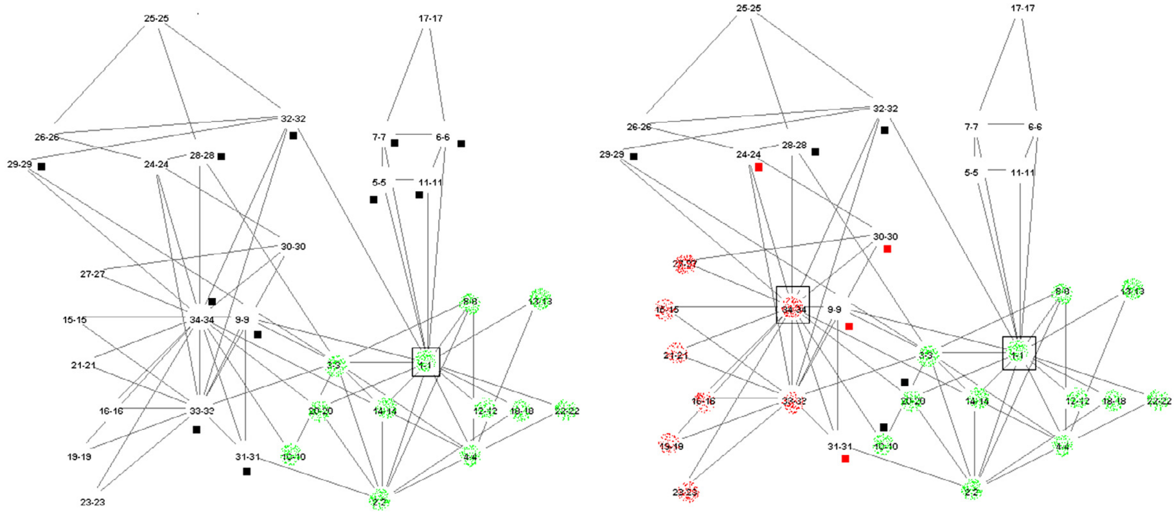

Fig. 2. The friendship network from Zackary's karate club $\alpha = 0.4$, $\beta = 0.2$, $\gamma = 3$: the third and fourth step (from left to right): the second community labeled with 34 is started to be build. Community center is in a black square. Nodes added to community are shown in circuits of the same color as community center. All possible nodes are marked with point of the same color as center and all other neighbors are marked with a black point. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the previous step k' preliminary communities each having only one element. Then we will need $\log n$ iterations to merge all communities to form one community including all nodes and each iteration can be implemented with time complexity $O(m)$. So, the time complexity of MME would be $O(m \ \log(n))$ on sparse graphs.

4.3. Multiresolution community detection

Multi-scale (called also multiresolution) community detection attempts to identify joins of the most relevant communities and so obtain multiresolution partitions. We can change two parameters (α , β Eq. (10)), that can increase or decrease the possibility of the node to be a possible member of a community. High values of thresholds decrease the possibility of a new node to become a member of a community and the result is a lot of small communities that are not stable. They are merged in the finalization step of MME algorithm and the result is a partition with less communities than obtained using normal parameters. For small parameters the possibility that neighbor nodes are possible candidates is higher and bigger communities are formed.

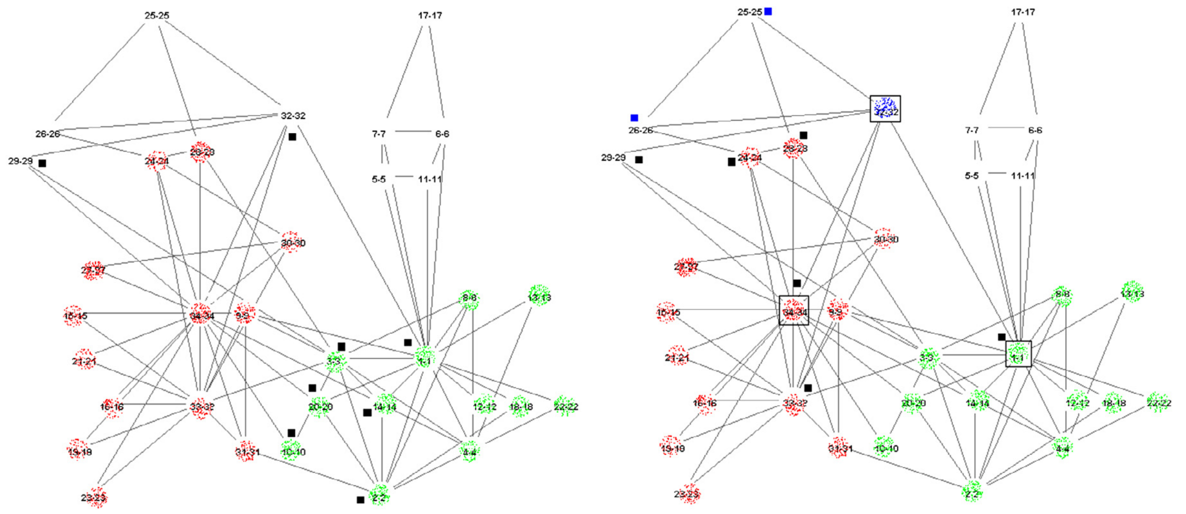

Fig. 3. The friendship network from Zackary's karate club $\alpha = 0.4$, $\beta = 0.2$, $\gamma = 3$: steps 5 and 6 (from left to right): the third preliminary community with the center node 32 is started to be build. Center is marked with a black square. Nodes added to community are shown in circuits of the same color as center. All possible nodes are marked with a point of the same color as center and all other neighbors are marked with a black point. One preliminary community is formed with only one node 29. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

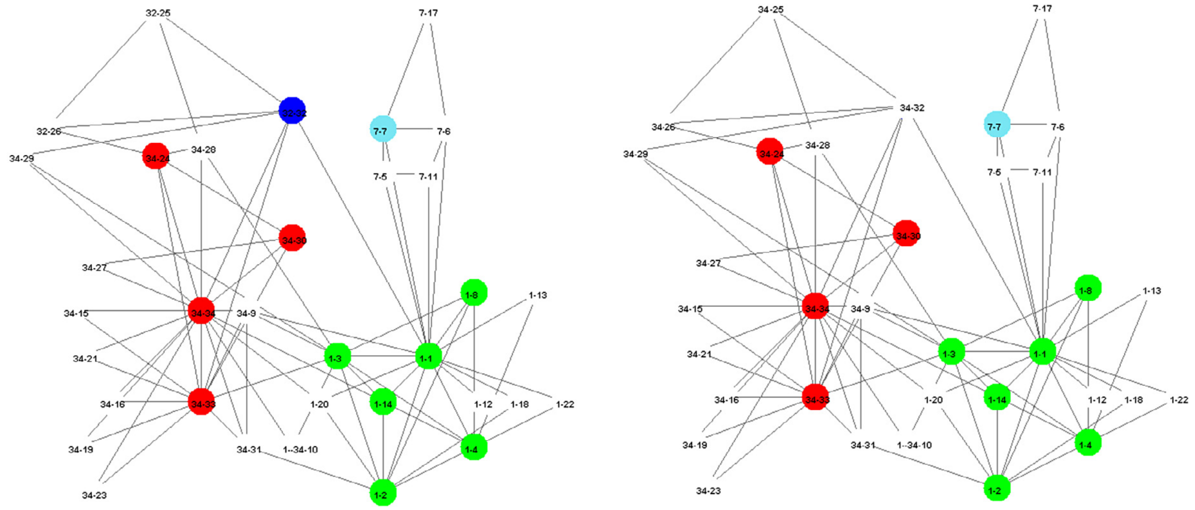

Fig. 4. The friendship network from Zackary's karate club $\alpha = 0.4$, $\beta = 0.2$, $\gamma = 3$: the final partition of Zackary karate club on the right after merging communities 34 and 32 shown in the left graph. The first number in each node is community label and the second is a node label. All nodes forming the same community have the same community number. Core nodes are colored while peripheral nodes are white. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5. Results

In this section we compare the results obtained by the proposed algorithm to find communities on some synthetic and some popular real-world datasets.

5.1. Artificial networks composed of cliques

To evaluate the proposed method for detecting communities of unbalanced size and prove that the method avoids the resolution limit problem we created some artificial networks composed of cliques.

Our example is a network of identical cliques connected by single edges: a ring of 5 cliques of 4 nodes connected with one link as illustrated in Fig. 5. The Newman modularity optimization method combines two or more cliques together when the number of cliques K is greater than square number of edges: $(K + K * n * (n - 1)/2)$, where n is the number of nodes in

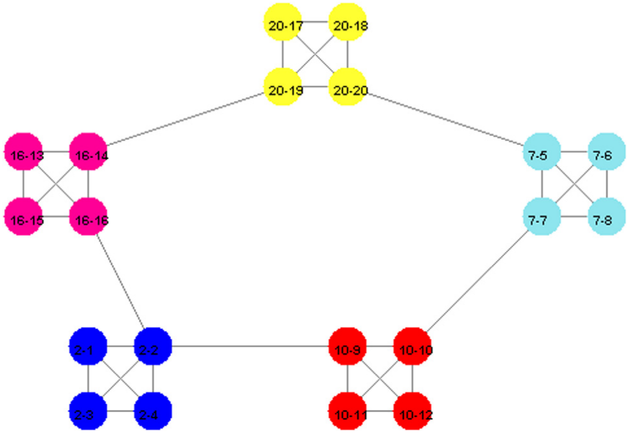

Fig. 5. Ring of cliques for normal parameters: $\alpha = 0.4$, $\beta = 0.2$, $\gamma = 3$.

each clique. For normal and other resolution parameters all 5 communities are identified. For resolution parameter greater than 2 each node forms a separate community, but they are unstable and merged then in 5 stable communities.

Beside artificial networks, we considered well-known real social networks used in the literature as a test bed for several models.

5.2. Zackary's karate club network

A first real-world dataset we considered is Zackary's karate club dataset [25]. 34 nodes correspond to the members of a karate club and edges correspond to connections between members inside the club. The club was broken into two groups to support the administrator and the instructor. For resolution parameters 0.4 and 0.2 (named normal) four preliminary communities and three communities in final partition are identified (see Fig. 4). Fig. 6 shows the two groups for resolution parameters $\alpha = 0.4$ and $\beta = 0.0$. The first has 18 nodes and is denoted by starting node 34 and the second community has 16 members and is denoted by starting node 1. The difference among partitions obtained by different community identification methods is a node 3. Node 3 is identified more times in the second group denoted with 1, than in the first denoted with 34. The proposed method puts node 3 in the second community with label 1. Node 10 is shared by both communities. For resolution parameters 0.5, 0.5 three big communities are identified and some small (see Fig. 7), which are then merged with big communities and the resulted 3 communities are shown in Fig. 8. The modularity of partition discovered by the proposed method for Zackary's karate club dataset for parameters 0.4 and 0.2 is better than modularities of other two considered methods (see Table 1).

5.3. Dolphin social network

Dolphin dataset [26] can be divided into two groups. The structure discovered by the proposed algorithm and resolution parameters α and β 0.2 and 0.0 is shown in Fig. 9. The first community of 21 dolphins is denoted by starting node 14 and the second containing 41 dolphins is denoted by starting node 15. For resolution parameters α and β 0.4 and 0.2 five communities are identified shown in Fig. 10. The modularity of partition discovered by the proposed method for dolphin dataset is better than modularities of other two considered methods (see Table 2).

5.4. Les miserables

Les Miserables is a co-appearance network of characters in the Victor Hugo's novel "Les Miserables" [27]. The dataset is very dense. But 7 communities are identified using normal resolution parameters as shown in Fig. 11, with starting nodes 1, 12, 24, 26, 27, 30 and 49. The communities can be described as follows: C1-Bishop Myriel and the characters he met, C12-the central group with Jean Valjean(12), C24-students and Fantine, C26-the evil innkeeper Thenardier, C27-the family of Marius, C30-protagonists of the Champmathieu affair, C49-Marius, Gavroche and the revolutionaries. The communities computed by the proposed model agree with the other methods and gives the best modularity between the compared methods in Table 2.

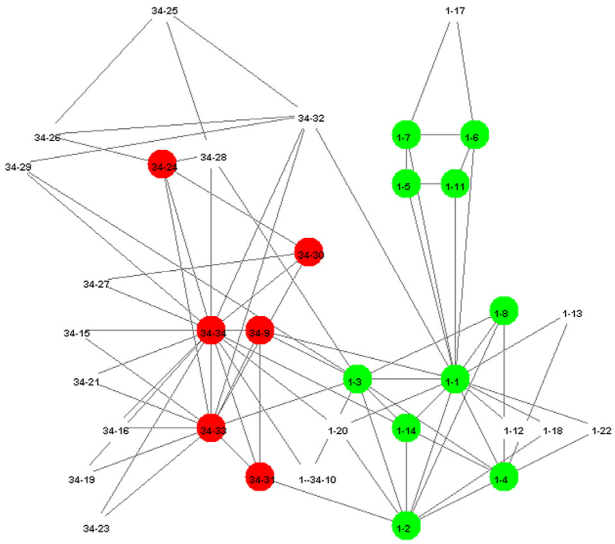

Fig. 6. The friendship network of Zackary's karate club: for parameters $\alpha = 0.4$, $\beta = 0.0$, $\gamma = 3$ two communities are formed with starting nodes 1 and 34. Core nodes of the same community are colored with the same color, while peripheral nodes are white. Each node has community label and then node label. All nodes of the same community have the same community label. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

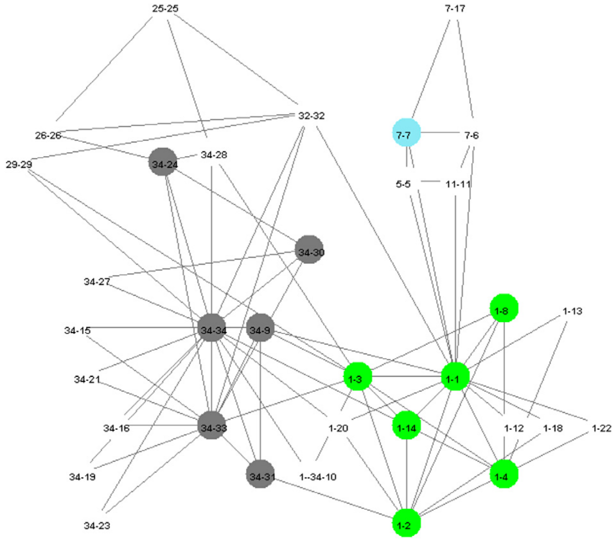

Fig. 7. The friendship network of Zackary's karate club: for resolution parameters $\alpha = 0.5$, $\beta = 0.5$, $\gamma = 3$ three big communities are formed and 4 communities with one node 25, 26, 29, 32 in community detection phase of MME algorithm. Core nodes of the same community are colored with the same color, while peripheral nodes are white. Each node has community label and then node label. All nodes of the same community have the same community label. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5.5. Political books

Books dataset is about frequent co-purchasing of books about US politics compiled by Valdis Krebs [28]. Our algorithm MME identifies two large communities and two small for normal resolution parameter shown in Fig. 12. Communities have 43, 37, 9 and 7 elements. Books in the largest community with 43 nodes represent conservative books and the community with 37 elements contain liberal books and two small communities represent neutral books. In [29], this network was decomposed into four communities and the partitioning is in accordance also to the representation of this network by Krebs. The modularity of resulted partition is better than molarities of other two considered methods (see Table 1).

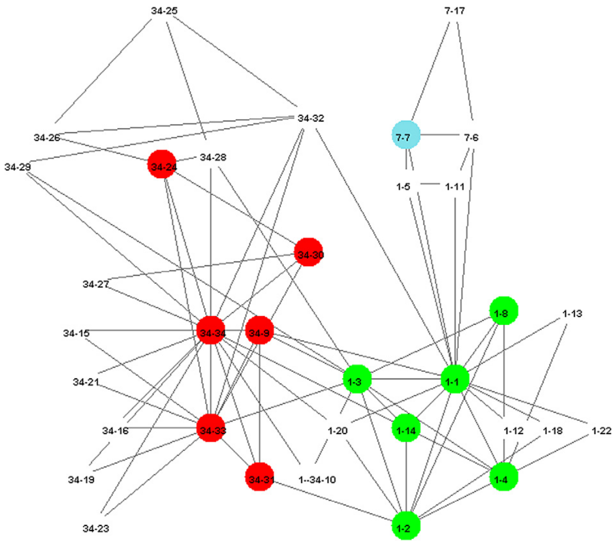

Fig. 8. The friendship network from Zackary's karate club: for resolution parameters $\alpha = 0.5$, $\beta = 0.5$, $\gamma = 3$ three big communities are formed (after finalization phase). Core nodes of the same community are colored with the same color, while peripheral nodes are white. Each node has community label and then node label. All nodes of the same community have the same community label. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

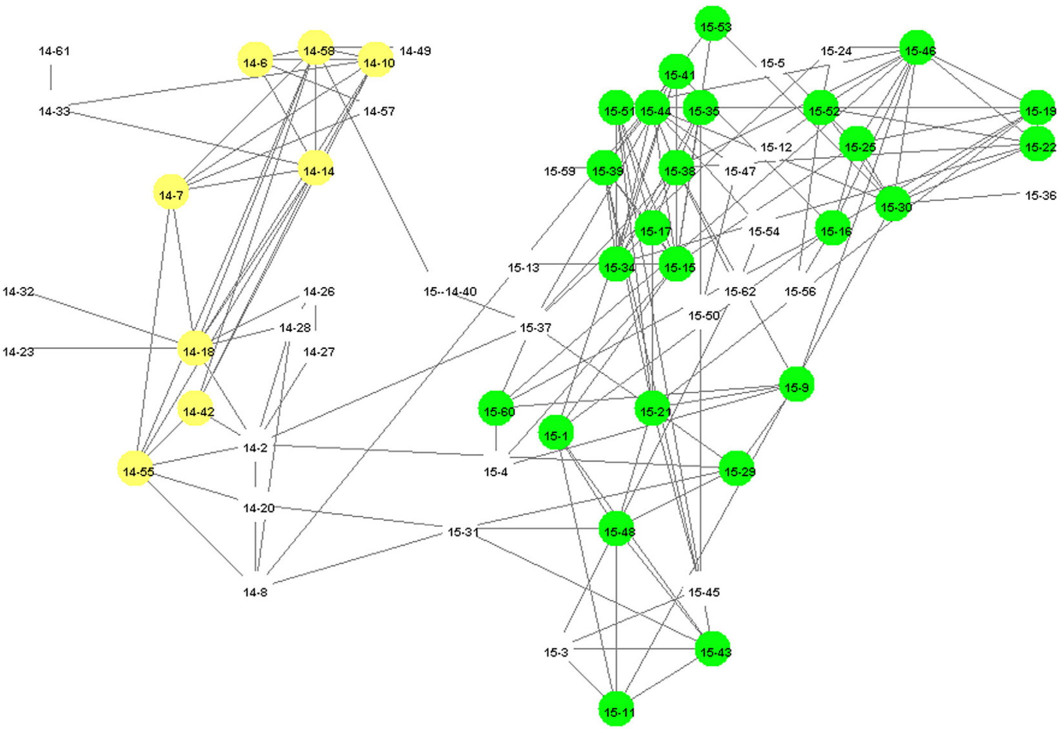

Fig. 9. The dolphin network. The two communities are formed with starting nodes 14 and 15 for parameters $\alpha = 0.2$, $\beta = 0.0$, $\gamma = 2$. Core nodes of the same community are colored with the same color, while peripheral nodes are white. Each node has community label and then node label. All nodes of the same community have the same community label. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5.6. Collage football network

We evaluated the proposed method in an example with more communities using U.S. college football network dataset [9]. The nodes of a network correspond to the teams, while the edges represent the games between any two teams. Teams in

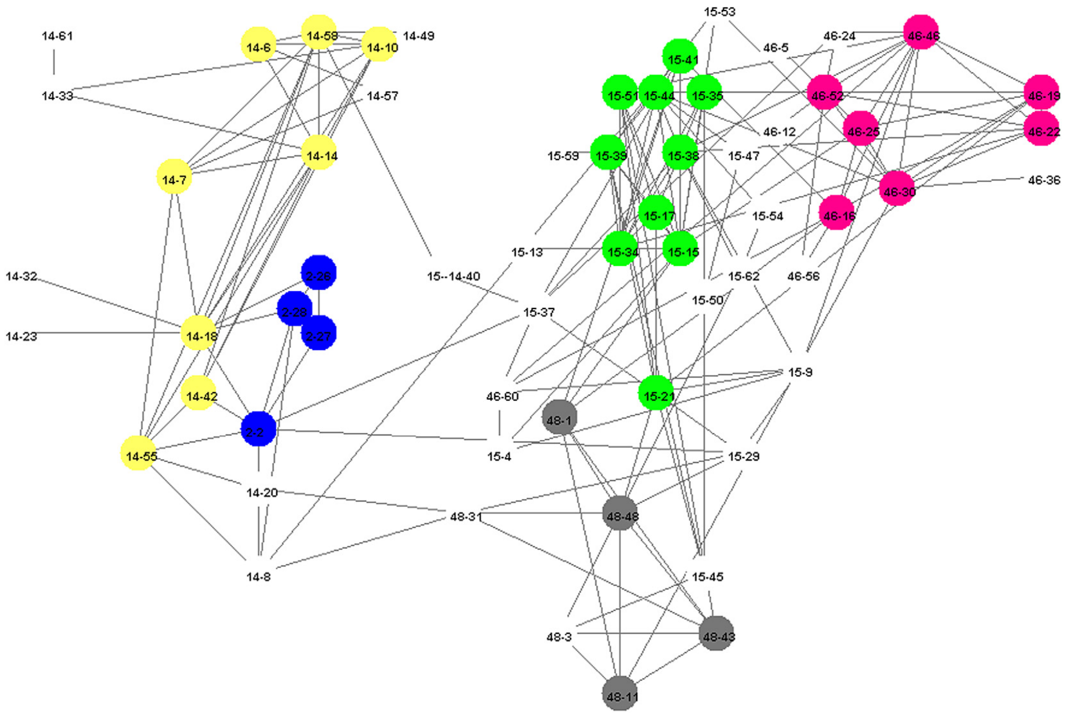

Fig. 10. The dolphin network. Five communities are formed for parameters $\alpha = 0.4$, $\beta = 0.0$, $\gamma = 2$ Core nodes of the same community are colored with the same color, while peripheral nodes are white. Each node has community label and then node label. All nodes of the same community have the same community label. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

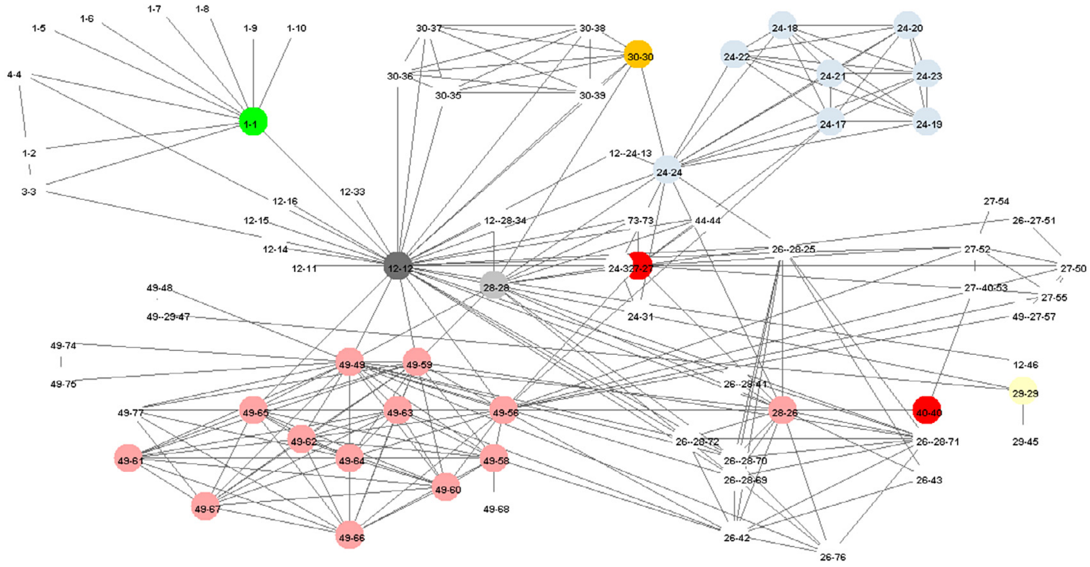

Fig. 11. Seven communities identified in Les Miserables dataset for parameters: $\alpha = 0.4$, $\beta = 0.3$, $\gamma = 8$. Core nodes of the same community are colored with the same color, while peripheral nodes are white. Each node has community label and then node label. All nodes of the same community have the same community label. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the same conference have more games among each other. One community corresponds to each conference. There are five independent teams (Utah State-90, Navy-80, Notre Dame-82, Connecticut-42 and Central Florida-36).

For normal resolution parameters the proposed algorithm identifies 12 communities shown in Fig. 13.

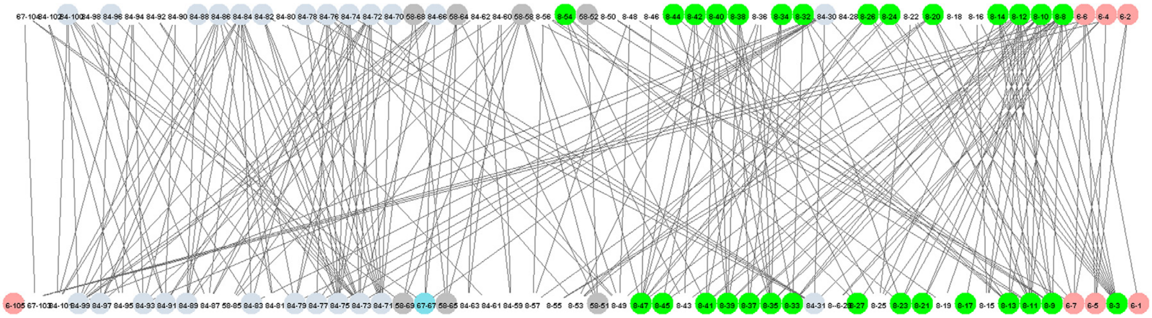

Fig. 12. Communities for the network political books for $\alpha = 0.2,\ \beta = 0.2,\ \gamma = 2$. Core nodes of the same community are colored with the same color, while peripheral nodes are white. Each node has community label and then node label. All nodes of the same community have the same community label. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

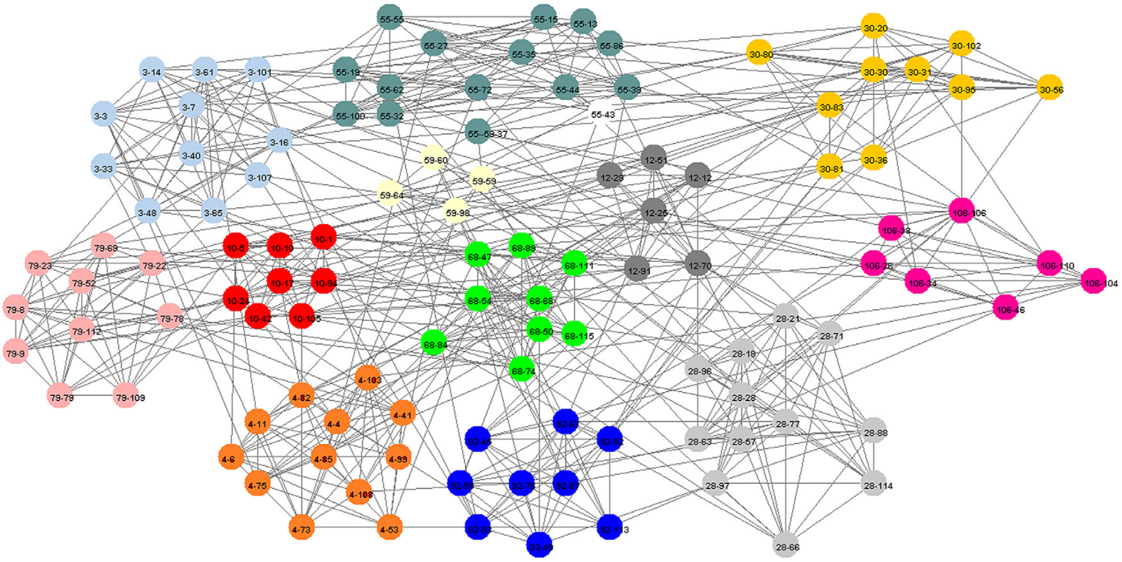

Fig. 13. Communities for Collage football network for $\alpha = 0.4,\ \beta = 0.2,\ \gamma = 2$. Core nodes of the same community are colored with the same color, while peripheral nodes are white. Each node has community label and then node label. All nodes of the same community have the same community label. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Comparison of results of Clauset et al.'s heuristic (CNM) [30], Newman's spectral divisive heuristic (divisive spectral heuristic DSH) and the proposed method (MME for normal resolution parameters 0.4 and 0.2) [29]. m denotes the number of communities and Q the modularity value of the best found solution. v is the number of nodes and e is the number of edges of the networks.

Dataset name	v	e	m CNM	Q CNM	m DSH	Q DSH	m MME	Q MME
Karate	34	78	3	0.381	4	0.393	3	0.4098
Dolphin	62	159	4	0.496	5	0.491	5	0.499
Les miserables	77	254	5	0.501	9	0.514	8	0.547
Political books	105	441	4	0.502	5	0.467	4	0.523
Football	115	613	7	0.577	8	0.493	12	0.601

6. Conclusion

Complex networks can be decomposed into subgraphs called communities. Several fast local community detection methods have been proposed using several conditions that should be satisfied by communities in networks by various authors. A too strong conditions for similarity of each node to community have a tendency to create a large number of small, but not necessary also relevant communities. We have proposed the MME algorithm which detects all communities using more different measures that can fast identify communities of different sizes and densities. More used measures provide better results. The MME extracts communities one at a time, allowing arbitrary structure in the remainder of network that can contain weakly or tightly connected nodes. The proposed method is able to identify communities both overlapping

Table 2
Number of iterations of MME algorithm (steps 2 and 3) for 6 well-known testing real-world datasets.

Dataset	Number of nodes	Number of edges	Number of iterations of algorithm (steps 2 and 3) for normal parameters
Karate	34	78	14
Dolphin	62	159	36
Political books	105	441	25
Les miserables	77	254	17
Football	115	613	40
Ring of cliques	20	35	10

and non-overlapping. The algorithm can detects communities at different resolutions and reveals rich information on input networks. We will apply the proposed method for large scale online networks.

Acknowledgments

We would like to express our deepest gratitude to the anonymous reviewers for their valuable suggestions and corrections of the paper.

This work was supported by the Slovenian Research Agency (grant number P2-0041,J2-8176).

References

[1] G. Chen, X. Wang, L. Xiang, Introduction To Complex Networks, John Wiley and Sons, 2012.
[2] M.A. Porter, J.P. Onnela, P.J. Mucha, Communities in networks, Notices Amer. Math. Soc. 56 (9) (2009) 1082–1097.
[3] S. Fortunato, Community detection in graphs, Phys. Rep. 486 (2010) 75–174.
[4] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J. Wiener, Comput. Netw. 33 (1) (2000) 309–320.
[5] D.J. Watts, S.H. Strogatz, Collective dynamics of small-world, networks, Nature 393 (1998) 4401–7442.
[6] J. Chen, Z. Saad, Dense subgraph extraction with application to community detection, IEEE Trans. Knowl. Data Eng. 24 (7) (2012) 1216–1230.
[7] Z. Dourisboure, F. Geraci, M. Pellegri, Extraction and classification of dense communities in the web, in: Proceedings of WWW, 2007.
[8] G. Palla, I. Dernyi, I. Farkas, T. Vicsek, Uncovering the overlapping community structure of complex networks in nature and society, Nature 435 (2005) 814–818.
[9] M. Girvan, M.E.J. Newman, Community structure in social and biological networks, Proc. Natl. Acad. Sci. USA 99 (2002) 7821–7826.
[10] I. Kivimaki, B. Lebichot, J. Saramaki, M. Saerens, Two betweenness centrality measures based on Randomized Shortest Paths, Sci. Rep. (6) (2016) 19668.
[11] M. Rosvall, C.T. Bergstrom, Maps of random walks on complex networks reveal community structure, Proc. Natl. Acad. Sci. USA 105 (4) (2007) 1118–1123.
[12] M.E.J. Newman, M. Grivan, Finding and evaluating community structure in networks, Phys. Rev. E 69 (2004) 026113.
[13] S. Fortunato, M. Barthemy, Resolution limit in community detection, Proc. Natl. Acad. Sci. USA 104 (1) (2006) 36–41.
[14] A. Clauset, Finding local community structure in networks, Phys. Rev. E 72 (2005) 026132.
[15] J.P. Bagrow, E.M. Bollt, Local method for detecting communities, Phys. Rev. E 72 (4) (2005) 046108.
[16] Y. Wu, R. Jin, J. Li, X. Zhang, Robust local community detection: On free rider effect and its elimination, Proc. VLDB Endow. 8 (7) (2015) 798–815.
[17] T. Laarhoven, E. Marchiori, Local network community detection with continuous optimization of conductance and weighted kernel k-means, J. Mach. Learn. Res. 17 (2016) 1–28.
[18] J. Shi, J. Malik, Normalized cuts and image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 22 (8) (2000) 888–905.
[19] U.N. Raghavan, R. Albert, S. Kumara, Near linear time algorithm to detect community structures in large-scale networks, Phys. Rev. E 76 (2007) 036106.
[20] K. Rizman Žalik, B. Žalik, Network clustering by advanced label propagation algorithm. V: IC3K 2011 : SciTePress - Science and Technology Publications, 2011, pp. 444–447.
[21] A. Lancichinetti, S. Fortunato, J. Kertesz, Detecting the overlapping and hierarchical community structure in complex networks, New J. Phys. 11 (2009) 033015.
[22] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, D. Parisi, Defining and identifying communities in networks, Proc. Natl. Acad. Sci. USA 101 (2004) 2658–2663.
[23] K. Rizman Žalik, B. Žalik, A local multiresolution algorithm for detecting communities of unbalanced structures, Physica A 407 (2014) 380–393.
[24] L.C. Freeman, Centrality in social networks: Conceptual clarification, Social Networks 1 (3) (1997) 215–239.
[25] W.W. Zachary, An information flow model for conflict and fission in small groups, J. Anthropol. Res. 33 (1977) 452–473.
[26] D. Lusseau, K. Schneider, O.J. Boisseau, P. Haase, E. Slooten, S.M. Dawson, Behav. Ecol. Sociobiol. 54 (2003) 396–405.
[27] D.E. Knuth, The Stanford GraphBase: A Platform for Combinatorial Computing, Addison-Wesley, Reading, MA, 1993.
[28] V. Krebs, Social network of political books, 2004. www.visualcomplexity.com.
[29] M. Newman, Modularity and community structure in networks, Proc. Natl. Acad. Sci. USA 103 (23) (2006) 8577–8582.
[30] A. Clauset, M. Newman, C. Moore, Finding community structure in very large networks, Phys. Rev. E 70 (2004) 066111.