

Generalized network community detection

Lovro Šubelj and Marko Bajec

University of Ljubljana, Faculty of Computer and Information Science,
SI-1001 Ljubljana, Slovenia,
`[name.surname]@fri.uni-lj.si`

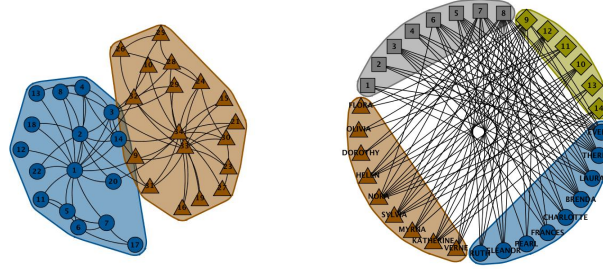
Abstract. Community structure is largely regarded as an intrinsic property of complex real-world networks. However, recent studies reveal that networks comprise even more sophisticated modules than classical cohesive communities. More precisely, real-world networks can also be naturally partitioned according to common patterns of connections between the nodes. Recently, a propagation based algorithm has been proposed for the detection of arbitrary network modules. We here advance the latter with a more adequate community modeling based on network clustering. The resulting algorithm is evaluated on various synthetic benchmark networks and random graphs. It is shown to be comparable to current state-of-the-art algorithms, however, in contrast to other approaches, it does not require some prior knowledge of the true community structure. To demonstrate its generality, we further employ the proposed algorithm for community detection in different unipartite and bipartite real-world networks, for generalized community detection and also predictive data clustering.

Keywords: link-density community, link-pattern community, propagation, community detection, data clustering

1 Introduction

Over a decade of research in network analysis has revealed a number of common properties of complex real-world networks [57,13]. *Community structure* [18]—the occurrence of cohesive modules of nodes—is of particular interest as it provides an insight into not only structural organization but also functional behavior of various real-world systems [35,2]. The analysis of communities has thus been the focus of many recent endeavors [15,38], while community structure analysis is also considered as one of the most prominent areas of network science [8,38].

However, most of the past work was constrained to communities characterized by higher density of links—*link-density communities* [18] (Fig. 1(a)). In contrast to the latter, recent studies reveal that networks comprise even more sophisticated modules than classical cohesive communities [34,1,37,52]. In particular, real-world networks can also be naturally partitioned according to common patterns of connections among nodes—into *link-pattern communities* [28,34] (Fig. 1(b)). Link-pattern communities can in fact be related to relevant functional roles in various complex systems [37,52], moreover, they also provide a



(a) Zachary's karate network [59] (b) Davis's women network [9]

Fig. 1. Comparison between (a) link-density and (b) link-pattern communities.

further comprehension of real-world network structure that is obscure under classical frameworks. Note that link-density communities could be seen as a special case of link-pattern communities, although several fundamental differences exist [52]. In particular, link-pattern communities do not correspond to densely connected groups of nodes, while generally do not even feature connectedness. The latter actually implies low transitivity—clustering coefficient [57]—for the nodes in link-pattern communities, which contradicts with small-world phenomena [57]. However, recent work suggests that best link-pattern communities might indeed emerge in parts of networks that exhibit low values of clustering (e.g., technological networks), where small-world property does not generally hold [50].

Recently, Šubelj and Bajec [52] have proposed a general propagation algorithm that can reveal arbitrary network modules ranging from link-density to link-pattern communities. Their algorithm does not require any prior knowledge of the true structure, though they introduce a community parameter that models the nature of each community according to the measure of network bottlenecks—conductance [5]. We advance the latter by proposing a more adequate modeling strategy based on node clustering coefficient [57]. The resulting algorithm is evaluated on various synthetic benchmark networks with planted partition, on random graphs and also resolution limit examples. It is shown to be comparable to current state-of-the-art, whereas, the proposed strategy also greatly improves on the approach of Šubelj and Bajec [52] (on these networks). Furthermore, to demonstrate its generality, we also employ the algorithm for community detection in different unipartite and bipartite real-world networks, for generalized community detection and predictive data clustering.

The rest of the paper is structured as follows. In Section 2 we briefly review relevant related work, with emphasis on the community detection literature. Section 3 introduces the proposed algorithm, while the empirical evaluation with formal discussion is done in Section 4. The performance on various real-world examples is presented in Section 5, and conclusions are made in Section 6.

2 Related work

Despite the wealth of the literature on classical communities in recent years [15], only a small number of authors have considered more general link-pattern communities. Nevertheless, authors have recently proposed different algorithms based on stochastic blockmodels [1,20], mixture models [34,47], model selection [28,37], data clustering [27] and other [36]. However, in contrast to the propagation algorithm proposed in this paper, and that in [52], all other approaches require some prior knowledge of the true structure (e.g., the number of communities). The latter indeed seriously limits their use in practice. Note that authors have also analyzed vertex similarity based on common patterns of connections [4,26]—commonly referred to as *structural equivalence*—whereas, some of the research on classical communities also apply for link-pattern counterparts [19,45].

It ought to be mentioned that link-pattern communities are known as *block-models* [58] in social networks literature. These have been extensively studied in the past, however, their main focus and employed formulation differs from ours.

3 Model-based propagation

Let the network be represented by an undirected graph $G(N, L)$, where N is the set of nodes of the graph and L is the set of its links (edges). Furthermore, let w_{nm} be the weight of the link between nodes $n, m \in N$. Moreover, let c_n denote the community (label) of node $n \in N$, and let Γ_n be the set of its neighbors.

The proposed model-based propagation algorithm is, as the algorithm in [52], based on the label propagation principle of Raghavan et al. [39]. In the following, we thus first introduce the latter.

Label Propagation. Label propagation algorithm [39] (LPA) reveals link-density communities by exploiting the following procedure. First, each node $n \in N$ is labeled with a unique label (i.e., $c_n = l_n$). Then, at each iteration, the node adopts the label shared by most of its neighbors (with respect to link weights). Hence,

$$c_n = \operatorname{argmax}_l \sum_{m \in \Gamma_n^l} w_{nm}, \quad (1)$$

where Γ_n^l is the set of neighbors of node n that share label l (ties are broken uniformly at random¹). Due to the existence of many intra-community links, relative to the number of inter-community links, cohesive modules of nodes form a consensus on some label after a few iterations. Thus, when the algorithm converges—a local equilibrium is reached—disconnected sets of nodes sharing the same label are classified into the same community. Due to extremely fast structural inference of label propagation, the algorithm exhibits near linear complexity and can easily scale to networks with millions of nodes and links [53,39].

Note that, to address issues with oscillations of labels in some networks (e.g., bipartite networks), label updates in Eq. (1) occur in a random order [39].

¹ When node's current label is among most frequent, the node retains its label.

General Propagation. Šubelj and Bajec [52] have argued that label propagation cannot be directly applied for the detection of link-pattern communities, as the bare nature of propagation requires connected (and cohesive) groups of nodes. However, when one considers second order neighborhoods, and propagates labels through nodes' neighbors, link-pattern communities indeed correspond to cohesive modules of nodes (see Fig. 1(b)). Based on the above they have proposed general propagation algorithm [52] (GPA) that is presented in the following.

Let δ_c be a community parameter that models the nature of community c , $\delta_c \in [0, 1]$. Assume δ_c equals 1 and 0 for link-density and link-pattern communities, respectively (to be properly defined later). Label propagation in Eq. (1) is then advanced into a general community detection algorithm as

$$c_n = \operatorname{argmax}_l \left(\delta_l \sum_{m \in \Gamma_n^l} w_{nm} b_m d_m + (1 - \delta_l) \sum_{m \in \Gamma_n^l \setminus \Gamma_n | s \in \Gamma_n} w_{nm}^s b_m \tilde{d}_m \right) \quad (2)$$

where $w_{nm}^s = \frac{w_{ns} w_{sm}}{s_s}$ and s_n is the strength of node $n \in N$ (i.e., $s_n = \sum_{m \in \Gamma_n} w_{nm}$). In the case of link-density communities (left-hand side of Eq. (2)), the labels are propagated among the neighbors as before, whereas, in the case of link-pattern communities (right-hand side of Eq. (2)), the labels are propagated through nodes' neighbors—between the nodes at distance two. Thus, the algorithm can indeed reveal either link-density or link-pattern communities, or different mixtures of both, when they are clearly depicted in the network's topology. (Note that in [52] the algorithm was presented for unweighted networks.)

Node balancers b_n [51] and diffusion values d_n, \tilde{d}_n [53,52] in Eq. (2) improve the algorithm's stability and accuracy, respectively. More precisely, random label update orders (see above) severely hamper the robustness of the approach, and consequently also the stability of the identified community structure [54]. In particular, nodes that are updated at the beginning exhibit higher propagation preferences than those that are updated towards the end [51]. Thus, balancers b_n are utilized to counteract for the randomness introduced by update orders—lower and higher preferences are given to the nodes updated first and last, respectively.

Let i_n denote a normalized position of node $n \in N$ in some random order, $i_n \in (0, 1]$. Then, node balancers are set according to

$$b_n = \frac{1}{1 + e^{-\mu(i_n - \lambda)}}, \quad (3)$$

where λ and μ are parameters of the algorithm. Intuitively, we fix λ to $\frac{1}{2}$, while μ is set to 2 based on some preliminary experiments (see Section 4). Node balancers can also be modeled with a linear function as $b_n = i_n$, however, introduction of the above parameters allows for a distinct control over the algorithm. In particular, analysis in Section 4 reveals that increasing μ improves the stability of the algorithm, although the computational time thus also increases. Note also that setting μ to 0 yields a classical label propagation where all b_n are equal.

To further boost the community detection strength of the algorithm, defensive preservation of communities is employed through diffusion values d_n ,

\tilde{d}_n [53,52]. Here higher diffusion values—propagation preferences—are given to core nodes of each (current) community, while lower values are given to their border nodes. The latter results in an immense ability of detecting communities, even when they are only weakly depicted in the network’s topology [53]. At each iteration, diffusion values are estimated by means of a random walker utilized on each (current) community. Hence,

$$d_n = \sum_{m \in \Gamma_n^{c_n}} d_m / k_m^{c_n} \quad (4)$$

and

$$\tilde{d}_n = \sum_{m \in \Gamma_s^{c_n} \setminus \Gamma_n | s \in \Gamma_n} \frac{\tilde{d}_m}{\sum_{s \in \Gamma_m} k_s^{c_n}}, \quad (5)$$

where $k_n^{c_n}$ is the intra-community degree of node $n \in N$ (all d_n, \tilde{d}_n are initialized to $\frac{1}{|N|}$). Besides deriving an estimate of the core and border of each community, the main rationale here is to formulate propagation—diffusion—within each community, to estimate the current state of label propagation, and then to adequately alter the dynamics of the process. Analysis in Section 4 reveals that defensive preservation of communities significantly improves the detection strength of the algorithm, while for further discussion and analysis see [53].

Despite the discussion above, the core of the algorithm is in fact represented by a community modeling strategy implemented through parameters δ_c . Šubelj and Bajec [52] have proposed to measure the conductance [5] of each community, to determine whether it better conforms with link-density or link-pattern regime. Conductance $\Phi(c)$ of community c is defined as a relative size of the corresponding network cut—ratio of inter-community links—thus it is a measure of network bottlenecks. Hence, at each iteration, they simply set $\delta_c = 1 - \Phi(c)$, while all δ_c are initialized to $\frac{1}{2}$. The main weakness of their strategy is that each community is considered independently of other. Thus, in the following, we propose a more adequate community modeling strategy based on the properties of complex real-world networks.

Model-based Propagation. Community modeling strategy of Šubelj and Bajec [52] considers merely the nature of each respective community, whereas all other communities are disregarded. Although no proper empirical study exists, in an ideal case, link-pattern communities would link to other link-pattern communities rather than to other link-density communities. The latter follows from the fact that the concerned links would else obviously decrease the quality of the respective link-density community—make it a link-pattern community. Thus, we propose a community model based on the hypothesis that the neighbors’ communities should be of the same type—either link-density or link-pattern—as the concerned node’s community. Hence,

$$\delta_c = \frac{1}{|N^c|} \sum_{m \in \Gamma_n | n \in N^c} \frac{\delta_{c_m}}{k_n}, \quad (6)$$

where k_n is the degree of node $n \in N$ and N^c is the set of nodes in community c .

We also argue that an adequate initialization of community parameters δ_c is of vital importance (exact results are omitted). Otherwise, the algorithm can easily get trapped in some local stable—probably suboptimal—fixed point that is hard to escape from. However, Eq. (6) cannot be directly employed at the beginning, as all nodes still reside in their own communities. We thus refine the above hypothesis such that the node’s neighbors should not only reside in the same type of the community, but in the same respective community. The latter immediately implies that the neighbors of the nodes in link-density communities should also link to each other, whereas the opposite holds for the nodes in link-pattern communities. Hence, for each node $n \in N$, one could initially set δ_{c_n} to C_n , where C_n is a node clustering coefficient [57] defined as the probability that two neighbors of node n also link to each other—network transitivity. It ought to be mentioned that recent work suggests that transitivity—rather than homophily—gives rise to the modular structure in real-world networks [17].

However, consider a node with very high degree—a hub node. Hubs commonly appear in link-density communities [19], still, due to a large number of links, they would only rarely experience high values of clustering coefficient (the opposite would in fact imply a large clique). Also, as most networks are disassortative by degree [31], hubs tend to link to low degree nodes that cannot provide for high clustering of the hub node [48]. Indeed, in many real-world networks node clustering coefficient roughly follows $C_n \sim k_n^{-1}$ [56,40,48], where k_n is the degree of node $n \in N$. Hence, we model initial communities as (assume $C_n > 0$)

$$\delta_{c_n} = \begin{cases} 1 & \text{for } C_n > \alpha k_n^{-1} + \beta, \\ \rho & \text{otherwise,} \end{cases} \quad (7a)$$

$$(7b)$$

where α and β are estimated from the network using ordinary least squares, and ρ is a parameter. We set ρ to $\frac{1}{4}$ based on some preliminary experiments.

Eq. (7) and Eq. (6) define the proposed model-based propagation algorithm (MPA), which is else (almost) identical to the algorithm in [52] (see Alg. 1). However, the evaluation on synthetic and real-world networks in Section 4 and Section 5, respectively, reveals that the proposed approach significantly outperforms that in [52]. For a thorough evaluation, we also analyze two variations of the basic approach that fix all community parameters δ_c to either 1 or 0. The approaches thus result in a fully link-density or link-pattern community detection algorithms, and are denoted MPA(D) and MPA(P), respectively.

4 Evaluation and discussion

In the following we evaluate the proposed algorithm on different synthetic benchmark networks with planted partition, and also on random networks. The results are assessed in terms of three different measures of community significance, borrowed from the field of information theory and community detection literature.

Let \mathcal{C} be a partition extracted by an algorithm and let \mathcal{P} be the known partition of the network (corresponding random variables are C and P , respectively).

Algorithm 1 Model-based propagation algorithm (MPA)

Input: Graph $G(N, L)$ and parameters λ, μ, ρ **Output:** Communities C

```
{Community initialization.}
for  $n \in N$  do
   $c_n \leftarrow l_n$  {Unique label.}
   $\delta_{c_n} \leftarrow$  {Model according to Eq. (7).}
   $d_n, \tilde{d}_n \leftarrow 1/|N|$ 
end for
{Model-based propagation.}
while not converged do
  shuffle( $N$ )
  for  $n \in N$  do
    {General propagation.}
     $b_n \leftarrow 1/(1 + e^{-\mu(i_n - \lambda)})$ 
     $c_n \leftarrow \operatorname{argmax}_l \left( \delta_l \sum_{m \in \Gamma_n^l} w_{nm} b_m d_m + (1 - \delta_l) \sum_{m \in \Gamma_n^l \setminus \Gamma_n} w_{nm}^s b_m \tilde{d}_m \right)$ 
    {Re-estimation.}
     $d_n \leftarrow \sum_{m \in \Gamma_n^{c_n}} d_m / k_m^{c_n}$  and  $\tilde{d}_n \leftarrow \sum_{m \in \Gamma_n^{c_n} | s \in \Gamma_n} \tilde{d}_m / \sum_{s \in \Gamma_n} k_s^{c_n}$ 
  end for
  for  $c \in C$  do
    {Community modeling.}
     $\delta_c \leftarrow 1/|N^c| \sum_{m \in \Gamma_n | n \in N^c} \delta_{c_m} / k_n$  {Omitted on first iteration.}
  end for
end while
return  $C$ 
```

First—normalized mutual information [8] (NMI)—has become a de facto standard in the recent literature. NMI of \mathcal{C} and \mathcal{P} is defined as $\frac{2I(\mathcal{C}, \mathcal{P})}{H(\mathcal{C}) + H(\mathcal{P})}$, where $I(\mathcal{C}, \mathcal{P})$ is the mutual information, and $H(\mathcal{C})$, $H(\mathcal{P})$ and $H(\mathcal{C}|\mathcal{P})$ are standard and conditional entropies. NMI of identical partitions equals 1, and is 0 for independent ones, $\text{NMI} \in [0, 1]$. Second, we also consider normalized variation of information [30, 22] (NVOI), which is a symmetric local measure that has the properties of a distance in the space of partitions. NVOI of \mathcal{C} and \mathcal{P} equals $\frac{H(\mathcal{C}|\mathcal{P}) + H(\mathcal{P}|\mathcal{C})}{\log |N|}$, therefore, in contrast to NMI, lower values represent better correlation between partitions, $\text{NVOI} \in [0, 1]$. Last, for a better comprehension, we also adopt a more intuitive measure—fraction of correctly classified nodes [18] (FCC)—that is commonly adopted within community detection literature. The node is considered correctly classified, if it resides in the same community as at least one half of the nodes in its true community. Again, $\text{FCC} \in [0, 1]$.

Community detection algorithms introduced in Section 3 are compared against a greedy agglomerative optimization [32, 7] of modularity Q [33] (denoted MO(G))—a classical link-density community detection algorithm—and a mixture model with expectation-maximization [10] proposed by Newman and Leicht [34] (denoted MM(EM)). The latter can detect arbitrary network modules and is currently among state-of-the-art approaches for generalized community detection [34, 37].

However, it demands the correct number of communities to be known ahead of time, which puts the algorithm in significant advantage compared to others [23]. For simplicity, we limit the number of iterations to 100 for all algorithms.

GN2 Benchmark. The algorithms are first applied to a class of benchmark networks [37] that is in fact a generalization of a classical benchmark proposed by Girvan and Newman [18]. Networks comprise four communities of 32 nodes, whereas, two communities correspond to classical link-density modules, while the other two form a bipartite structure of link-pattern communities. Average degree is fixed to 16, while the community structure is controlled by a mixing parameter θ , $\theta \in [0, 1]$. When θ is 0, all links are set according to the designed community structure, while for θ equal 1, the networks are completely random.

The results are shown in Fig. 2. Observe that for small values of θ only MPA and MPA(P) can accurately reveal the planted structure in these networks. However, when θ increases, the performance of MPA is similar to that of a classical community detection algorithm (e.g., MO(G) or MPA(D)). MM(EM) can detect communities to some extent until $\theta \leq \frac{1}{3}$ (dashed lines in Figs. 2, 3)—when, for the nodes within link-density communities, there are twice as many links that conform with the planted structure than randomly placed links. Note also that twice as many links are needed to define a link-pattern community, compared to a respective link-density community, which would yield the same threshold at $\theta = \frac{1}{6}$ for these networks (solid lines in Figs. 2, 3). Thus, MPA accurately extracts planted link-density and link-pattern communities in these networks, as long as they are clearly depicted in the network’s topology. Note also that community modeling strategy within MPA seems more adequate than that of GPA.

SB Benchmark. GN2 benchmark provides a rather unrealistic testbed due to homogeneous degree and community size distributions. We address the latter by proposing a class of simple benchmark networks with heterogeneous community sizes. Networks comprise three communities of 16, 32 and 24 nodes, respectively (see network in Fig. 8(a)). The latter two again form a bipartite structure of link-pattern communities, while the third community corresponds to a classical cohesive module. Links are placed according to the designed community structure such that the average degree of the nodes in the first and third community is fixed to 16. The latter implies an average degree of 8 for the nodes in the second community. Furthermore, we also add some number of links uniformly at random for each node—denoted node confusion degree κ , $\kappa \geq 0$.

The results appear in Fig. 3. The performance of the algorithms is rather similar to that on GN2 benchmark (note different scales in Figs. 2, 3). Only MPA can accurately reveal the planted structure for small values of κ , while the model within GPA again seems to fail. Observe that MM(EM) can extract communities equally well, even when κ equals 16—only $\frac{1}{3}$ of the links for the nodes in the second community still agrees with the intrinsic structure, thus, the communities are only marginally defined. The latter clearly demonstrates that knowing an exact number of communities indeed presents a significant advantage.

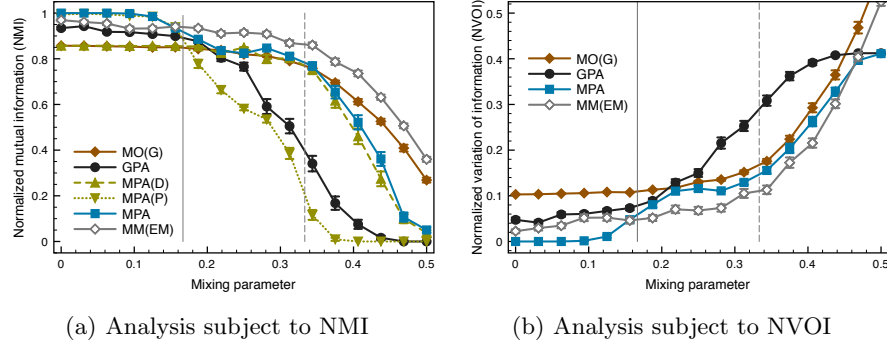


Fig. 2. Analysis on GN2 benchmark networks [37]. The values are estimates over 100 network realizations, while error bars show standard error of the mean.

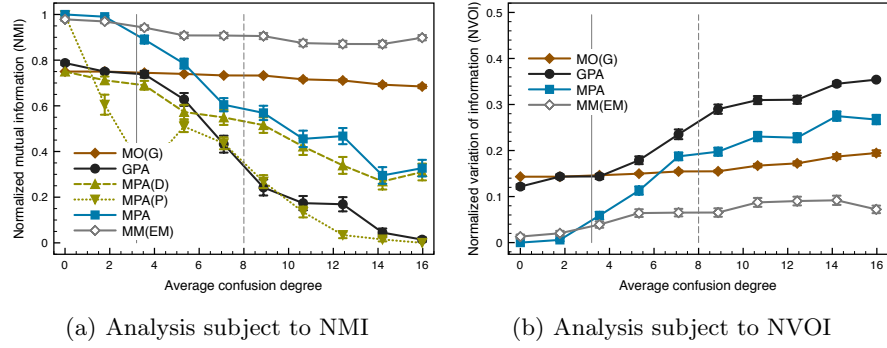


Fig. 3. Analysis on SB benchmark networks. The values are estimates over 100 network realizations, while error bars show standard error of the mean.

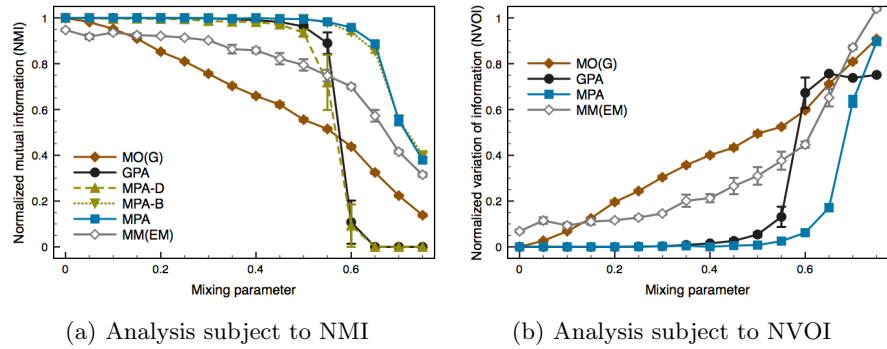


Fig. 4. Analysis on LFR benchmark networks. The values are estimates over 10 network realizations, while error bars show standard error of the mean. To ensure convergence, μ is set to $\frac{1}{2}$.

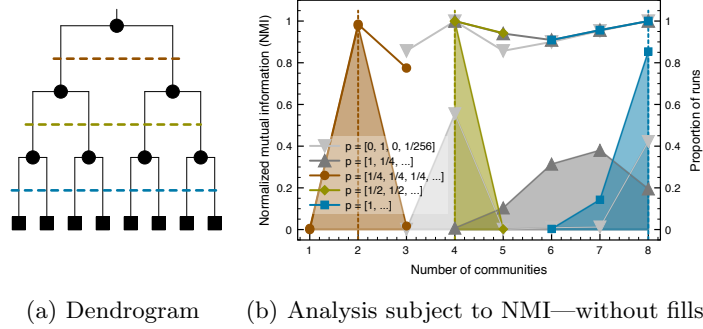


Fig. 5. Analysis on HN benchmark networks [6]. The values are estimates over 1000 network realizations, while missing values of p in the legend equal $\frac{1}{16}$. See also text.

LFR Benchmark. To enable easier comparison with previous literature on community detection, we also apply the algorithms to a class of standard benchmark networks with scale-free degree and community size distributions proposed by Lancichinetti et al. [25]. The size of the networks is set to 1000, while community sizes range between 10 and 50 nodes. Note that all communities here correspond to a link-density regime. As before, the quality of the planted structure is controlled by a mixing parameter θ , $\theta \in [0, 1]$. For comparison, we also analyze two variations of MPA that do not employ either balanced propagation or defensive preservation of communities (denoted MPA-D and MPA-B, respectively).

Results in Fig. 4 show that MPA most accurately reveals the planted structures in these networks, while it also significantly outperforms the other generalized community detection algorithm MM(EM). Observe also that defensive preservation of communities greatly improves the algorithm’s community detection strength. Comparing the results with an analysis of over ten state-of-the-art approaches for classical community detection conducted in [24], we conclude that, at least on these networks, MPA performs similarly as the best algorithms analyzed there. These are hierarchical modularity optimization of Blondel et al. [3], model selection technique of Rosvall and Bergstrom [44], spectral algorithm proposed by Donetti and Muñoz [11] and multi-resolution spin model of Ronhovde and Nussinov [43].

HN Benchmark. Next, we also analyze the proposed algorithm on a class of benchmark networks with a hierarchical structure [6]. In particular, networks are constructed according to a community dendrogram in Fig. 5(a), where leaves correspond to eight modules of 16 nodes, while each node d of the dendrogram is also associated with a probability p_d , $p_d \in [0, 1]$. The nodes of the network are linked with the probability associated with the lowest common ancestor in the community dendrogram. Varying the values of p_d can infer (almost) arbitrary hierarchical structure of either link-density or link-pattern communities. However,

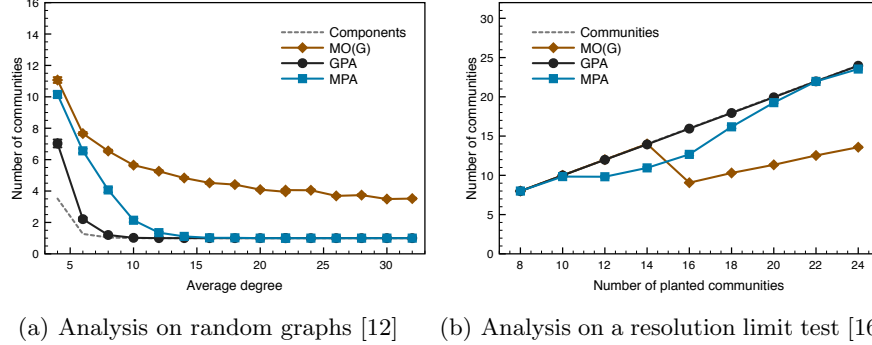


Fig. 6. Analysis on (a) random graphs [12] and (b) resolution limit test networks [16]. The values are estimates over 100 network realizations, while error bars are smaller than the symbol sizes.

due to simplicity, we associate each level of the nodes with the same probability p_d . Thus, denote $p = [p_1, p_2, p_3, p_4]$ to be the vector of respective probabilities for the nodes from the lowest to the highest level of the hierarchy, respectively.

The performance of MPA on five realizations of the above benchmark can be seen in Fig. 5. Values of NMI were estimated such that each revealed partition was compared against (only) three intrinsic community structures—represented by dashed lines in Fig. 5(a)—and the best correspondence was reported. (Note that the results are thus actually rather pessimistic.) Observe that MPA can accurately reveal the planted structure in all five cases—see legend in Fig. 5(b)—which further confirms the adequacy of the proposed community model. More precisely, in the first case, the intrinsic network structure results in a hierarchy of link-pattern communities, whereas, in the second case, communities are in fact defined on two levels of the designed hierarchy. In each of the last three cases, the communities corresponds to a single level of the hierarchy. Thus, MPA can indeed be employed for the detection of arbitrary community structure.

Random Graphs. We also apply the algorithms to Erdős-Rényi random graphs [12] that presumably have no community structure. We fix the number of nodes to 128 and vary the average degree from 4 to 32. The results are shown in Fig. 6(a). Note that, in contrast to MO(G), neither MPA nor GPA reports any community structure for these networks—all nodes are classified into a single community.

Resolution Limit. We further analyze the algorithms on a resolution limit [16]—existence of an intrinsic scale within the algorithm, below which the communities are no longer recognized—test benchmark networks [16]. Hence, the networks consist of cliques with 4 nodes that are linked into a ring. Results in Fig. 6(b) reveal that neither MPA nor GPA is seriously attributed to the resolution limit issues, whereas, the opposite holds for MO(G). Although some fluctuations are indeed observed for MPA, these are not as severe as in the case of modularity [16].

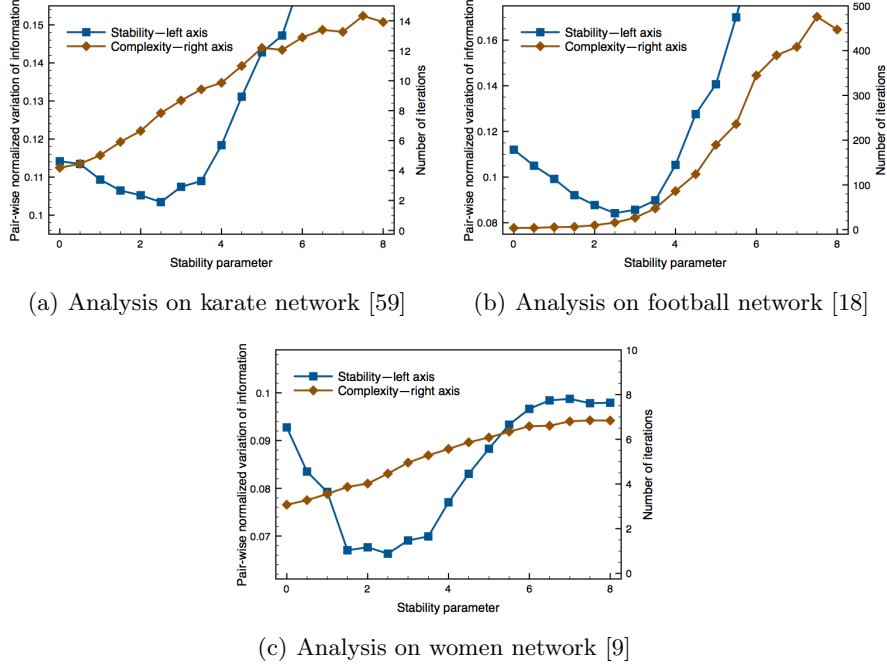


Fig. 7. Analysis of stability and complexity of MPA on three real-world networks from Table 1. The values are estimates over at least 100 runs (note different scales).

Algorithm Stability. As previously discussed, random label update orders severely hamper the stability of label propagation, and thus also the robustness of the revealed community structure [54]. Hence, balanced propagation [51] is employed, yet this introduces two parameters λ and μ (Section 3). Value of λ is intuitively fixed to $\frac{1}{2}$ (see Eq. (3)), while parameter μ in fact controls the stability of the algorithm. In Fig. 7 we analyze MPA with respect to stability parameter μ on three real-world networks from Table 1. Plots show pair-wise distance between revealed community structures, and also the number of iterations for the algorithm to converge (note different scales). Observe that increasing μ improves the stability of MPA in all three networks, however, the number of iterations also increases. Furthermore, as one would expect, when μ exceeds a certain threshold, pair-wise distance between community structures notably increases—some number of nodes already gets completely disregarded due to propagation preferences close to 0 (see Eq. (3))—while the number of iterations can also increase substantially (see Fig. 7(b)). The transition occurs at around $\mu \approx 4$ for these networks, thus, for the analysis throughout the paper, μ is set to 2 (if not stated otherwise). It ought to be mentioned that balanced propagation can also improve community detection strength of the basic label propagation [51] (see above).

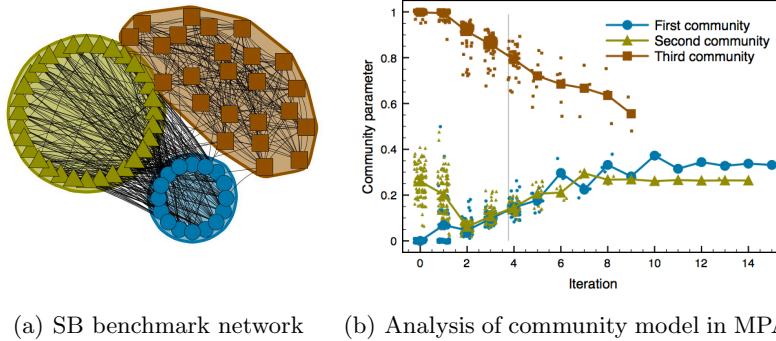


Fig. 8. Analysis of community modeling strategy of MPA on SB benchmark networks. Node shapes represent planted network communities and are consistent among figures. The values are estimates over 100 network realizations, while, for an adequate analysis, we set ρ to 0 and increase μ (Section 3). See also text.

Community Modeling. For a comprehensive analysis, we also directly analyze the proposed community modeling strategy of MPA on SB benchmark networks with confusion degree set to 2 (see above). In particular, we measure the average value of community parameter δ_c (Section 3) for the nodes in each of the planted network communities. Results in Fig. 8 show that, at least for these networks, values of community parameter δ_c clearly distinguish between link-density and link-pattern regime—average δ_c is close to 1 and 0 for the nodes in link-density and link-pattern communities, respectively. Note also that, due to lower average degree, values of δ_c are initially higher for the larger of the two link-pattern communities. However, before the algorithm converges—average number of iterations is shown by a horizontal line in Fig. 8(b)—community model in MPA infers the same average value of δ_c for both link-pattern communities. Note also that GPA cannot properly model communities planted in these networks (Fig. 3).

Computational Complexity. Basic label propagation and its advances exhibit near linear time complexity $O(|L|)$ [39], where $|L|$ is the number of links in the network. In particular, the exact complexity was estimated to around $O(|L|^{1.2})$ [53]. Similarly, the proposed model-based propagation MPA exhibits complexity near $O(k|L|)$, where k is the average degree in the network. Although a thorough empirical analysis is out of scope of this paper, based on the results in [53] (and above), we estimate that MPA should scale up to networks with a million links—accessible on a standard desktop computer within an hour.

Final Remarks. The above analysis on different benchmark networks and random graphs indeed confirms that MPA can reveal arbitrary composites of either link-density or link-pattern communities, as long as they are clearly depicted in the network’s topology. Moreover, the proposed community modeling strategy also seems more adequate than the approach proposed by Šubelj and Bajec [52]

Table 1. Analysis on real-world networks subject to NMI estimated over 1000 runs (10 runs for software networks). Corporate network is reduced to the largest component, while the known partition is also limited to 86 corporate nodes—we thus set μ to $\frac{1}{2}$.

Network	Nodes	Links	Comm.	MO(G)	GPA	MM(EM)	MPA
Zachary’s karate club [59]	34	78	2	0.6925	0.7155	0.7870	0.8949
American college football [18]	115	616	12	0.7547	0.8769	0.8049	0.8919
Davis’s southern women [9]	32	89	4		0.7338	0.8332	0.8084
Scottish corpor. interlocks [46]	217	348	8		0.6634	0.5988	0.6411
Java (org namespace) [49]	709	3571	47	0.5029	0.5190	—	0.5187
Java (javax namespace) [49]	1595	5287	107	0.7048	0.7369	—	0.7386

for all networks considered. Further note that, although MPA is mostly outperformed by MM(EM) on the benchmarks above, the latter should be attributed to the fact that MM(EM) is advised about the number of communities. However, this currently cannot be properly estimated for large networks [23]. Moreover, MPA also performs significantly better on real-world networks in Section 5.

5 Real-world examples

In the following we further employ the proposed algorithm for community detection in different unipartite and bipartite social networks—classical and fully link-pattern community detection, respectively—and also for a generalized community detection and predictive data clustering. All of the networks considered below are regarded as unweighted and undirected.

Community Detection We first consider two classical networks for community detection—a network of social interactions between members of a karate club analyzed by Zachary [59], and a network of interplays in the 2000 NCAA American football schedule proposed in [18]—and two well-known bipartite networks—a network of social collaborations between women in Natchez, Mississippi collected by Davis [9], and a network of corporate interlocks in Scotland between 1904 and 1905 introduced in [46] (see Table 1). All these networks have known natural community structures that results from earlier studies (see also Fig. 1).

Propagation algorithms—MPA and GPA—most accurately reveal the true community structure for main of these networks (Table 1), whereas, community modeling strategy of MPA again seems more adequate than that of GPA. Note also that most values of NMI for MPA in Table 1 are considerably high.

Next, we also consider two software class dependency networks representing **org** and **javax** namespaces of Java language compiled in [49]. Here, the natural community structure should coincide with respective software packages [49], while these are expected to conform with link-density and also link-pattern

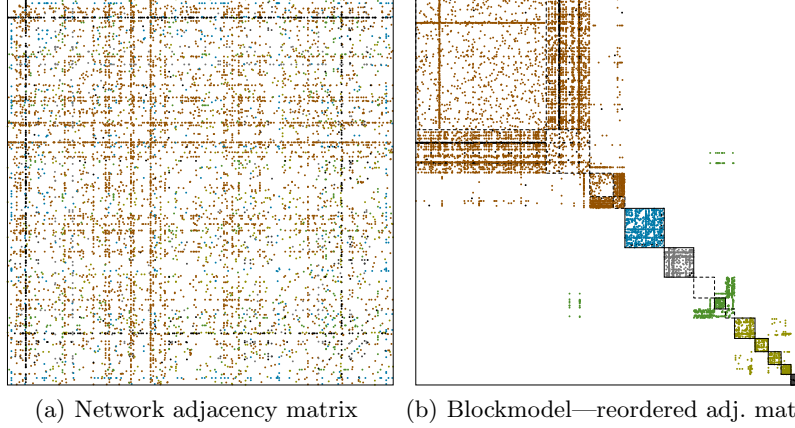


Fig. 9. Community structure of Java software network revealed with MPA (b). Only communities with more than 24 nodes are shown, still, the structure contains 1020 nodes and 4184 links. Link colors correspond to high-level software packages—`javax.swing`, `javax.management`, `javax.xml`, `javax.print`, `javax.naming`, `javax.lang` and other—while each dot was enlarged five times for better visibility.

regime [52]. Again, propagation algorithms most accurately extract the true network structures (Table 1), whereas MM(EM) fails completely. In Fig. 9 we also show the community structure of `javax` network revealed with MPA that obtains $NMI = 0.7431$. Observe how communities rather agree with high-level software packages, whereas, the majority of the links in the network is consistent with the revealed structure. Interestingly, some packages contain mainly link-pattern communities (e.g., `javax.swing`), while others are composed of only link-density communities (e.g., `javax.xml`).

Data Clustering To apply community detection algorithm for data clustering, the respective dataset must first be represented by a network using some measure of similarity. According to [42], we adopt the inversed Chebyshev distance, with initial $[0, 1]$ -normalization. In order to obtain a sparse network, links must also be thresholded accordingly. (Due to simplicity, we consider only unweighted versions of the algorithms.) Note that the resulting network thus commonly decomposes into several connected components, however, community detection algorithm can still be employed to further partition these components (see Table 2).

We employ community detection to predict class variables of two famous datasets—Iris plants dataset introduced by Fisher [14], and Ecoli protein localization sites dataset [21]. For comparison, in Table 2 we also report the results for a classical partitional clustering algorithm K-Means [29] (denoted KM). Observe that MPA obtains extremely promising results on these datasets, while it also significantly outperforms MM(EM) and KM that are both advised about the

Table 2. Analysis of data clustering on two real-world datasets subject to NMI and FCC, respectively (estimated over 100 runs).

Dataset	Items	Classes	Links	Comp.	KM	MM(EM)	MPA
Iris plants dataset [14]	150	3	2405	2	0.8234	0.8113	0.8264
					0.8227	0.8196	0.8983
Ecoli protein dataset [21]	336	8	14685	4	0.5835	0.0797	0.6251
					0.2530	0.0277	0.4164

number of communities. Still, the results could be further improved in various ways. (Note that low NMI for MM(EM) on Ecoli dataset is not entirely evident.)

6 Conclusions

The paper proposes an enhanced community modeling strategy for a recently introduced general propagation algorithm [52]. The resulting algorithm can detect arbitrary network modules—ranging from link-density communities to link-pattern communities—while, in contrast to most other approaches, it requires no apriori knowledge about the true structure (e.g., the number of communities). The algorithm was evaluated on various benchmark networks with planted partition, on random graphs and resolution limit test networks, where it is shown to be at least comparable to current state-of-the-art. Moreover, to demonstrate its generality, the algorithm was also employed for community detection in different unipartite and bipartite social networks, for generalized community detection and data clustering. The results imply that the proposed community model provides an adequate approximation of the real-world network structure, although, recent work suggests that network clustering and degree mixing could be even further utilized within the model [48,17,55,41]. The latter will be considered for future work. (For supporting website see <http://lovro.lpt.fri.uni-lj.si/>.)

Acknowledgments

This work has been supported by the Slovene Research Agency ARRS within Research Program No. P2-0359.

References

1. Airoldi, E.M., Blei, D.M., Fienberg, S.E., Xing, E.P.: Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.* 9, 1981–2014 (2008)
2. Arenas, A., Díaz-Guilera, A., Pérez-Vicente, C.J.: Synchronization reveals topological scales in complex networks. *Phys. Rev. Lett.* 96(11), 114102 (2006)
3. Blondel, V.D., Guillaume, J., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. *J. Stat. Mech.* P10008 (2008)

4. Blondel, V.D., Gajardo, A., Heymans, M., Senellart, P., Dooren, P.V.: A measure of similarity between graph vertices: Applications to synonym extraction and web searching. *SIAM Rev.* 46(4), 647–666 (2004)
5. Bollobás, B.: *Modern graph theory*. Springer (1998)
6. Clauset, A., Moore, C., Newman, M.E.J.: Hierarchical structure and the prediction of missing links in networks. *Nature* 453(7191), 98–101 (2008)
7. Clauset, A., Newman, M.E.J., Moore, C.: Finding community structure in very large networks. *Phys. Rev. E* 70(6), 066111 (2004)
8. Danon, L., Díaz-Guilera, A., Duch, J., Arenas, A.: Comparing community structure identification. *J. Stat. Mech.* P09008 (2005)
9. Davis, A., Gardner, B.B., Gardner, M.R.: *Deep south*. Chicago University Press (1941)
10. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc.* 39(1), 1–38 (1977)
11. Donetti, L., Munoz, M.A.: Detecting network communities: A new systematic and efficient algorithm. *J. Stat. Mech.* P10012 (2004)
12. Erdős, P., Rényi, A.: On random graphs i. *Publ. Math. Debrecen* 6, 290–297 (1959)
13. Faloutsos, M., Faloutsos, P., Faloutsos, C.: On power-law relationships of the internet topology. *Comput. Commun. Rev.* 29(4), 251–262 (1999)
14. Fisher, R.A.: The use of multiple measurements in axonomic problems. *Annals Eugenics* 7, 179–188 (1936)
15. Fortunato, S.: Community detection in graphs. *Phys. Rep.* 486(3-5), 75–174 (2010)
16. Fortunato, S., Barthelemy, M.: Resolution limit in community detection. *P. Natl. Acad. Sci. USA* 104(1), 36–41 (2007)
17. Foster, D.V., Foster, J.G., Grassberger, P., Paczuski, M.: Clustering drives assortativity and community structure in ensembles of networks. e-print arXiv:10122384v1 (2010)
18. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. *P. Natl. Acad. Sci. USA* 99(12), 7821–7826 (2002)
19. Guimerà, R., Amaral, L.A.N.: Functional cartography of complex metabolic networks. *Nature* 433(7028), 895–900 (2005)
20. Hofman, J.M., Wiggins, C.H.: Bayesian approach to network modularity. *Phys. Rev. Lett.* 100(25), 258701 (2008)
21. Horton, P., Nakai, K.: A probabilistic classification system for predicting the cellular localization sites of proteins. In: *Proceedings of the International Conference on Intelligent Systems for Molecular Biology*. pp. 109–115 (1996)
22. Karrer, B., Levina, E., Newman, M.E.J.: Robustness of community structure in networks. *Phys. Rev. E* 77(4), 046119 (2008)
23. Karrer, B., Newman, M.E.J.: Stochastic blockmodels and community structure in networks. *Phys. Rev. E* 83, 016107 (2011)
24. Lancichinetti, A., Fortunato, S.: Community detection algorithms: A comparative analysis. *Phys. Rev. E* 80(5), 056117 (2009)
25. Lancichinetti, A., Fortunato, S., Radicchi, F.: Benchmark graphs for testing community detection algorithms. *Phys. Rev. E* 78(4), 046110 (2008)
26. Leicht, E.A., Holme, P., Newman, M.E.J.: Vertex similarity in networks. *Phys. Rev. E* 73(2), 026120 (2006)
27. Lin, C., Koh, J., Chen, A.L.P.: A better strategy of discovering link-pattern based communities by classical clustering methods. In: *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*. pp. 56–67 (2010)

28. Long, B., Xu, X., Zhang, Z., Yu, P.: Community learning by graph approximation. In: Proceedings of the International Conference on Data Mining. pp. 232–241 (2007)
29. MacQueen, J.B.: Some methods for classification and analysis of multivariate observations. In: Proceedings of Berkeley Symposium on Mathematical Statistics and Probability. pp. 281–297 (1967)
30. Meila, M.: Comparing clusterings: An information based distance. *J. Multivariate Anal.* 98(5), 873–895 (2007)
31. Newman, M.E.J.: Assortative mixing in networks. *Phys. Rev. Lett.* 89(20), 208701 (2002)
32. Newman, M.E.J.: Detecting community structure in networks. *Eur. Phys. J. B* 38(2), 321–330 (2004)
33. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. *Phys. Rev. E* 69(2), 026113 (2004)
34. Newman, M.E.J., Leicht, E.A.: Mixture models and exploratory analysis in networks. *P. Natl. Acad. Sci. USA* 104, 9564–9569 (2007)
35. Palla, G., Derényi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. *Nature* 435, 814 (2005)
36. Park, Y., Moore, C., Bader, J.S.: Dynamic networks from hierarchical bayesian graph clustering. *PLoS One* 5(1), e8118 (2010)
37. Pinkert, S., Schultz, J., Reichardt, J.: Protein interaction networks: More than mere modules. *PLoS Comput. Biol.* 6(1), e1000659 (2010)
38. Porter, M.A., Onnela, J., Mucha, P.J.: Communities in networks. *Not. Am. Math. Soc.* 56(9) (2009)
39. Raghavan, U.N., Albert, R., Kumara, S.: Near linear time algorithm to detect community structures in large-scale networks. *Phys. Rev. E* 76(3), 036106 (2007)
40. Ravasz, E., Barabási, A.: Hierarchical organization in complex networks. *Phys. Rev. E* 67(2), 026112 (2003)
41. Ren, F., Cheng, X., Shen, H.: Modeling the clustering in citation networks. e-print arXiv:11044209v1 (2011)
42. Rodrigues, F.A., De Arruda, G.F., Costa, L.: A complex networks approach for data clustering. e-print arXiv:11015141v1 (2011)
43. Ronhovde, P., Nussinov, Z.: Local resolution-limit-free potts model for community detection. *Phys. Rev. E* 81(4), 046114 (2010)
44. Rosvall, M., Bergstrom, C.T.: Maps of random walks on complex networks reveal community structure. *P. Natl. Acad. Sci. USA* 105, 1118–1123 (2008)
45. Rosvall, M., Bergstrom, C.T.: An information-theoretic framework for resolving community structure in complex networks. *P. Natl. Acad. Sci. USA* 104(18), 7327–7331 (2007)
46. Scott, J., Hughes, M.: The anatomy of Scottish capital: Scottish companies and Scottish capital, 1900–1979. Croom Helm (1980)
47. Shafiei, M., Chipman, H.: Mixed-membership stochastic block-models for transactional networks. In: Proceedings of the IEEE International Conference on Data Mining. pp. 1019–1024 (2010)
48. Soffer, S.N., Vázquez, A.: Network clustering coefficient without degree-correlation biases. *Phys. Rev. E* 71(5), 057101 (2005)
49. Subelj, L., Bajec, M.: Community structure of complex software systems: Analysis and applications. *Physica A* 390(16), 2968–2975 (2011)
50. Subelj, L., Bajec, M.: On clustering assortativity and community structure of complex real-world networks. in preparation (2011)

51. Subelj, L., Bajec, M.: Robust network community detection using balanced propagation. *Eur. Phys. J. B* 81(3), 353–362 (2011)
52. Subelj, L., Bajec, M.: Ubiquitousness of link-density and link-pattern communities in real-world networks. e-print arXiv:11043162v2 (2011)
53. Subelj, L., Bajec, M.: Unfolding communities in large complex networks: Combining defensive and offensive label propagation for core extraction. *Phys. Rev. E* 83(3), 036103 (2011)
54. Tibély, G., Kertész, J.: On the equivalence of the label propagation method of community detection and a potts model approach. *Physica A* 387(19-20), 4982–4984 (2008)
55. Van Mieghem, P., Ge, X., Schumm, P., Trajanovski, S., Wang, H.: Spectral graph analysis of modularity and assortativity. *Phys. Rev. E* 82(5), 056113 (2010)
56. Vázquez, A., Pastor-Satorras, R., Vespignani, A.: Large-scale topological and dynamical properties of the internet. *Phys. Rev. E* 65(6), 066130 (2002)
57. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. *Nature* 393(6684), 440–442 (1998)
58. White, H.C., Boorman, S.A., Breiger, R.L.: Social structure from multiple networks: I. block models of roles and positions. *Am. J. Sociol.* 81, 730–779 (1976)
59. Zachary, W.W.: An information flow model for conflict and fission in small groups. *J. Anthropol. Res.* 33(4), 452–473 (1977)