Eur. Phys. J. B **67**, 277–284 (2009)
DOI: 10.1140/epjb/e2008-00418-0

THE EUROPEAN PHYSICAL JOURNAL B

Regular Article

Detection of node group membership in networks with group overlap

E.N. Sawardecker[1,a], M. Sales-Pardo[1,2,3], and L.A.N. Amaral[1,2,b]

[1] Department of Chemical and Biological Engineering, Northwestern University, Evanston, IL 60208, USA
[2] Northwestern Institute on Complex Systems, Northwestern University, Evanston, IL 60208, USA
[3] Northwestern University Clinical and Translational Sciences Institute, Chicago, IL 60611, USA

Received 5 August 2008
Published online 20 November 2008 –

Abstract. Most networks found in social and biochemical systems have modular structures. An important question prompted by the modularity of these networks is whether nodes can be said to belong to a single group. If they cannot, we would need to consider the role of "overlapping communities." Despite some efforts in this direction, the problem of detecting overlapping groups remains unsolved because there is neither a formal definition of overlapping community, nor an ensemble of networks with which to test the performance of group detection algorithms when nodes can belong to more than one group. Here, we introduce an ensemble of networks with overlapping groups. We then apply three group identification methods – modularity maximization, k-clique percolation, and modularity-landscape surveying – to these networks. We find that the modularity-landscape surveying method is the only one able to detect heterogeneities in node memberships, and that those heterogeneities are only detectable when the overlap is small. Surprisingly, we find that the k-clique percolation method is unable to detect node membership for the overlapping case.

PACS. 89.75.Fb Structures and organization in complex systems

1 Introduction

Real-world networks including man-made and natural networks are strongly modular, that is, the pattern of connections among nodes is not homogeneous [1,2]. The modularity of a network is a consequence of the fact that there are groups of nodes in the network that preferentially connect to one another [1–8]. However, the assignment of nodes into those groups still remains a challenging task because, typically, nodes also connect to nodes that are not in their group [3,6,8]. Additionally, nodes may hold membership in more than one group [6,8–11], resulting in groups that "overlap."

The question of whether there are nodes that belong to more than one group and how important overlapping groups are to the network's organization is especially relevant in social and biochemical systems, in which typically nodes are thought to belong to more than one group. Consider, for instance, the network of scientific collaborations within an institution: people with joint appointments would be expected to appear in more than one group. Or, consider the network of physical interactions between proteins: topological modules are thought to have a strong correlation with biological function [12]. Since many proteins are known to have more than one function, one would expect these proteins to belong to more than one group.

Recently, methods to uncover the hierarchical organization of networks [13,14] have been proposed, opening the possibility of performing multi-scale analysis on social, biological, and economical systems for which large amounts of data are available. However, a potential caveat of these methods is that they do not take into account the fact that nodes could hold membership in more than one group. Modularity maximization methods [7,15,16], which have been successful at finding correlations between network function and structure, suffer from the same problem. *The impact of neglecting overlapping groups has not been assessed so far, since there is neither a formal definition of overlapping group nor a set of models on which to test overlap identification algorithms.*

Here, we introduce an ensemble of networks [17] that have overlapping groups by construction [11]. We then apply three different group detection methods – modularity maximization [1,15], k-clique percolation [8], and modularity-landscape surveying [13] – to these networks. We find that the modularity-landscape surveying method is the only one able to detect heterogeneities in node memberships, and that these heterogeneities are detectable provided the overlap is significantly smaller than the size of the modules involved.

[a] e-mail: e-sawardecker@northwestern.edu
[b] e-mail: amaral@northwestern.edu

2 Model networks

Consider a network comprised of N nodes and L edges. Let $\mathcal{N} = \{n_i\colon i = 1, \ldots, N\}$ be the set of nodes and $\mathcal{G} = \{g_x\colon x = a, b, \ldots\}$ be the set of groups in which the nodes can hold membership. Specifically, let $\mathcal{G}_i \subset \mathcal{G}$ be the set of groups in which node n_i holds membership. Without loss of generality, we assume that all groups with identical membership lists have been merged and that all groups have at least two members.

Here, we focus on the ensemble of random networks in which the probability p_{ij} of the edge (n_i, n_j) being present in a network is a function *solely* of the set of co-memberships of the two nodes $\mathcal{M}_{ij} = \mathcal{G}_i \cap \mathcal{G}_j$. We restrict our attention to the case where p_{ij} is a non-decreasing function of the cardinality of $\mathcal{M}_{ij}$. For undirected networks, this is the most plausible case. Indeed, most clustering algorithms used to investigate the modular structure of networks have at their basis the assumption that this is the appropriate case. Such an assumption is supported by the fact that those methods return plausible results for those networks [7,17–20].

We consider the case where $p_{ij} = p_l$, where $l = ||\mathcal{M}_{ij}||$ is the cardinality of $\mathcal{M}_{ij}$, and, further, we assume that $p_0 < p_1 \leq p_2 \leq \cdots$. An implication of our choice for p_{ij} is that if one selects a sub-set of $\mathcal{N}$ in which all pairs of nodes have non-empty co-membership sets, then there will be more edges connecting these nodes than one would expect to find by chance. Or, in other words, we expect to find more edges connecting these nodes than if all links had the same probability $p = 2L/N(N-1)$ of being present. In contrast, if one selects a sub-set of $\mathcal{N}$ in which all pairs of nodes have empty co-membership sets, then there will be fewer edges connecting these nodes than one would expect to find by chance. These facts directly suggest that the maximization of a modularity function such as that proposed by Newman and Girvan [1] will enable one to identify node membership in modular networks (see [13,21,22] for caveats to this argument).

The ensemble of networks we focus on comprises two distinct sub-ensembles. The first sub-ensemble, which we denote *transitive*, conforms to a transitive relationship among co-membership sets. That is, if the co-membership set $\mathcal{M}_{ij}$ is non-empty and the co-membership set $\mathcal{M}_{ik}$ is also non-empty, then the co-membership set $\mathcal{M}_{jk}$ must also be non-empty. The second sub-ensemble, which we denote *non-transitive*, does not conform to a transitive relationship among co-membership sets.

Networks in the transitive sub-ensemble have the property that every node must hold membership in only one group (if identical groups have been collapsed). In contrast, networks in the non-transitive sub-ensemble have some nodes that hold membership in more than one group. Most module detection algorithms in the literature deal only with the transitive sub-ensemble [7,15,20,23], that is, nodes are divided into "exclusive groups." A notable exception is the work of Palla et al. [8], which highlights the possibility that a network will contain nodes belonging to more than one group, thus allowing for "overlapping communities." Regretfully, Palla et al. [8] do not define ensembles of networks with overlapping groups.

Another significant exception is the work of Sales-Pardo et al. [13], which determines community structure even when hierarchical levels of structure exist. These hierarchical levels of structure indicate that nodes may belong to more than one group, but only when multiple layers are considered. Here, we define an ensemble of networks in which most nodes hold membership in a single group, while a small fraction of nodes hold membership in two or three groups (Fig. 1).

3 Community detection

3.1 Description of the methods

Let us now address the question of detectability of the memberships of individual nodes. Ideally, one wishes to detect all group memberships from the topology of the network alone. For the case of transitive networks, it has already been shown that when p_1, the probability that two nodes belonging to the same group are connected, is not much larger than p_0, it is impossible to extract the correct membership assignment from the network structure alone [15]. Here, we focus on the detection of node membership for the ensemble of non-transitive networks described above. We consider three different classes of group detection algorithms: modularity maximization [1,15], k-clique percolation [8], and modularity-landscape surveying [13].

Modularity maximization methods are the current "gold standard" for group identification [23]. In this approach, nodes are classified into groups that maximize the number of within group edges compared to those that would be expected from chance alone [1,2,4,5,7]. Some of the proposed algorithms, such as spectral decomposition, are extremely fast and can handle networks comprised of hundreds of thousands of nodes [7]. However, this approach is clearly geared toward networks with transitive membership structures since every node must be classified into a single group.

The k-clique percolation method introduced by Palla et al. [8] is based on the observation that networks sometimes contain connected cliques of the same size [8]. In this method, a group comprises chains of "adjacent" k-cliques – where two k-cliques are adjacent if they share $k-1$ nodes. A strength of this approach is that nodes can be classified into more than one group, making it a priori well-suited to investigate non-transitive networks. Two limitations of this this approach, however, are that different values of k will result in different group membership patterns, and that sparse networks might contain a very small number of cliques with $k > 2$.

The modularity-landscape surveying method [13], or MLS, is based upon the observation that the modularity landscape is very rugged and has many local maxima, which means that there are many partitions of nodes into groups characterized by high values of the modularity function. In analogy to disordered physical systems whose landscapes are also rugged [24–26], one expects that the

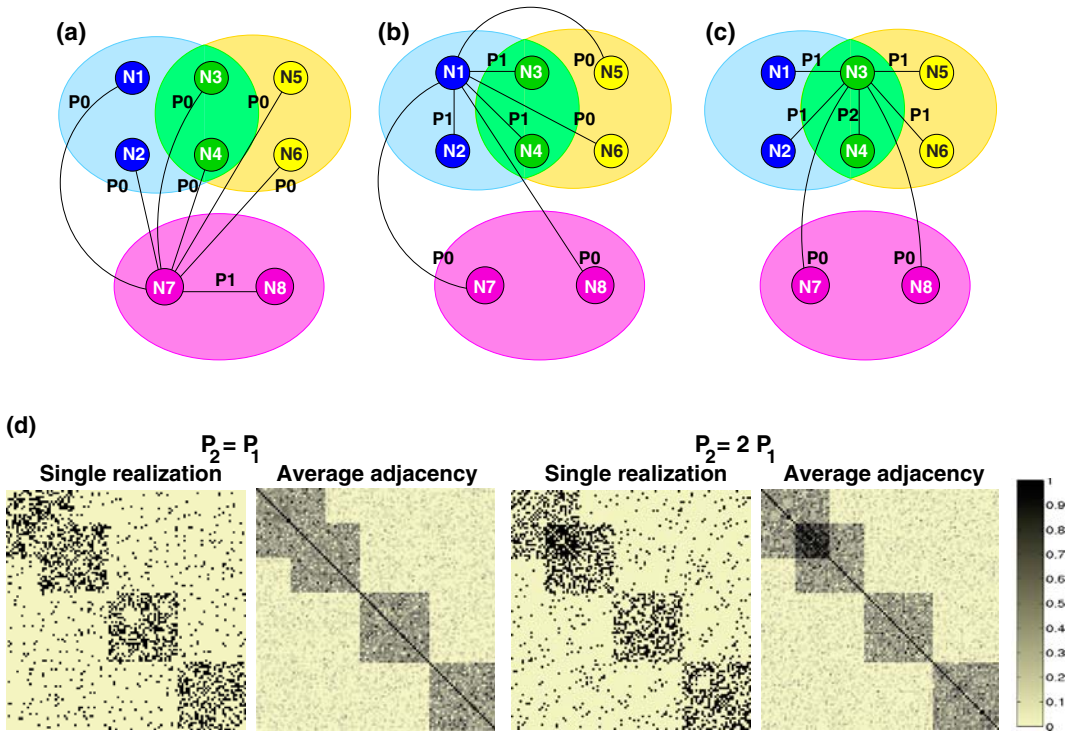

Fig. 1. (Color online) Random network models with overlapping groups. Consider a network with eight nodes, n_1 through n_8. We assign these nodes to three different groups which we indicate by three different colors, blue, yellow, and purple. Moreover, we allow the blue and yellow groups to overlap; nodes n_3 and n_4 belong to both yellow and blue groups. We highlight these overlapping nodes in green. (a), Purple nodes connect to other purple nodes with probability p_1 and to blue, green, and yellow nodes with probability $p_0 < p_1$. (b), Blue nodes connect with other nodes in the blue group (n_1, n_2, n_3, n_4) with probability p_1, and to nodes outside their group (yellow and purple nodes) with probability p_0. Note that yellow nodes have the same properties as blue nodes. (c), Green nodes connect to blue and yellow nodes with the same probability p_1. Since green nodes are not in the same group as purple nodes, the probability of having a connection between a green and purple node is p_0. For generality, we assume that green nodes connect between themselves with probability p_2. (d), Adjacency matrices for the model networks with $p_2 = p_1$ and $p_2 = 2p_1$. We show both the adjacency matrix for a single realization of the model network and the average adjacency matrix, that is the fraction of times two nodes are connected in the ensemble of networks defined by the model. Note that $p_2 = 2p_1$ has an overlapping region that is more densely connected, hence the average adjacency matrix has more "black." We show results for networks of 112 nodes divided into four groups of 32 nodes. Two of the four groups overlap by sharing 16 nodes and the corresponding edges. We select p_1 and p_0 such that for nodes not in the overlapping region, the average degree z is 16 and the ratio r between the external-degree and internal-degree is $r = 0.125$.

set of all local maxima conveys the relevant contribution to the system's physical properties. Thus, the method samples all the partitions $\mathcal{P}$ corresponding to local maxima with probabilities proportional to the size of their basins of attraction. Then, it builds a co-classification matrix $\mathbf{A}$, in which each element A_{ij} corresponds to the expected fraction of the time in which a pair of nodes (n_i, n_j) is classified in the same group. As discussed above, this method does not restrict nodes to hold membership in a single group, but rather, by collecting statistics, it reports the likelihood that two nodes are members of the same group (or sets of groups), and therefore it is a priori suitable to identify node memberships in networks with both transitive and non-transitive memberships.

3.2 Random network ensembles

In order to investigate the detectability of the membership structure of a network, we generate random networks in which nodes can belong to five groups, $\mathcal{G} = \{g_a, g_b, g_c, g_d, g_e\}$. We consider the cases in which most of the nodes belong to a single group, $\{g_a\}, \{g_b\}, \{g_c\}$, or $\{g_d\}$, and the remaining nodes belong to two groups, $\{g_a, g_b\}$, or to three groups, $\{g_a, g_b, g_e\}$. We then assume that if two nodes have membership in the same group, they will be linked with probability $p_1 > p_0$. Similarly, if two nodes have membership in the same two groups, they will be connected with probability $p_2 \geq p_1$. Note that as $r \equiv p_0/p_1$ approaches one, the harder it becomes to detect the co-membership structure of a network. Also, since some nodes belong to multiple groups, the degree of these overlapping nodes will be larger than that for a node belonging to only one group. Henceforth, we denote the average degree of a node belonging to only one group by z.

If a node holds membership in only one group, say g_a, then it belongs to group A. If some nodes hold membership in two groups $\{g_a, g_b\}$, then they belong to groups A and

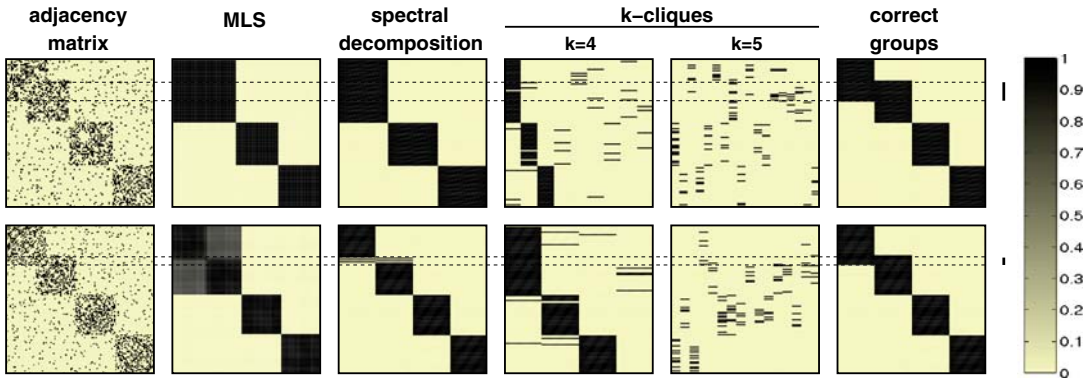

Fig. 2. (Color online) Output of methods for group detection. We apply three different group detection algorithms, the spectral decomposition method, the modularity-landscape surveying method, and the k-clique method to model networks with overlapping groups (Fig. 1). We show results obtained for a typical network realization for $p_2 = p_1$ and for two overlap sizes: large overlap ($s = 0.25$ – first row), and small overlap ($s = 0.125$ –second row). We generate these networks with the same parameters z and r that we use for the networks in Figure 1. The first column shows the adjacency matrix. The second column shows the co-classification matrix obtained with the modularity-landscape surveying method (see text). The third column displays the group classification matrix obtained with the spectral decomposition method. The next two columns show the group classification matrices obtained with the k-clique method for $k = 4$ and $k = 5$. The final column shows the expected group classification. The colorbar is the same as in Figure 1. In the group classification matrices, each row corresponds to a node, and each column corresponds to a group. If a node belongs to the group, a black band appears in that group's column. If a node belongs to more than one group, multiple bands will appear; these are nodes that the belong to the "overlap" between groups according to the algorithm. Dotted black lines indicate the nodes that by construction hold membership in more than one group and thus comprise the overlap in the model network. The modularity-landscape surveying method indicates a region of membership heterogeneity such that nodes in the overlapping groups have more in common with each other than with nodes in the non-overlapping groups (grey region), but this effect is dependent on the value of s. The spectral decomposition method yields three or four groups, depending on the value of s. The k-clique method yields at least four groups for $k \geq 4$, some of which only contain a few nodes.

B. If the latter case is true, then groups A and B overlap. We define the "overlap size" s as

$$s = \frac{||A \cap B||}{||A \cup B||}, \tag{1}$$

that is, the number of nodes in both A and B divided by the combined size of groups A and B. Thus, an important issue regarding co-membership detection when nodes hold membership in more than one group is how the size of the overlap affects the accuracy in detecting group membership.

3.3 Mutual information

To quantify the similarity between two partitions of nodes, we calculate the mutual information between the two partitions [23]:

$$M_I = \frac{-2\sum_{P,Q} N_{ij} \ln\left(\frac{N_{ij}N}{N_i N_j}\right)}{\sum_P N_i \ln\left(\frac{N_i}{N}\right) + \sum_Q N_j \ln\left(\frac{N_j}{N}\right)}, \tag{2}$$

where P is the list of groups in the first partition, Q is the list of groups in the second partition, N is the total number of nodes, N_i is the number of nodes in group g_i in the first partition, N_j is the number of nodes in g_j in the second partition, and N_{ij} is the number of nodes that are both in g_i and g_j. Note that this expression is symmetric; thus, it is an unbiased metric to compare the similarity of two partitions.

If the partitions are identical, $M_I = 1$, whereas if the two partitions are totally uncorrelated, $M_I = 0$. Note, however, that for the case in which each node is placed into a separate group, one has $M_I = M_I^* > 0$. We thus report $m_I = \frac{M_I - M_I^*}{M_I^*}$, so that values of m_I greater than zero indicate significant accuracy.

4 Results

To compare the performance of the methods for the ensemble of model networks with overlapping groups previously introduced, we generate ten networks for each set of parameter values and apply the three group detection algorithms to each network. Figure 2 displays the typical results obtained for a network with parameters $r = 0.125$, $z = 16$, $p_2 = p_1$ or $p_2 = 2p_1$, and $s = 0.125$ or $s = 0.25$.

To determine the accuracy of each method, we compare the partitions returned by each method to the known division of nodes into groups. Specifically, we use the normalized mutual information m_I, which quantifies the amount of information that two different partitions share [23]. Figure 3 displays the average m_I versus z for $p_2 = p_1$ and $s = 0.125$ or $s = 0.25$ and different values of r.

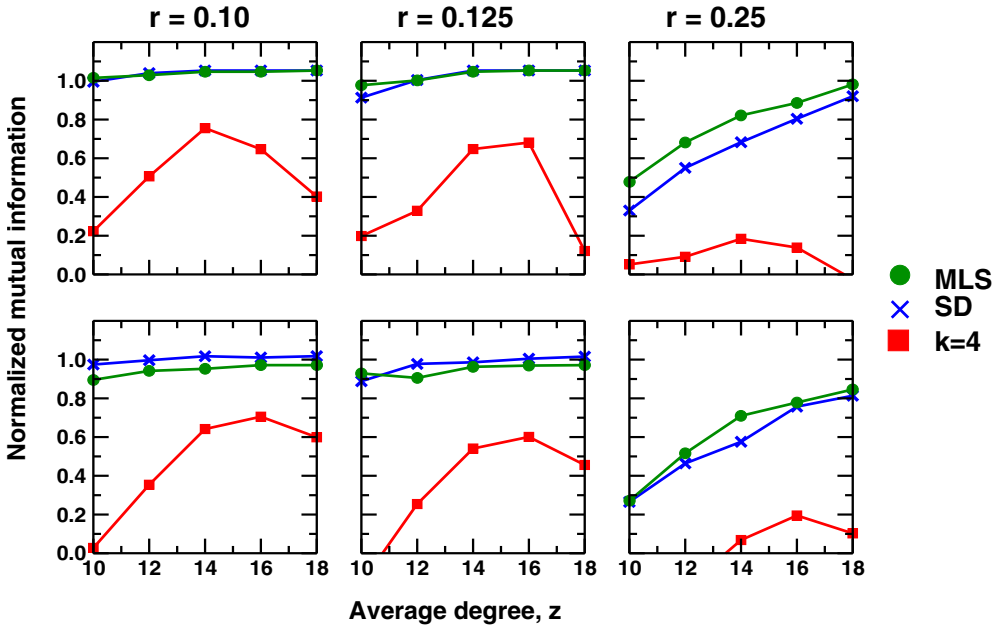

Fig. 3. (Color online) Performance of modularity and k-clique methods for group detection. We generate ten networks for each set of parameter values, z and r. We then compare the groups we obtain from the k-clique, spectral decomposition, and modularity-landscape surveying methods to the expected group assignment and compute m_I (see text). We show the normalized mutual information, m_I, versus average node degree z for networks with $p_2 = p_1$ and a large overlap (first row) and a small overlap (second row) (Fig. 1). For the modularity-landscape surveying method, we computed m_I with regards to the topmost level only. The first column of plots corresponds to $r = 0.1$, the second column to $r = 0.125$, and the third column to $r = 0.25$. Note that $m_I > 0$ indicates that the method considered yields significant information on the network's group structure. For the k-clique method, $m_I \leq 0$ indicates either the lack of detectable cliques of that size or the tendency to put all nodes in the same group. Note that for the spectral decomposition and modularity-landscape surveying methods, performance increases with increasing average node degree z and decreasing r.

Since the average degree of a node should strongly affect the ability of each group detection method to detect the known group structure, we systematically investigate degree effects. We expect that, as degree increases, the difficulty of detection should decrease. Also, as the number of nodes having membership to two or more groups increases, the difficulty of detection should increase.

4.1 Modularity maximization

The results obtained with the spectral decomposition method exhibit different behaviors depending on the size of the overlap: for small overlaps ($s = 0.125$), the method identifies four groups, whereas for large overlaps ($s = 0.25$), it identifies three groups, such that the two overlapping groups are combined into a single one, as predicted by Fortunato and Barthélemy [21]. Note that there are no significant differences between the cases $p_2 = p_1$ and $p_2 = 2p_1$.

4.2 Modularity-landscape surveying

In contrast, the modularity-landscape surveying method is able to uncover more information about the underlying organization of the nodes in the network than either the modularity maximization or k-clique percolation methods. Even for small overlaps, the algorithm is able not only to identify densely interconnected groups of nodes, but is also able to detect that the overlapping groups have more in common with each other than with the remaining groups (Fig. 2).

4.3 k-clique percolation

The results obtained with the k-clique method depend strongly on the value of k. For $k = 3$, the method is unable to detect the modular structure of the networks; it places all the nodes into a single group. For $k = 4$, the two overlapping groups are mostly combined into one group for both large and small overlaps. Finally, for $k = 5$, the algorithm does not identify any sizable group of nodes in the network. In fact, the signal provided by the k-clique method is weaker than that provided by the adjacency matrix. Surprisingly, even though the k-clique method allows nodes to belong to more than one group, *the nodes placed in multiple groups do not in general correspond to the nodes belonging to the overlapping groups* (Fig. 2).

We find that the accuracy of the k-clique method is always much smaller than that of the spectral decomposition and modularity-landscape surveying methods. In fact,

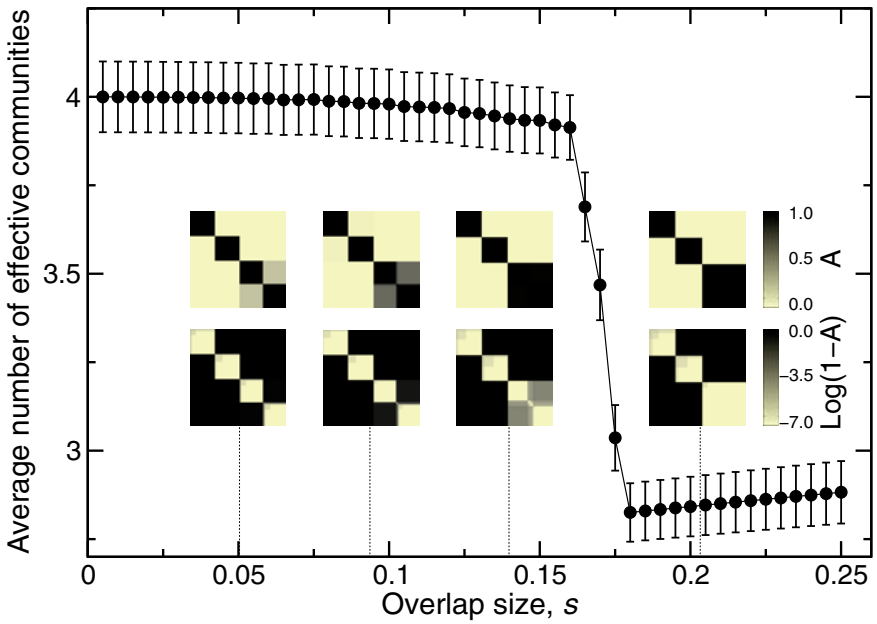

Fig. 4. (Color online) Effect of overlap size. We plot the average number of groups detected using the spectral decomposition method for modularity maximization (black circles) versus the overlap size s (Eq. (1)) for model networks with groups of 100 nodes, $z = 50$, $r = 0.1$, and $p_2 = 2p_1$. We show the average and standard error obtained for forty networks. We also show the co-classification matrix **A** obtained from the modularity-landscape surveying method for a typical network for four different s values. In the co-classification matrices, each row/column represents a node and each element A_{ij} corresponds to the fraction of the time two nodes are classified in the same partition for the maxima in the modularity landscape (see text). Each element is colored following the color code on the right hand side. To illustrate the differences for co-classification values close to one, we also plot $\log(1 - A_{ij})$. Again, matrix elements are colored following the color code on the right hand side. Note that while the spectral decomposition method can only detect whether the overlap is large enough to transition between four to three groups, the modularity-landscape surveying method is able to capture the greater affinity between the nodes in two of the groups in the organization of the network.

in order for the k-clique method to return results that are significant, one must have $r < 0.25$. Moreover, for $k = 4$, the accuracy of the method decreases as the density of edges increases.

For low edge densities, the network does not contain any 5-cliques, so the groups identified for $k = 5$ are unreliable. These results point to a severe limitation of the k-clique method: similar networks require different k values in order to yield meaningful results, and even when group detection is meaningful the method always performs significantly worse than modularity based methods.

4.4 Overlap detectability

These results suggest that the detection of overlapping groups may be essentially impossible when the overlap is large. However, for small overlaps, the modularity-landscape surveying method is able to detect heterogeneities in node group membership. The question that arises is thus how small should the overlap be in order to be detected and whether detection may ever be unambiguous. To answer this question, we analyze model networks with groups comprising 100 nodes, $z = 16$, $r = 0.1$, and $p_2 = 2p_1$, for a wide range of overlap sizes $s = 0.02, \ldots, 0.25$ (Figs. 4 and 6). The spectral decomposition method shows a transition from identifying four groups ($s < 0.175$) to identifying three groups ($s > 0.2$). For $s < 0.175$, the modularity-landscape surveying method is able to detect the signature of heterogeneities in node membership. However, the "signal" fades as s increases.

Like the spectral decomposition method, the modularity-landscape surveying method also indicates that there are three different groups for $s > 0.2$. Additionally, it is impossible to detect the overlap between groups from the collection of edges alone. The signal is only distinct for $s \leq 0.1$, and even then it is not clear whether one can distinguish between the case in which two groups overlap and the case in which a group comprises two sub-groups (Fig. 5) [13].

5 Conclusions

The ability to detect overlapping communities within real-world networks would greatly enhance understanding of phenomena such as synchronization [27]. However, our analysis reveals that the group detection methods in the literature are not entirely equipped to handle such information. In some cases, these methods may require tunable parameters, such as in the k-clique percolation method [8] and the method of Gfeller et al. [10].

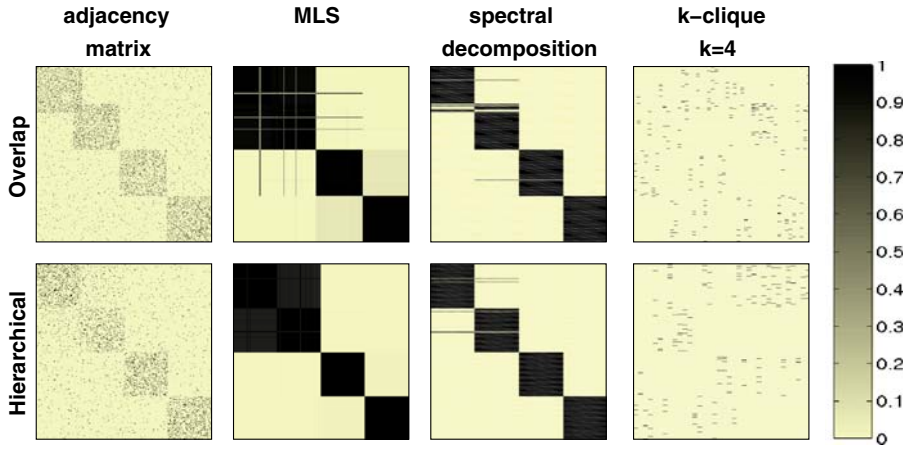

Fig. 5. (Color online) Comparison between overlapping networks and networks with a hierarchical structure. We compare results from applying the three different methods (spectral decomposition, k-clique percolation, and modularity-landscape surveying) to (i) networks with a small overlap, and (ii) networks with a hierarchical structure. We show typical results for (i) a network with overlapping groups of 100 nodes, $z = 16$, $r = 0.1$, $p_2 = 2p_1$, and $s = 0.1$ (top row – see Fig. 1 for details); (ii) a network with hierarchical structure such that at the top level there are three groups (two of 100 nodes and one of 200 nodes), and the largest group is comprised of two sub-groups of 100 nodes. We construct the latter network such that $z = 16$ and $r = 0.1$, for nodes in the groups with flat organization. For edges involving nodes inside the large group, we link nodes (i, j) with probability: (i) p_1 if they belong to the same sub-group, (ii) with $p_2 < p_1$ if they do not belong to the same sub-group, but belong to the same group at the top level; and, (iii) with p_0 if they do not belong to the same group at the top level. Note that p_0 is the same for any pair of edges running across groups, and that p_1 and p_2 are selected such that $z = 16$. We show results for the case $r_2 = p_2/p_1 = 1/3$. There are very few differences in the results for both networks. The spectral decomposition method finds four groups for both cases. The k-clique percolation method with $k = 4$ subtly outlines the underlying group structure of the network shown by the adjacency matrix. Results for $k = 5$ are not shown because there are no cliques of that size for sparse networks. Note that the signal detected by the k-clique percolation method has significantly decreased compared to that for the smaller networks in Figure 2. In contrast, the signal detected by the modularity-landscape surveying method has not decreased. Note that results for hierarchical and overlapping networks are very similar.

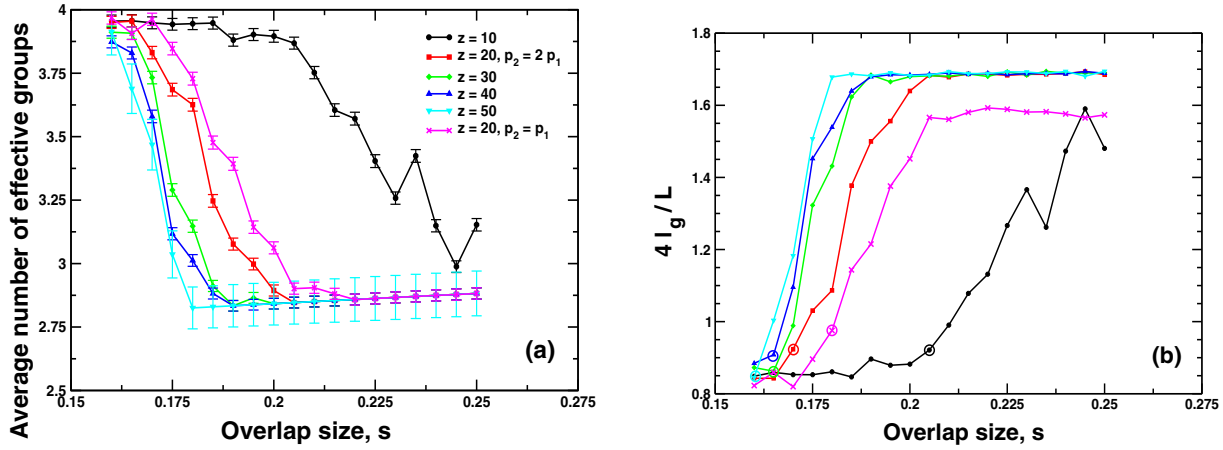

Fig. 6. (Color online) (a) This graph shows the average number of effective groups detected by spectral decomposition for a range of overlap sizes and for different average degrees. The average degrees investigated were $z = 10,\ 20,\ 30,\ 40,\ 50$. Forty networks were constructed for all cases except the $z = 50$ case, which only had ten networks for every overlap size. Each network consists of four groups, two of which share nodes. As the network becomes more dense with increasing degree, the transition from the detection of approximately four groups to three groups occurs abruptly at an overlap size of $s = 0.16$. (b) This graph shows the resolution of limit detectability. In the paper by Fortunato and Barthélemy, they determined that for a module to be detectable via spectral decomposition, the number of links within the module, l_g, should be less than $\frac{L}{4}$ where L is the total number of links in the network. We tested these conditions for several networks with overlapping groups by construction. The size of the overlap varied from $s = 0.16$ to $s = 0.25$. Finally, forty networks were generated for each set of parameters except for $z = 50$, in which ten networks were generated. In this case, we plot $4\frac{l_g}{L}$ for each overlap size and degree, and see that the transition of detectability occurs over the same range as is indicated in the average number of effective groups. The transition sites are circled.

The promising method of Nepusz et al. [11] aims to obtain the global organization of a network while determining which nodes act as "bridges" between communities. This method captures some of the same information as the modularity-landscape surveying method, but requires additional centrality calculations to correct for the "bridgeness" score. Furthermore, even the recently proposed modularity-landscape surveying method, which can detect small overlaps, is not able to unambiguously differentiate overlapping groups from hierarchically-organized groups.

We thank R. Guimerà, M.J. Stringer, and M.E.J. Newman for comments. L.A.N.A. gratefully acknowledges the support of the Keck Foundation, an NIH/NIGMS K-25 award, and of NSF.

Appendix A

To characterize how the degree and the size of the overlap affect group detection, we generate forty networks composed of four groups of 100 nodes each for every set of conditions tested. We tested the degree of a non-overlapping node at values of $z = 10, 20, 30, 40, 50$, and we tested the size of the overlap, s, for $s = 0.16$ through $s = 0.25$. Each of these networks was generated with $p_2 = 2p_1$, except for the case $z = 20$, for which we also studied networks with $p_2 = p_1$.

Since the spectral decomposition method is very fast and is considered the gold standard for group detection, we applied this method on each of the networks generated (Fig. 6a). We expected the spectral decomposition method to detect four groups for low overlap sizes for every degree tested, and that it would detect three groups at higher overlap sizes or at higher node degree. We expected also that the higher the degree, the faster the transition between four detectable groups and three detectable groups. Since some of the groups reported by the method were very small, we calculated the effective number of groups, $N_{effective}$

$$N_{effective} = \frac{1}{\sum_{\mathcal{G}} (\frac{S_i}{N})^2} \tag{A.1}$$

where $\mathcal{G}$ is the list of groups in the partition of the network returned by the method, S_i is the number of nodes in a group within $\mathcal{G}$, and N is the total number of nodes in the network. For the networks with an average degree of $z = 20$ and $p_2 = p_1$, we find a transition from detecting four groups to detecting three groups for $s \approx 0.18$, while for $z = 50$, the transition occurs by $s = 0.16$.

However, we wanted to further investigate the detectability for the degrees and overlap sizes chosen. Specifically, we examine the detection resolution limits as outlined by Fortunato and Barthélemy in [21]. In their paper, Fortunato and Barthélemy indicate that in order for a group to be unambiguously detectable by spectral decomposition, it must meet two criteria: (i) the number of links within group g, or l_g, should be less than the total number of links L divided by four, or $l_g < \frac{L}{4}$; and (ii) the ratio of links leaving the module, l_g^{out}, to links within the module should be less than two, or $a = \frac{l_g^{out}}{l_g} < 2$. All of the detected groups for each of the networks satisfied the second condition. Figure 6 shows $4 \times \frac{l_g}{L}$ versus s for each degree. In this figure, a value less than one indicates that the resolution detection limit is satisfied. Comparing Figures 6a and 6b, we see that as soon as the resolution detection limit is violated, the number of detected groups decreases, so our results are consistent with those of Fortunato and Barthélemy [21].

References

1. M.E.J. Newman, M. Girvan, Phys. Rev. E **69**, 026113 (2004)
2. R. Guimerà, L.A.N. Amaral, J. Stat. Mech.: Theor. Exp. P02001 (2005)
3. D.J. Watts, P.S. Dodds, M.E.J. Newman, Science **296**, 1302 (2002)
4. L. Donetti, M.A. Muñoz, J. Stat. Mech.: Theor. Exp. P10012 (2004)
5. R. Guimerà, M. Sales-Pardo, L.A.N. Amaral, Phys. Rev. E **70**, 025101 (2004)
6. J. Reichardt, S. Bornholdt, Phys. Rev. Lett. **93**, 218701 (2004)
7. J. Duch, A. Arenas, Phys. Rev. E **72**, 027104 (2005)
8. G. Palla, I. Derényi, I. Farkas, T. Vicsek, Nature **435**, 814 (2005)
9. J. Baumes, M. Goldberg, M. Magdon-Ismail, Efficient Identification of Overlapping Communities, in *Intelligence and Security Informatics: IEEE International Conference on Intelligence and Security Informatics, ISI 2005*, edited by P. Kantor, G. Muresan, F. Roberts, D. Zeng, F.Y. Wang, H. Chen, R. Merkle, IEEE (Springer, Berlin, 2005), Vol. 3495 of Lecture Notes in Computer Science, pp. 27–36
10. D. Gfeller, J.C. Chappelier, P.D.L. Rios, Phys. Rev. E **72**, 056135 (2005)
11. T. Nepusz, A. Petróczi, L. Négyessy, F. Bazsó, Phys. Rev. E **77**, 016107 (2008)
12. J.B. Pereira-Leal, A.J. Enright, C.A. Ouzounis, Proteins **54**, 49 (2004)
13. M. Sales-Pardo, R. Guimerà, A.A. Moreira, L.A.N. Amaral, Proc. Natl. Acad. Sci. USA **104**, 15224 (2007)
14. A. Clauset, C. Moore, M.E.J. Newman, Nature **453**, 98 (2008)
15. R. Guimerà, L.A.N. Amaral, Nature **433**, 895 (2005)
16. M.E.J. Newman, Proc. Natl. Acad. Sci. USA **103**, 8577 (2006)
17. M.E.J. Newman, Phys. Rev. E **68**, 026121 (2003)
18. R. Guimerà, S. Mossa, A. Turtschi, L.A.N. Amaral, Proc. Natl. Acad. Sci. USA **102**, 7794 (2005)
19. M.B. Hastings, Phys. Rev. E **74**, 035102 (2006)
20. M.E.J. Newman, Phys. Rev. E **74**, 036104 (2006)
21. S. Fortunato, M. Barthélemy, Proc. Natl. Acad. Sci. USA **104**, 36 (2007)
22. J.M. Kumpula, J. Saramäki, K. Kaski, J. Kertész, Eur. Phys. J. B **56**, 41 (2007)
23. L. Danon, A. Díaz-Guilera, J. Duch, A. Arenas, J. Stat. Mech.: Theor. Exp. P09008 (2005)
24. F.H. Stillinger, T.A. Weber, Phys. Rev. A **25**, 978 (1982)
25. F.H. Stillinger, T.A. Weber, Science **225**, 983 (1984)
26. F.H. Stillinger, T.A. Weber, Phys. Rev. B **31**, 5262 (1985)
27. A. Arenas, A. Díaz-Guilera, C.J. Pérez-Vicente, Phys. Rev. Lett. **96**, 114102 (2006)