Eur. Phys. J. B **72**, 671–677 (2009)
DOI: 10.1140/epjb/e2009-00397-6

THE EUROPEAN
PHYSICAL JOURNAL B

Regular Article

Comparison of methods for the detection of node group membership in bipartite networks

E.N. Sawardecker[1,a], C.A. Amundsen[2], M. Sales-Pardo[1,3,4], and L.A.N. Amaral[1,3,5,b]

[1] Department of Chemical and Biological Engineering, Northwestern University, Evanston, IL 60208, USA
[2] Department of Chemical and Biological Engineering, University of Wisconsin – Madison, Madison, WI 53706, USA
[3] Northwestern Institute on Complex Systems, Northwestern University, Evanston, IL 60208, USA
[4] Northwestern University Clinical and Translational Sciences Institute, Chicago, IL 60611, USA
[5] HHMI, Northwestern University, Evanston, IL 60208, USA

Received 29 May 2009 / Received in final form 5 October 2009
Published online 25 November 2009 – © EDP Sciences, Società Italiana di Fisica, Springer-Verlag 2009

Abstract. Most real-world networks considered in the literature have a modular structure. Analysis of these real-world networks often are performed under the assumption that there is only one type of node. However, social and biochemical systems are often bipartite networks, meaning that there are two exclusive sets of nodes, and that edges run exclusively between nodes belonging to different sets. Here we address the issue of module detection in bipartite networks by comparing the performance of two classes of group identification methods – modularity maximization and clique percolation – on an ensemble of modular random bipartite networks. We find that the modularity maximization methods are able to reliably detect the modular bipartite structure, and that, under some conditions, the simulated annealing method outperforms the spectral decomposition method. We also find that the clique percolation methods are not capable of reliably detecting the modular bipartite structure of the bipartite model networks considered.

PACS. 89.75.Fb Structures and organization in complex systems

1 Introduction

Real-world networks including man-made and natural networks are strongly modular; in other words, the pattern of connections among nodes is not homogeneous [1,2]. In some instances, 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]. The ability to detect these homophilic groups is an important task as the modular structure can affect the dynamics of the system [9,10]. Furthermore, each module may possess different structural properties, and thus global average network properties may misrepresent the structure of the system [11]. Although community identification in unipartite networks is now well understood, a thorough analysis of the equivalent problem for bipartite networks has not been made yet.

The question of detecting the organization of a bipartite network is especially relevant in social and biochemical systems, in which nodes often can be said to come from two distinct groups. Consider, for instance, scientific collaboration networks: one set of nodes includes all authors, while the other set of nodes comprises the set of papers [12–15]. The edges, therefore, connect authors to their publications, and the modular structure relates to the communities of collaborators (for the author node set) and communities of research topics (for the paper node set). Similarly, protein-protein interaction networks reflect the physical binding interactions between "bait" proteins and "library" proteins [16–19]. Modules, then, might indicate groups of functionally similar proteins.

Here, we test four different group detection methods in bipartite networks – modularity maximization via simulated annealing [11,20], modularity maximization via spectral decomposition [21,22], k-clique percolation [8], and biclique percolation [23] – to ensembles of modular random bipartite networks. We use the mutual information between method-generated partitions and the original partition to systematically quantify the accuracy of the four group detection methods. We find that the modularity-maximization methods are the only ones that reliably detect node membership in these bipartite networks.

The organization of this paper is as follows: in Section 2 we describe the modular random bipartite networks; in Section 3 we describe the four different community detection methods for bipartite networks analyzed in this paper and review the definition of the mutual information function, which we use to assess method performance. In

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

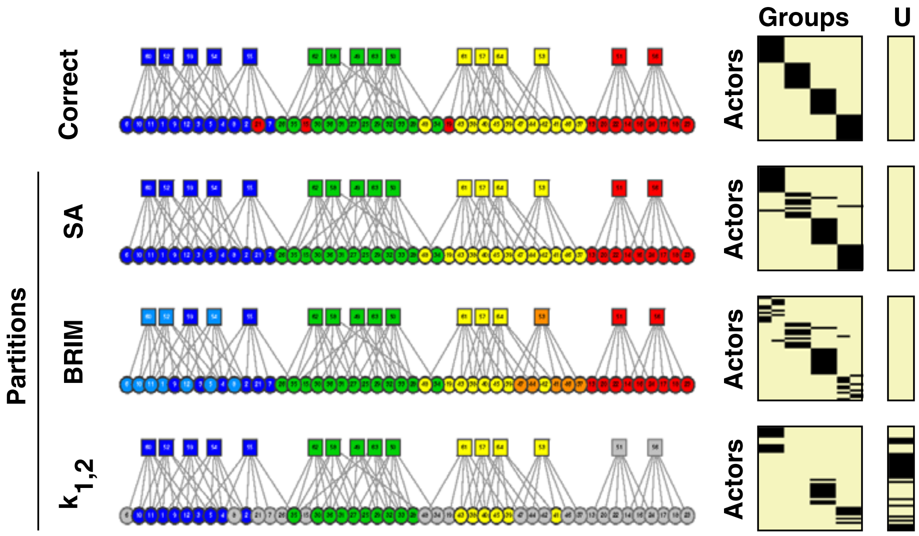

Fig. 1. (Color online) Community detection in bipartite networks. We create a bipartite network with 16 teams and 48 actors divided equally among 4 "colors". We set $m_a = 6$ and $p = 0.9$. The different panels show the network partitions according to the true colors of actors and teams (top row). The second row shows the results for the simulated annealing modularity maximization (SA), the third row gives the structure for the BRIM algorithm (BRIM), and the fourth row gives the results for the biclique percolation (Biclique) method with $k_{a,b} = k_{1,2}$. The second column shows the corresponding group membership matrices for the actor structure. The third column shows nodes that were not classified into a group. The BRIM method subdivides two of the modules while otherwise obtaining the correct result. The biclique percolation method, shown here for $k_{a,b} = k_{1,2}$, fails to classify almost half of the actors into a group (shown in grey).

Section 4 we present our results and in Section 5 we discuss the implications of our work.

2 Model networks

We generate modular random bipartite networks according to the model proposed by Guimerà et al. [11]. For simplicity and consistency with the nomenclature in the literature, we denote the two sets of nodes in the bipartite network as the set of *actors* and the set of *teams*. Given a bipartite network, we are interested in identifying groups (modules) of actors that are strongly connected to each other through co-participation in many teams.

We start by partitioning N actors onto N_M modules (Fig. 1). We assign the "color" c_s to the S_s nodes in module s. We then create N_T teams. Team a is assigned a color c_a and a team size m_a. For each of the m_a positions in team a, we select an actor at random with probability p from module s such that $c_s = c_a$; with probability $1-p$ we select an actor at random from modules for which $c_s \neq c_a$. We refer to p as the team homogeneity. When $p = 1$, all the members of the team have the same color c_a.

3 Community detection

We next address the question of detectability of the membership of individual nodes. Ideally, one wishes to detect all group memberships from the topology of the network alone. We consider two methods each within two classes of group detection algorithms: modularity maximization via simulated annealing [11,20] and via spectral decomposition [21,22], and clique percolation via k-clique cluster formation [8] and biclique cluster formation [23].

3.1 Description of the methods

Modularity maximization methods are the current "gold standard" for group identification in unipartite networks [10,24,25]. In these approaches, nodes are classified into groups that maximize the number of edges within the group compared to the total number of edges than can be formed from the same set of nodes [1,2,4,5,7]. Some of the proposed algorithms, such as spectral decomposition, can analyze networks comprised of hundreds of thousands of nodes [7]. Due to the fact that spectral decomposition methods return a local maximum, the identity of which is dependent on algorithm initialization, they are frequently less accurate than the slower simulated annealing approach [21]. The bipartite recursively induced modules (BRIM) method employs a strictly local search method; the solution found is not guaranteed to be globally optimal. A different initialization could lead to a solution with a higher modularity than the solution that was found. The initialization used for this study was to create two groups, and then to randomly and evenly divide the nodes among groups as they were added.

The clique percolation methods are based on the observation that networks sometimes contain connected cliques of the same size [8]. In this method, a group comprises clusters of "adjacent" cliques – two k-cliques are adjacent if they share $k-1$ nodes. It was initially reported that

k-clique percolation is applicable not only for networks in which nodes may belong to multiple groups, but also for bipartite networks when one considers the projection of the bipartite graph onto one set of nodes, either the actors or teams [8].

We have recently demonstrated that k-clique percolation often does not assign nodes to any group or assigns nodes to the incorrect group [25]. For example, sparse networks might contain a very small number of cliques with $k > 2$, preventing the k-clique method from classifying any nodes. Moreover, different values of k result in different group membership patterns, and it is not clear how to select the k value that best reveals the network structure [25].

Nonetheless, we study the k-clique method here because bipartite networks will, by construction, lead to projections onto a single set of nodes with large numbers of cliques. Moreover, the team size suggests a natural value for k: $k \geq m_a$.

The biclique percolation method is similar to the k-clique percolation method in that it constructs communities from adjacent cliques, but it can operate on the original set of nodes within the bipartite network [23]. Formally, a biclique, denoted by $k_{a,b}$, is adjacent to another $k_{a,b}$ biclique if they share at least $a-1$ actors and $b-1$ teams [23]. Biclique group assignments are not necessarily symmetric; in other words, the modules obtained using $k_{a,b}$ are not necessarily the same as those obtained using $k_{b,a}$. This difference in structure reflects the different connectivity patterns that could exist for actors and teams.

The limitations of biclique percolation are the same as for the k-clique percolation method, namely, that there is no clear criterion for selecting a and b, that different values of $k_{a,b}$ result in different group membership patterns, and that sparse networks might contain a very small number of cliques with $k_{a,b} > k_{2,1}$.

3.2 Method accuracy

To quantify the similarity between two partitions of nodes, we calculate the mutual information between the two partitions [24]:

$$MI = \frac{-2 \displaystyle\sum_{i \in P, j \in Q} N_{ij} \ln\left(\frac{N_{ij} N}{N_i N_j}\right)}{\displaystyle\sum_{i \in P} N_i \ln\left(\frac{N_i}{N}\right) + \sum_{j \in Q} N_j \ln\left(\frac{N_j}{N}\right)}, \qquad (1)$$

where P is the set of groups in the first partition, Q is the set 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 group 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 equation (1) is symmetric; thus, it is an unbiased metric to compare the similarity of two partitions. If the partitions are identical, $MI = 1$, whereas if the two partitions are totally uncorrelated, $MI = 0$.

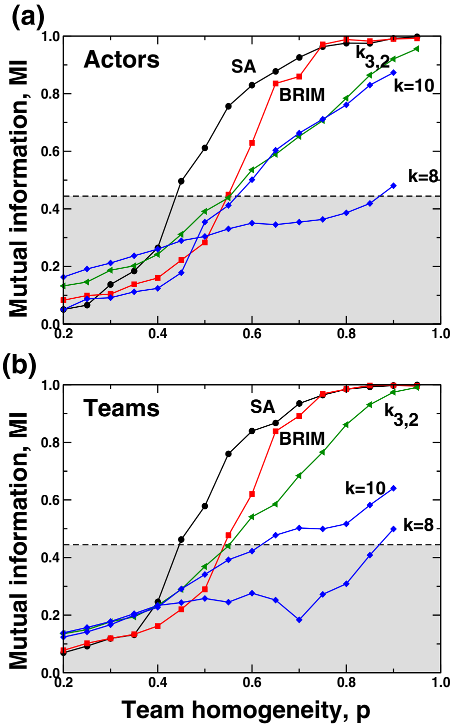

Fig. 2. (Color online) Effect of team homogeneity on algorithm accuracy. Networks are generated with team size $m_a = 8$. The dotted line is the mutual information for a network in which there is one community for each actor (team); therefore, any partition that results in a mutual information value above the dotted line reflects meaningful identification of community structure. (a) Mutual information for actor partitions. The k-clique percolation method was run on the projection of the network onto the set of actors. Our results support the idea that for the k-clique, a natural rule of thumb to select a value of k for the actor modules is to set $k \geq m_a$. (b) Mutual information for team partitions. Here, the k-clique percolation method was run on the projection of the network onto the set of teams. No natural rule of thumb for the k-clique percolation method presents itself for the analysis of teams. The k-clique percolation method increases in accuracy with increasing clique size, but it never reaches the accuracy of the modularity maximization methods.

4 Results

We compare the performance of the methods for the ensemble of modular random bipartite networks previously introduced. To this end, we first determine the accuracy of each method by calculating the mutual information of the partitions returned by each method and the known division of nodes into groups. In addition, we plot the group membership matrices G, in which the actors are the rows of the matrix and the modules are the columns [25]. Elements of the group membership matrix are shown in

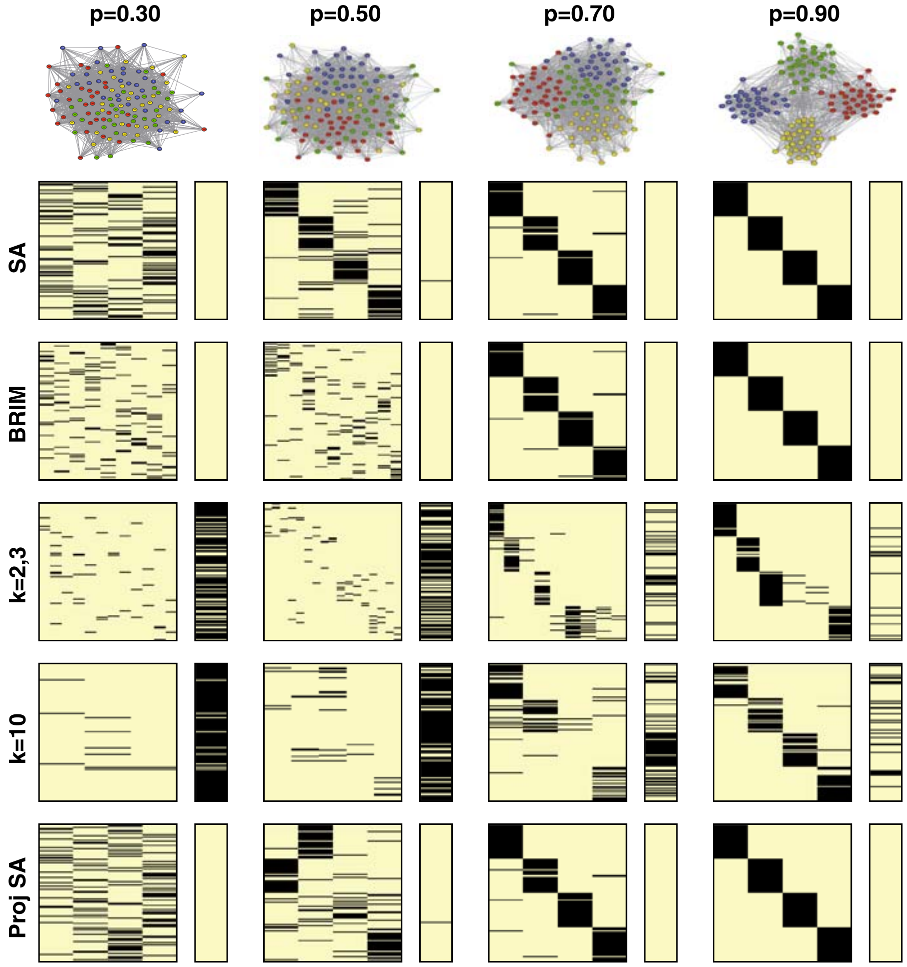

Fig. 3. (Color online) Group membership matrices reflect dependence on level of team homogeneity. Each group membership matrix shown here represents the calculated membership of nodes for networks generated with $p = 0.30, 0.50, 0.70$, and 0.90, and with team size $m_a = 8$. The first row shows the network projected on the set of actors, where actors are color coded by group. The second row shows the results for the simulated annealing modularity maximization (SA), the third row gives the structure for the BRIM algorithm (BRIM), the fourth row gives the results for the biclique percolation (Biclique) method with $k_{a,b} = k_{2,3}$, and the fifth row contains the results from k-clique percolation with $k = 10$. The k-clique percolation results correspond to the structure of the projection of network edges on the set of actors as determined by k-clique percolation. As a comparison, the final row shows the results from running a unipartite simulated annealing modularity maximization on the same projections (Proj SA). Note the degree to which the modularity maximization methods (SA, BRIM, and Proj SA) outperform the clique percolation methods.

black if an actor belongs to that module, and are shown in yellow otherwise.

4.1 Effect of team homogeneity

We studied the accuracy of the four methods as a function of p for $m_a = 8$ and $m_a = 14$. For $m_a = 14$ the projected networks were so dense that the k-clique percolation method had difficulty detecting all the cliques, much less the final structure.

Actor modules. As shown in Figure 2a, the SA method performs well for $p > 0.5$, while the BRIM method yields meaningful results for $p > 0.6$. The biclique method performs adequately only for $p > 0.7$ and $k_{a,b} = k_{3,2}$, as does the k-clique percolation for $k = 10$. Notably, the clique percolation methods never reach the level of accuracy obtained with the modularity maximization for the same range of parameter values.

Because the k-clique percolation method was the only method applied to the network projection, we ran a unipartite modularity maximization method (Proj SA) on the network projection for comparison (Fig. 3, last row). For large p, we find that the unipartite modularity maximization is nearly as accurate as the bipartite modularity maximization methods.

Team modules. The SA and BRIM methods perform well for $p > 0.5$ and $p > 0.6$, respectively. The biclique method begins to yield meaningful partitions for

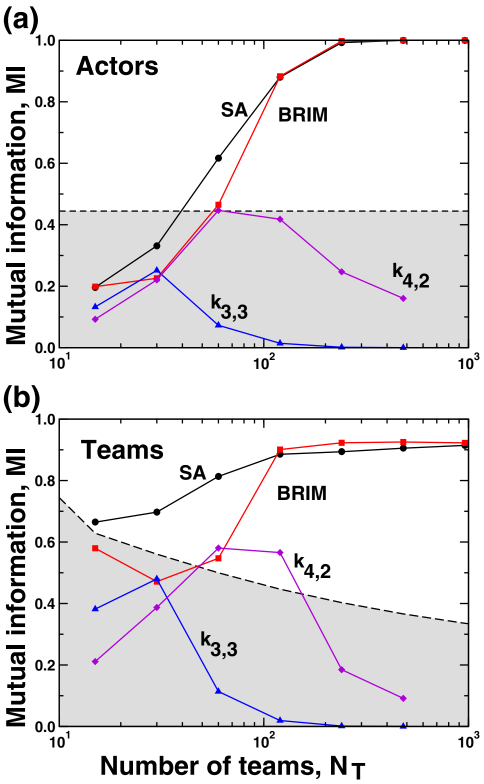

Fig. 4. (Color online) Effect of number of teams on algorithm accuracy. The dotted line is the mutual information for a network in which there is one community for each actor (team); therefore a partition that results in a mutual information value above the dotted line reflects meaningful community structure. (a) Mutual information for actor partitions. (b) Mutual information for team partitions. Note that the accuracy of the biclique method decreases as more information–that is, data on more teams–becomes available.

$p > 0.75$ and $k_{a,b} = k_{3,2}$, but never reaches the accuracy of the modularity maximization methods. Remarkably, the k-clique method works even worse for teams than for actors, perhaps because team clusters are highly heterogeneous (Fig. 2b).

4.2 Effect of number of teams

We next investigated the performance of the community detection algorithms as a function of N_T (Fig. 4). We considered the case $N_M = 4$, $S_S = 32$ for all s, $m_a = 14$, and $p = 0.5$.

Actor modules. For the partition of actors, both the SA and BRIM methods yield meaningful groups of the actors for $N_T = 60$, but the SA method outperforms BRIM for $N_T < 120$. Surprisingly, the accuracy of the biclique method decreases as the number of teams increases. For $N_T = 960$, the biclique percolation method is no longer able to uncover the community structure of the networks.

Team modules. The mutual information for the groups of teams suggests that the SA method detects meaningful team structure for networks with $N_T \geq 30$. The BRIM method only begins to detect meaningful team structure for $N_T = 120$, at which point it performs almost as accurately as the SA method. Again, the biclique method performs poorly, and can only detect some team structure for $N_T = 60, 120$ for $k_{a,b} = k_{4,2}$.

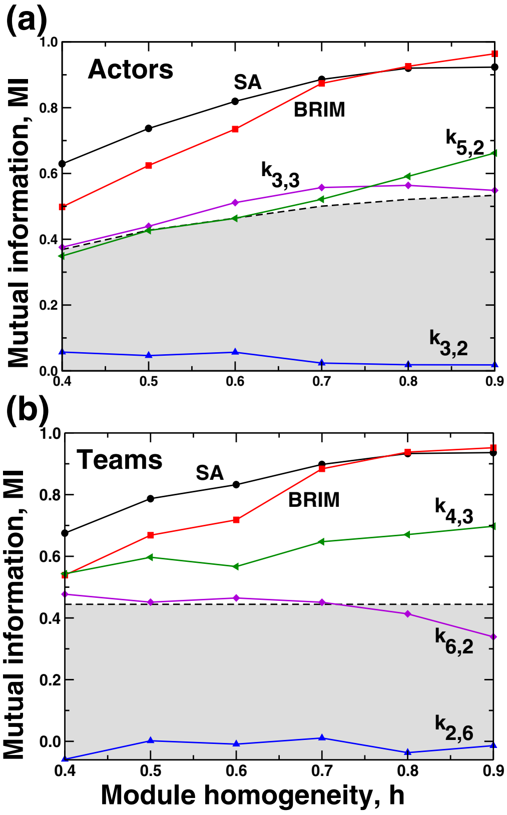

Fig. 5. (Color online) Effect of module size homogeneity on algorithm accuracy. The dotted line is the mutual information for a network in which there is one community for each actor (team); therefore a partition that results in a mutual information value above the dotted line reflects meaningful community structure. There are $N_M = 6$ modules per network, and sizes are determined by h, the module size homogeneity. (a) Mutual information for actor partitions. (b) Mutual information for team partitions. Note that the biclique method has difficulty detecting meaningful actor partitions. The biclique method is able to yield meaningful team partitions for $k_{a,b} = k_{4,3}$, although it is not as accurate as the SA or BRIM methods for $h > 0.4$.

4.3 Effect of module size homogeneity

Next, we investigated the accuracy of the three bipartite community detection algorithms for $N_M = 6$ and for different values of the module size homogeneity h. The module size homogeneity was calculated by first ordering the modules according to size, $S_1 \geq S_2 \geq ... \geq S_{N_M}$, and then taking the ratio of consecutive module sizes: $h = \frac{S_{i+1}}{S_i} < 1$. We considered the case $N_T = 128$, $m_a = 14$

and $p = 0.5$.

Actor modules. The SA and BRIM methods significantly outperform the biclique method for the actor partitions for all $h \in [0.4, 0.9]$. For $h \ll 1$, the SA method is the most accurate for both the set of actors and the set of teams (Fig. 5). The biclique method only yields meaningful groups for the set of actors for $h > 0.8$ and for $k_{a,b} = k_{5,2}$.

Team modules. The mutual information for the groups of teams reveals that the SA and BRIM methods again return meaningful group structure for all $h \in [0.4, 0.9]$. The biclique method also returns meaningful team structure for $h \in [0.4, 0.9]$ and for $k_{a,b} = k_{4,3}$. However, the accuracy of the biclique method is consistently less than the accuracy of the SA and BRIM methods, for both actors and teams.

4.4 Effect of heterogeneity of team sizes

Real-world networks are typically inhomogeneous in both team and module composition. For example, a scientific paper can have anywhere from one to several hundred authors [15]. We therefore test the detection ability of the SA and BRIM methods for the case when there is a distribution of team sizes within the network. We generate modular random bipartite networks with $N_M = 4$, $S_s = 32$, $N_T = 128$, $h = 1.0$, and $p = 0.5$. We draw team sizes from a geometric distribution with mean team size μ, which is the discrete counterpart to an exponential distribution [11]. When the team sizes are varied within the network, the SA method more accurately detects the structure for both actors and teams than the BRIM method (Fig. 6).

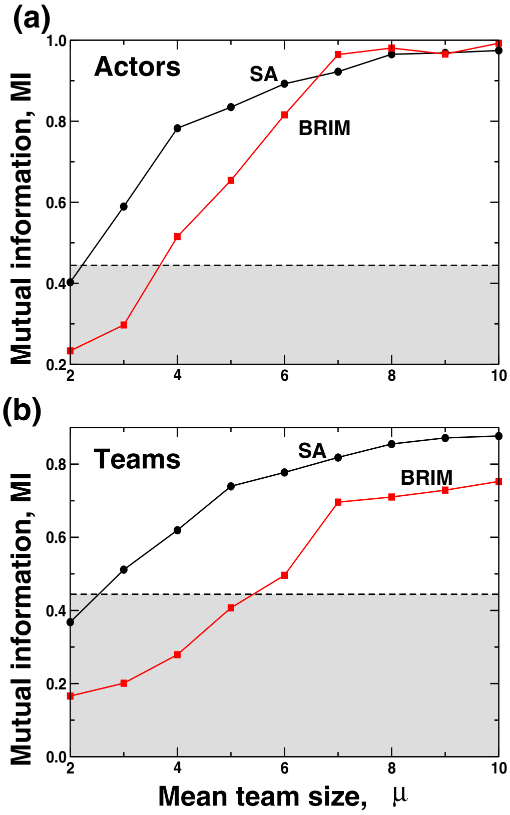

Fig. 6. (Color online) Effect of heterogeneity of team sizes on algorithm accuracy. Here, networks are generated with a geometric distribution of team sizes based on the mean team size, μ. The dotted line is the mutual information for a network in which there is one community for each actor (team); therefore, a partition that results in a mutual information value above the dotted line reflects meaningful community structure. (a) Mutual information for actor partitions. (b) Mutual information for team partitions. Note that the simulated annealing results are more accurate than the BRIM method for these networks.

5 Conclusions

Our analysis strongly suggests that modularity maximization methods are also the gold standard for community detection in bipartite networks.

While the biclique and k-clique methods both accurately detect regions of high clustering, these regions are very localized and do not reliably convey information about the global organization of the bipartite network. In addition, their tunable parameters of clique sizes are not straightforward to interpret, thus selecting the “most” accurate community structure is only possible when one knows the correct answer a priori.

Interestingly, the k-clique method better detects the structure of the projection of the bipartite network than it detects the structure of the overlapping unipartite networks for which it was designed [25]. Our analysis shows that the k-clique method is most accurate for $k = 10$, which is only slightly larger than the team size, $m_a = 8$. In other words, bipartite networks with uniform team sizes may have a natural range of k values that better uncover the underlying structure.

Both modularity maximization methods, SA and BRIM, display great accuracy in determining the modular structure under a broad range of conditions. Since both methods are based on modularity maximization, one expects that the detection of communities is also affected by a resolution limit as it is in unipartite graphs [26]. We find, however, that the limit in the detection is a much weaker condition in the case of bipartite graphs. Specifically, the limit derived in [26] for unipartite graphs depends on the number of links, while the analogous limit derived for the SA approach to bipartite networks depends instead on the number of teams. The greatest advantage of BRIM over SA is speed. The SA method is too time consuming to allow the study of networks with $N_T > 400$, whereas BRIM can quickly determine the modular structure of networks with $N_T \gg 100$. A weakness of the BRIM method is its sensitivity to initialization specifications, such as the number of nodes it randomly assigns to each module at the beginning of each iteration. Moreover, the BRIM method is not as accurate as the SA method for more challenging cases, such as when networks are generated with few teams or a wide distribution of team sizes.

We thank R. Guimerà, R.D. Malmgren, M.J. Stringer, J. Duch, S.M.D. Seaver, I.M. Sirer, A.L. Salazar, X. Zhang, A. Campanharo, P. Winter and P.D. McMullen for comments. L.A.N.A. gratefully acknowledges the support of the Keck Foundation, P50, NSF SciSIP award OP030388. M.S.-P. gratefully acknowledges the support of NSF SciSIP award OP030388 and of NIH CTSA grant 1 UL1 RR025741.

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. R. Guimerà, S. Mossa, A. Turtschi, L.A.N. Amaral, Proc. Natl. Acad. Sci. USA **102**, 7794 (2005)
10. M. Sales-Pardo, R. Guimerà, A.A. Moreira, L.A.N. Amaral, Proc. Natl. Acad. Sci. USA **104**, 15224 (2007)
11. R. Guimerà, M. Sales-Pardo, L.A.N. Amaral, Nature Phys. **3**, 63 (2007)
12. M.E.J. Newman, Proc. Natl. Acad. Sci. USA **98**, 404 (2001)
13. K. Börner, J.T. Maru, R.L. Goldstone, Proc. Natl. Acad. Sci. USA **101**, 5266 (2004)
14. R. Guimerà, B. Uzzi, J. Spiro, L.A.N. Amaral, Science **308**, 697 (2005)
15. M.J. Stringer, M. Sales-Pardo, L.A.N. Amaral, PLoS ONE **3**, e1683 (2008)
16. P. Uetz, L. Giot, G. Cagney, T.A. Mansfield, R.S. Judson, J.R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart et al., Nature **403**, 623 (2000)
17. H. Jeong, S.P. Mason, A.L. Barabási, Z.N. Oltvai, Nature **411**, 41 (2001)
18. S. Maslov, K. Sneppen, Science **296**, 910 (2002)
19. S. Li, C.M. Armstrong, N. Bertin, H. Ge, S. Milstein, M. Boxem, P.O. Vidalain, J.D.J. Han, A. Chesneau, T. Hao et al., Science **303**, 540 (2004)
20. R. Guimerà, L.A.N. Amaral, Nature **433**, 895 (2005)
21. M.J. Barber, Phys. Rev. E **76**, 066102 (2007)
22. M.E.J. Newman, Proc. Natl. Acad. Sci. USA **103**, 8577 (2006)
23. S. Lehmann, M. Schwartz, L.K. Hansen, Phys. Rev. E **78**, 016108 (2008)
24. L. Danon, A. Díaz-Guilera, J. Duch, A. Arenas, J. Stat. Mech.: Theor. Exp., P09008 (2005)
25. E.N. Sawardecker, M. Sales-Pardo, L.A.N. Amaral, Eur. Phys. J. B **67**, 277 (2009)
26. S. Fortunato, M. Barthélemy, Proc. Natl. Acad. Sci. USA **104**, 36 (2007)