A Letters Journal Exploring the Frontiers of Physics

February 2013

EPL, **101** (2013) 48001
doi: 10.1209/0295-5075/101/48001

www.epljournal.org

A spectral algorithm of community identification

Xiaofeng Gong[1,2], Kun Li[1,2], Menghui Li[1,2] and C.-H. Lai[2,3,4,5]

[1] *Temasek Laboratories, National University of Singapore - Singapore 117508*
[2] *Beijing-Hong Kong-Singapore Joint Center of Nonlinear and Complex Systems (Singapore), National University of Singapore - Singapore 117508*
[3] *Department of Physics, National University of Singapore - Singapore 117542*
[4] *Centre for Quantum Technologies, National University of Singapore - Singapore 117543*
[5] *Yale-NUS College - Singapore*

received 24 September 2012; accepted in final form 7 February 2013
published online 4 March 2013

PACS 89.75.Hc – Networks and genealogical trees
PACS 89.20.Hh – World Wide Web, Internet
PACS 05.10.-a – Computational methods in statistical physics and nonlinear dynamics

Abstract – A novel spectral algorithm utilizing multiple eigenvectors is proposed to identify the communities in networks based on the modularity Q. We investigate the reduced modularity on low-rank approximations of the original modularity matrix consisting of leading eigenvectors. By exploiting the rotational invariance of the reduced modularity, near-optimal partitions of the network can be found. This approach generalizes the conventional spectral network partitioning algorithms which usually use only one eigenvector, and promises better results because more spectral information is used. The algorithm shows excellent performance on various real-world and computer-generated benchmark networks, and outperforms the most known community detection methods.

Introduction. – Network models are widely used in diverse fields to describe complex systems where large numbers of objects are interconnected, ranging from natural science to engineering systems and the human society [1,2]. One common feature of many networks is the community structure. A community is usually thought of as a subset of nodes which are interconnected densely and only sparsely connected to the rest of the network [3–5]. Such a heterogeneous distribution of links is believed to be closely related to the functioning of the underlying system. For example, in the WWW network, web pages consist of communities [6] dealing with the same topic. For more complicated systems, finding closely connected components may shed light on the organizations of the systems and their functions. Therefore, community detection has become one of the fundamental problems in network science.

There are various ways to describe community structures in networks. Of particular importance is the concept of modularity Q [3–5], defined as the difference between the total number of edges within the communities and their expected number. It is a function of the partitions which divide the network into groups, with larger value indicating stronger community structure. The merit of the modularity function is that it makes the role of the null model explicit and clear, and thus can better cope with real networks which usually have prominent heavy-tailed degree distributions [7]. If the modularity Q is used as the benefit function, the problem of community detection becomes that of searching for a good partition of the network which maximizes Q. In recent years, many algorithms have been developed trying to find good partition and get large modularity [8–15]. One particular interesting approach is the spectral method [16,17].

Usually, in a spectral partitioning algorithm, an optimization problem hinging on certain benefit function is solved first in the relaxed continuous domain, dropping the discrete constraint (0 or 1) on elements of the partition matrix S [18]. A legitimate partition, which satisfies all constraints, is then obtained by finding a suitable discrete approximation of the intermediate results. A well-known example is the method which is based on the eigenvector associated with the second smallest eigenvalue of the graph Laplacian L [19], to bisect a network to achieve the minimum cut. In the context of community detection, the connection between a partition of a network and the spectral representation of the modularity Q has been studied in [16,17]. An effective spectral algorithm has been proposed by using the leading eigenvector of the modularity matrix recursively to find multiple communities in a network.

What makes the spectral partitioning methods appealing is their global-optimal property in the relaxed

continuous domain due to the extremal properties of the eigenvalues, which also implies a better achievable performance should more eigenvectors be used. However, there are few algorithms that incorporate multiple eigenvectors [14,20]. The main reason is that, except for the situation where only a single eigenvector is involved, finding the discrete approximation is highly non-trivial. Furthermore, it is usually unclear in what sense the discrete solution (the found partition) is optimal. It remains as a challenging problem to construct a good discrete approximation efficiently from multiple eigenvectors.

In this paper, we developed an efficient way to derive a near-optimal partition from multiple eigenvectors of the modularity matrix B to maximize the modularity[1]. We consider sets of B_k's which are rank-k approximations of the original modularity matrix by using k leading eigenvectors. The reduced modularity Q_k is the modularity value when any legitimate partition S is applied to B_k. By exploiting the invariance of Q_k under orthogonal rotations, we are able to derive the near-optimal partition S_k which maximizes Q_k. The best S_k which leads to the largest modularity Q is then selected as the final partition. When $k=1$, *i.e.*, only the first eigenvector is used, the procedure is the same as that in [16,17] to divide the network into two parts. In this sense, the proposed algorithm can be regarded as a generalization of the method in [16,17]. When more eigenvectors are used, multiple communities can be found simultaneously, and the number of communities is determined automatically.

Spectral representation of modularity. – Suppose a network with N nodes and M links is given. Let $A(N\times N)$ be the adjacency matrix, and $k_i=\sum_{j=1}^{N}A_{ij}$ the degree of node i. For an arbitrary partition $\mathcal{G}$ which divides the network into K non-overlapped groups, the modularity Q can be defined as [16]

$$Q=\frac{1}{2M}\sum_{ij}(A_{ij}-P_{ij})\delta(g_i,g_j), \tag{1}$$

where g_i is the community to which node i belongs, $P_{ij}=k_ik_j/2M$ resulted from the null model selected, and $\delta(r,s)=1$, if $r=s$, and 0 otherwise. Given the partition $\mathcal{G}$, one can further define an associated partition matrix $S(N\times k)=[s_1,s_2,\cdots,s_k]$,

$$S_{ij}=\begin{cases}1, & \text{if node } i \text{ belongs to community } j,\\ 0, & \text{otherwise.}\end{cases} \tag{2}$$

Each column s_k in S is an N-dimensional $\{0,1\}$ index vector. $\mathrm{Tr}(S^TS)=N$ and $S^TS=\mathrm{Diag}(n_1,n_2,\cdots,n_K)$ is a diagonal matrix, where n_k is the size of group k. The modularity Q can now be written as

$$Q=\frac{1}{2M}\mathrm{Tr}(S^T(A-P)S)=\mathrm{Tr}(S^TBS), \tag{3}$$

where the important modularity matrix $B=\frac{1}{2M}(A-P)$. Since B is real symmetric, it can be written as $B=\sum_{j=1}^{N}\lambda_ju_ju_j^T$, where $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_N$ are ordered eigenvalues of B. The eigenvectors $U=[u_1,u_2,\cdots,u_N]$ form an orthonormal basis of an N-dimensional vector space, therefore s_k can be written as a linear combination of u_j. Let $s_k=\sum_{j=1}^{N}c_{jk}u_j$, we have

$$Q=\sum_{j=1}^{N}\sum_{k=1}^{K}\lambda_j(u_j^Ts_k)^2=\sum_{j=1}^{N}\sum_{k=1}^{K}\lambda_jc_{jp}^2. \tag{4}$$

Equation (4) is the spectral representation of the modularity Q. It clearly shows that the major contributions to the modularity come from the projection of the index vectors onto the subspace spanned by the leading eigenvectors. For a good partition achieving large Q, the index vectors as columns in the partition matrix S necessarily have large projections onto the leading eigenvectors with positive eigenvalues.

Community detection algorithm using multiple eigenvectors. –

Rank-k approximation of B. Using k leading eigenvectors, a low-rank approximation of B can be constructed by $B_k=\sum_{j=1}^{k}\lambda_ju_ju_j^T, k\ll N$, which has a much simpler structure. We can define the reduced modularity as

$$Q_k=\mathrm{Tr}(S^TB_kS). \tag{5}$$

Then B_k is the optimal rank-k approximation of B regarding the modularity in the sense that the upper bound of the reduced modularity is the maximum according to eq. (4). The basic idea behind the proposed approach is to find the optimal partition S_k maximizing Q_k by exploiting the structure simplicity of B_k. The best among $S_k, k=1,2,\cdots,K$, which achieves the largest modularity Q, is then selected as the final partition, *i.e.*,

$$\begin{aligned}S_k &= \mathrm{argmax}_\alpha\mathrm{Tr}(S_\alpha^TB_kS_\alpha), \text{ and}\\ S_m &= \mathrm{argmax}_{k:\{1,2,\cdots,K\}}\mathrm{Tr}(S_k^TBS_k),\end{aligned} \tag{6}$$

where S_α is an arbitrary partition. The maximum achieved modularity is $Q_m=\mathrm{Tr}(S_m^TBS_m)$.

Let K be the number of eigenvectors involved, and define K scaled leading eigenvectors of B as $v_j=\sqrt{\lambda_j}u_j, (j=1,2,\cdots,K,\lambda_j\geqslant 0)$. Let $U_K=[u_1,u_2,\cdots,u_K]$, $V_K=[v_1,v_2,\cdots,v_K]$, $D_K=\mathrm{Diag}(\lambda_1,\cdots,\lambda_K)$, we obtain

$$Q_K=\mathrm{Tr}(S^TB_KS)=\mathrm{Tr}(S^TV_KV_K^TS). \tag{7}$$

Write the partition matrix S as

$$S=UC=[U_K,U_R][C_K^T,C_R^T]^T=U_KC_K^T+U_RC_R^T, \tag{8}$$

we have $V_K^TS=D_K^{1/2}U_K^TS=D_K^{1/2}C_K^T=\tilde{C_K}$, and therefore

$$Q_K=\|\tilde{C_K}\|_F^2, \tag{9}$$

where $\|X\|_F=(\sum_{ij}X_{ij}^2)^{1/2}$ is the Frobenius norm of X. The task is to find a partition S_K that maximizes $\|\tilde{C_k}\|_F$.

[1]The method proposed in [14] is different from the present algorithm, where the network data is first projected and analyzed using conventional clustering algorithms on the subspace spanned by several leading eigenvectors. The results are then mapped back to the original network to obtain the final partition.

When $K=1$, the optimal partition S_1 divides the network into 2 groups. In this case, S_1 can be easily constructed from the corresponding signs of the components of V_1. Specifically, for $i=1$ and 2, we have $s_{ki}=1$, if $v_{ki}\geqslant 0$, and 0 otherwise. This is exactly the conventional spectral partitioning method except that the first eigenvector of the modularity matrix B is used instead of the second eigenvector of the graph Laplacian matrix L. When this procedure is recursively applied on the resultant groups, it leads us to the method studied in [16]. When $K>1$, it is in general difficult to derive S_K from V_K directly. However, if V_K is in a special form (referred to as the canonical form), S_K can be constructed easily.

Canonical form. Let us define a matrix $G(N\times K)$ as in canonical form associated with optimal partition $S^G(N\times 2K)$, if in each row of G, there is one and only one non-zero element. Thus, by proper rearrangement, we are able to write G as

$$G^T=\begin{bmatrix} G_1^T & G_2^T & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & 0 & G_3^T & G_4^T & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & \cdots & 0 & G_{2K-1}^T & G_{2K}^T \end{bmatrix}, \qquad (10)$$

where $G_j(n_j\times 1)=[g_1^{(j)},g_2^{(j)},\cdots,g_{n_j}^{(j)}]^T$, and $g_i^{(j)}>0$, if $j=2k-1$, and $g_i^{(j)}<0$, if $j=2k$. Denote $\tilde{G}(N\times 2K)$ a block diagonal matrix

$$\tilde{G}^T=\begin{bmatrix} G_1^T & 0 & \cdots & \cdots & 0 \\ 0 & G_2^T & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & 0 & G_{2K}^T \end{bmatrix}, \qquad (11)$$

the associated optimal partition $S^G(N\times 2K)=[s_1,s_2,\cdots,s_{2K}]$ can be determined as follows:

$$S_{ij}^G=\begin{cases} 1, & \text{if } \tilde{G_{ij}}\neq 0, \\ 0, & \text{otherwise.} \end{cases} \qquad (12)$$

Simply put, partition S^G can be obtained by first dividing the network into K groups according to the non-zero elements in each row of G^T, and then each group is split into two smaller ones based on corresponding signs. It is easy to verify that $S^G=\text{argmax}_\alpha \parallel G^T S_\alpha \parallel_F^2$. Thus, if V_K in eq. (7) is in canonical form, S_K given by (12) is the optimal partition which achieves the largest possible Q_K for B_K.

Orthonormal rotation. To obtain the desired partition S_K, we need first to construct some matrix in canonical form to approximate V_K in eq. (7). The key observation is that Q_K in eq. (7) (or (9)) is invariant under orthogonal rotations. Geometrically, if each row of $\tilde{C_K}$ is regarded as a vector in the K-dimensional space, Q_K represents the sum of the squared lengths of these vectors, which is invariant under orthonormal rotations of the axes. Formally, $Q_K=\|\tilde{C_K}\|_F^2=\|V_K^TS\|_F^2=\|R^TV_K^TS\|_F^2=\|(V_KR)^TS\|_F^2$, where $R(K\times K)$ is an arbitrary orthogonal rotation. This suggests an efficient way to construct the canonical matrix to approximate V_K, *i.e.*, we fist orthonormally transform V_K to the desired form as close as possible, and then the canonical matrix can be constructed simply by keeping only elements with the largest absolute value in each row and zeroing all others.

Given a non-ideal matrix, we hope to transform C to $\tilde{C}$ by an orthonormal rotation, where $\tilde{C}$ has only one large element and many zero (or near-zero) elements in each row. This can be achieved by minimizing the sum of squared cross terms $\sum_i\sum_j C_{ij}^2C_{ik}^2$, where $i=1,2,\cdots,N$, and $j,k=1,2,\cdots,K$ and $j\neq k$. Since $T=\sum_{ij}C_{ij}^2$ is invariant under an orthogonal rotation, so is T^2. Therefore, minimizing the sum of squared cross terms is equivalent to maximize $\sum_{ij}C_{ij}^4$. To get a good canonical approximation of V_K, we therefore consider the orthogonal rotation which maximizes

$$F=\sum_{ij}v_{ij}^4. \qquad (13)$$

The orthogonal rotation that maximizes eq. (13) is known as the quartimax rotation [21,22], and is widely used in factor analysis to get a simpler representation of the factor structure. (The pseudo-code can be found in [21].) Usually, the choice of K will greatly influence the level of the obtained sparsity. When K is close to N, the transformed matrix $\tilde{C}$ will approach the identity matrix. In the problem of community identification however, it is always the case that $K\ll N$. Such a problem will not occur.

According to the construction of S_m, up to $2K$ communities can be made by K leading eigenvectors. Since in most practical situations, we only have a rough estimate on the number of communities at best, we thus need to scan a range of K to find out the best partition S_m which leads to the largest Q_m.

A refinement procedure. In some cases, the associated partition S_K obtained by the procedure described above consists of some very small groups. We find this is mainly due to the complicated community structure of the network where some nodes cannot be clearly classified. To get more reliable result, a refinement procedure is applied. The procedure is very simple and works as follows. After we obtain a partition, one node is picked out and reassigned to the community which results in the largest modularity value. The procedure is applied to each node systematically or randomly, and repeated until convergence is achieved. The procedure above is very fast and a few iterations are sufficient to obtain converged results. During the refinement procedure, the community number can decrease. This happens if every node initially in some group needs to be moved into other groups. We found that artificial small communities will be successfully removed by this refinement procedure, which give us larger modularity value and much better estimate of the true community number. The refinement procedure we applied is different from and much simpler than the Kernighan-Lin algorithm [17,23]. In our procedure, the order of each movement is not optimized, meaning that the first moved node is not necessarily the one which leads to the most significant increase of the modularity. This makes

Table 1: Comparison of modularity values achieved by the present algorithm and other previously published methods. The cited Q values for different methods are those appeared in the original papers. *Q(nc)* is the modularity obtained by our algorithm. The number *nc* in the bracket is the community number of the partition. The details of the tested networks can be found in the cited papers.

Network	GN [3]	DA [12]	EIG [17]	GHL [24]	*Q(nc)*
Karate club	0.401	0.4188	0.419	0.4198	0.4198(4)
Dolphin	0.52				0.5263(4)
Football	0.601			0.6044	0.6046(10)
Jazz	0.401	0.4452	0.442		0.4451(4)
Metabolic	0.403	0.4342	0.435		0.4363(10)

the procedure very fast (only $O(N)$). This refined procedure when applied independently to network community detection will not generate good results usually. Its high effectiveness to remove superficial small groups as part of the algorithm is largely dependent on the fact that we already have a good partition obtained by the spectral algorithm.

Test of the algorithm. –

Measures of performance. The performance of a community detection algorithm is usually measured by the similarity between the found partition and the real known structure according to some quantitative criterion. A widely adopted quantity to measure and compare performances of different methods is the normalized mutual information (NMI) [25,26]. NMI takes its maximum value of 1 if partition A is identical to partition B, and $\mathrm{NMI} = 0$ if the two partitions are statistically independent. However, as any community detection algorithm hinges on certain benefit function implicitly or explicitly, for example, the current algorithm is based on the modularity Q, a good partition found by the algorithm may not have high NMI value. Even though the global optimal partition was found by the algorithm (for example, the maximum modularity was achieved), the NMI may still be less than 1, unless the real partition happens to realize the maximum value of the benefit function, which is unlikely especially for real networks. Thus, to fairly judge an algorithm, both aspects need to be considered.

Tests on real networks. The algorithm is tested on various real networks. In table 1, we present the results for the modularity achieved by our algorithm compared to those obtained by other methods. It can be seen that the multiple-eigenvector–based algorithm works very well on these networks. One particularly interesting real example is the network of American college football teams [3], which clearly demonstrates the power of the proposed algorithm and the difficulty discussed above when evaluating the performance of a community detection algorithm by real networks. This network is a map of the schedule of Division I games for the 2000 season where 115 nodes represent the teams and 616 edges represent games between the two teams they connect. All teams are organized into 12 conferences each of which contains about 8–12 teams. Since games are usually more frequent between members of the same conference than between members of different conferences, most conferences can be regarded as communities[2]. When using the conferences as "true partition", the modularity is $Q_0 = 0.554$ and $\mathrm{NMI} = 1$.

Table 2: Confusion matrix of community assignment of the network of American college football teams. The names of conferences are listed in the leftmost column. (The conference of IA Independence cannot be regarded as a community as discussed in the text.) In the table, columns a–j represent the communities found by the algorithm. Each found community consist of teams from one or more conferences as indicated by the numbers in the corresponding column

	a	b	c	d	e	f	g	h	i	j	
Atlantic Coast							9				9
Big East				8							8
Big 10		11									11
Big 12			12								12
Conference USA						1			9		10
IA Independents				2				2		1	5
Mid American								13			13
Mountain West										8	8
Pac 10	10										10
SEC					12						12
Sunbelt					3					4	7
Western Athletic					1	8				1	10
	10	11	12	10	16	9	9	15	9	14	115

Table 2 shows the detailed results, where, except for the conferences of "IA Independents" and "Sunbelt", all other conferences are identified as communities with high accuracy. The modularity achieved by this partition is $Q = 0.605$ and $\mathrm{NMI} = 0.88$. Since the nodes of the "Sunbelt" conference are assigned into community e and j by the algorithm, it seems that a better partition (regarding to the true conference structure) could be constructed by moving all 7 nodes belonging to the "Sunbelt" conference from community e and j and assigning them to a new community k (now there are 11 communities totally). Unfortunately, the resultant modularity value is only $Q' = 0.587$ (but NMI increases to 0.93). Thus, a partition with lower modularity value is a better choice if measured by a known community structure (or NMI). This raises a problem in evaluating a community detection algorithm based on modularity Q using real networks as discussed above.

[2]However there are a few of them, for example, the conference of IA Independence, whose teams played more or nearly as many games against teams in other conferences than or as those in their own conference. So the conferences are not completely coincident with communities found based on the topological information of the network.

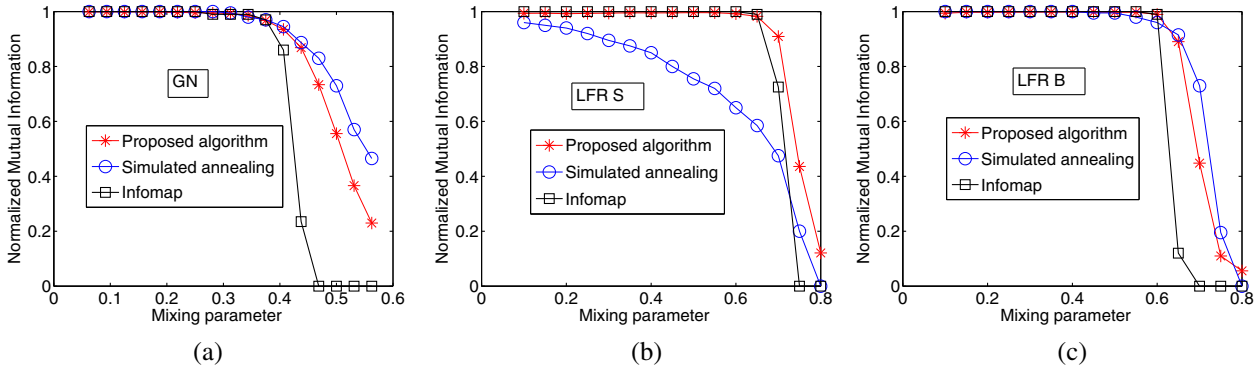

Fig. 1: (Colour on-line) Test results of the algorithm on the GN and LFR benchmarks. (a) Test results on 4-group GN benchmarks. (b) Test results on "Small" LFR benchmarks. (c) Test results on "Big" LFR benchmarks. Other parameters for (b) and (c): the average degree is 20, the maximum degree is 50, the exponent of the degree distribution is -2, and that of the community size distribution is -1. The size of the network is 1000. Each point on the curves corresponds to an average over 100 realizations. The performances of the simulated annealing and Infomap algorithm are also depicted for comparison [25].

Tests on benchmark networks. To evaluate and compare the performances of community detection algorithms more fairly, we resort to artificially generated benchmark networks. In this paper, benchmark networks generated by the stochastic block model [3,27] and the LFR model [28] are used to test the proposed algorithm. The stochastic model is described by two independent parameters p_{in} and p_{out}, which determine the connection probability between nodes in the same group and across different groups. The most popular example is the four-blocks network proposed in [3] (GN benchmark), where the network consist of 128 nodes, each with expected degree 16. The whole network is divided into four groups of 32. The LFR model is much more complicated. It generalizes the GN benchmark by introducing power law distributions for the degree and community size. An important difference between the GN and LFR model is that in the latter case, the connection probabilities of intra- and inter-group (corresponding to p_{in} and p_{out}) are no longer independent. A new parameter, the mixing ratio μ which expresses the ratio between the external degree of a node with respect to its community and the total degree of the node, is usually used to build and characterize the community structure. The LFR benchmark presents a more flexible test to various algorithms.

A comprehensive comparative study of the performances of various community detection algorithms on these two types of benchmark networks has been done in [25]. In our simulations, we use exactly the same parameter sets for both benchmarks as in [25]. The performance is also measured by NMI as in [25], which make it possible to compare directly the proposed algorithm with other methods without reproducing all results.

The test results of the proposed algorithm on GN benchmarks are shown in fig. 1(a). According to the results in [25], within the pool of analyzed methods, the best one for GN benchmark is the method of exhaustive modularity optimization via simulated annealing [10,11] and one of the best following ones is the Infomap algorithm [29]. It can be seen clearly from fig. 1(a), the algorithm proposed here has excellent performance and only slightly worse than that of the simulated annealing approach (cf., fig. 1 in [25]) which requires much more computing resources, and outperforms all the others especially when μ is relatively large (0.4–0.5).

Figures 1(b) and (c) illustrate the test results for the LFR benchmark. Two different ranges for the community sizes —"Small" and "Big"— are tested. In the "Small" category, the size of community is between 10 and 50 nodes and in the case of "Big", communities have between 20 and 100 nodes. Again, the algorithm shows very good performance, comparable to the best Infomap algorithm [16] and RN algorithm [15] reported (cf., fig. 2 in [25]), and clearly outperforms all the others. When μ is in the range 0.6–0.7, the performance of the present algorithm is even edged better than the Infomap.

The overall trend of performance with the changing of μ is similar to the other methods. For small values of μ, the communities are well separated and easily detected. The modularity values achieved by the found partition are almost the same as those of the true partition, and NMI is 1 (or very close to 1). When μ increases, the communities become entangled and harder to be identified. Interestingly, with μ increasing further, the algorithm actually starts to find partitions with larger modularity values than the true partition, even though the NMIs continue to decrease as the found partitions become irrelevant with the real one. This happens even when the community structure is still considered to be there (*e.g.*, $\mu < 0.75$ for GN benchmark).

Therefore, the sharp decrease of performance measured by NMI is largely caused by the limitation of the modularity Q to describe a weak community structure, as well as the property of NMI. Although related results are not reported in the literature, we believe that a similar phenomenon also exists in other modularity-based methods. Such intrinsic difficulty can only be resolved by applying new quantities which can describe the network community structure more accurately and more consistently.

One important problem when applying modularity is the "resolution limit" [30,31], in which the maximal modularity partition will fail to reveal small intuitive modular structures due to the coexistence of large-scale communities. Interesting algorithms have been proposed in [32,33] to mitigate the resolution problem by hierarchically screening the modular structure at different scales. The present algorithm aims to obtain the maximal modularity partition, and therefore has this resolution problem. We test our algorithm on a realization of the multi-scale benchmark network (400-13-13) used in [33]. The algorithm generates a 4-groups partition, with 2 small cliques hidden in one group. The modularity is $Q = 0.0836$ and $\mathrm{NMI} = 0.1109$, and the hidden multi-scale partition cannot be correctly revealed. However, since using the true partition, the modularity is only $Q = 0.0154$, this kind of result is not surprising since the algorithm is designed to find the maximal modularity partition. As discussed above, in such situations, modularity is no longer a suitable measure to describe a meaningful multi-scale structure. More subtle quantities, for example, the modified modularity with adjustable parameter proposed in [32,33], should be used.

Summary. – In this paper, we have developed a spectral algorithm to identify the communities in a network based on modularity Q. We introduced the reduced modularity and the associated canonical form. By exploiting the rotational invariance of the reduced modularity, we use the quartimax rotation to transform the scaled eigenvector matrix to the desired form and use it to construct a near-optimal partition for a low-rank approximation of the modularity matrix. The set of the derived partitions makes up the set of candidates, from which the final optimal partition can be selected and refined. This novel approach generalizes the conventional spectral community detection algorithms where usually only one eigenvector is involved, and therefore achieves better results because more spectral information is utilized. The algorithm has been tested on various real-world and computer-generated benchmark networks and achieves excellent results. While it is difficult if not impossible to fairly rank various community detection algorithms without a rigorous definition of what a community is, the test results on the GN and LFR benchmark show that the proposed algorithm stands among the best ones known.

REFERENCES

[1] Albert R. and Barabasi A. -L., *Rev. Mod. Phys.*, **74** (2002) 47.
[2] Newman M. E. J., *SIAM Rev.*, **45** (2003) 167.
[3] Girvan M. and Newman M. E. J., *Proc. Natl. Acad. Sci. U.S.A.*, **99** (2002) 7821.
[4] Newman M. E. J. and Girvan M., *Phys. Rev. E*, **69** (2004) 026113.
[5] Newman M. E. J., *Eur Phys. J. B*, **38** (2004) 321.
[6] Flake G. W., Lawrence S. R., Giles C. L. and Coetzee F. M., *IEEE Comput.*, **25** (2002) 66.
[7] Barabasi A. and Albert R., *Science*, **286** (1999) 509.
[8] Newman M. E. J., *Phys. Rev. E*, **69** (2004) 066133.
[9] Clauset A., Newman M. E. J. and Moore C., *Phys. Rev. E*, **70** (2004) 066111.
[10] Reichardt J. and Bornholdt S., *Phys. Rev. E*, **74** (2006) 016110.
[11] Guimera R., Pardo M. S. and Amaral L. A. N., *Phys. Rev E*, **70** (2004) 025101(R).
[12] Duch J. and Arenas A., *Phys. Rev. E*, **72** (2005) 027104.
[13] Blondel V. D., Guillaume J. L., Lambiotte R. and Lefebvre E., *J. Stat. Mech.: Theory Exp.* (2008) P10008.
[14] Donetti L. and Munoz M. A., *J. Stat. Mech.: Theory Exp.* (2004) P10012.
[15] Ronhovde P. and Nussinov Z., *Phys. Rev. E*, **80** (2009) 016109.
[16] Newman M. E., *Phys. Rev. E*, **74** (2006) 036104.
[17] Newman M. E., *Proc. Natl. Acad. Sci. U.S.A.*, **103** (2006) 8577.
[18] Chan Pak K., Schlag Martine D. F. and Zien Jason Y., *IEEE Trans. Computer-Aided Des.*, **13** (1994) 1088.
[19] Pothen A., Simon H. and Liou K. P., *SIAM J. Matrix Anal. Appl.*, **11** (1990) 430.
[20] Yu Stella X. and Shi Jianbo, in *Proceedings of the Ninth IEEE International Conference on Computer Vision (2003)*, Vol. **1**, pp. 313–319.
[21] Stegmann M. B., Sjostrand K. and Larsen R., *Proc. SPIE*, **6144** (2006) 61441G.
[22] Harman Harry H., *Modern Factor Analysis*, 2nd edition (University of Chicago Press) 1967.
[23] Kernighan B. W. and Shen Lin, *Bell Syst. Tech. J.*, **49** (1970) 291.
[24] Gustafsson M., Hönquest M. and Lombardi A., *Physica A*, **367** (2006) 559.
[25] Lancichinetti A. and Fortunato S., *Phys. Rev. E*, **80** (2009) 056117.
[26] Danon L., Guilera A. D., Duch J. and Arenas A., *J. Stat. Mech: Theory Exp.* (2005) P09008.
[27] Condon A. and Karp R. M., Random. Struct. *Algorithms*, **18** (2001) 116.
[28] Lancichinetti A., Fortunato S. and Radicchi F, *Phys. Rev. E*, **78** (2008) 046110.
[29] Rosvall M. and Bergstrom C. T., *Proc. Natl. Acad. Sci. U.S.A.*, **105** (2008) 1118.
[30] Fortunato S. and Barthèlemy M., *Proc. Natl. Acad. Sci. U.S.A.*, **104** (2007) 36.
[31] Good B. H., Montjoye Y. A. and Clauset A., *Phys. Rev. E*, **81** (2010) 046106.
[32] Arenas A., Fernández A. and Gómez S., *New J. Phys.*, **10** (2008) 053039.
[33] Granell C., Gómez S. and Arenas A., e-print arXiv:1201.2036 [physics. data-an] (2012).