ISSN 0378-4371

ELSEVIER

PHYSICA A

Recognized by the European Physical Society

STATISTICAL MECHANICS
AND ITS APPLICATIONS

Editors:
K.A. DAWSON
J.O. INDEKEU
H.E. STANLEY
C. TSALLIS

Available online at
ScienceDirect
www.sciencedirect.com

http://www.elsevier.com/locate/physa

Detecting the community structure in complex networks based on quantum mechanics

Yan Qing Niu, Bao Qing Hu, Wen Zhang, Min Wang

PII: S0378-4371(08)00609-2
DOI: 10.1016/j.physa.2008.07.008
Reference: PHYSA 11591

To appear in: *Physica A*

Received date: 9 April 2008
Revised date: 1 July 2008

Please cite this article as: Y.Q. Niu, B.Q. Hu, W. Zhang, M. Wang, Detecting the community structure in complex networks based on quantum mechanics, *Physica A* (2008), doi:10.1016/j.physa.2008.07.008

Detecting the community structure in complex networks based on quantum mechanics

Yan Qing Niu[a,b,*], Bao Qing Hu[a], Wen Zhang[c], Min Wang[a]

[a] *School of Mathematics and Statistics, Wuhan University, Wuhan, 430072, China*

[b] *School of Computer Engineering, Nanyang Technological University, 639798, Singapore*

[c] *School of Computer Science, Wuhan University, Wuhan, 430079, China*

Abstract

In this paper, we develop a novel method to detect the community structure in complex networks. The approach is based on the combination of the kernel-based clustering using quantum mechanics, the spectral clustering technique and the concept of the Bayesian information criterion. We test the proposed algorithm on Zachary's karate club network and the world of American college football. Experimental results indicate that our algorithm is efficient and effective at finding both the optimal number of clusters and the best clustering of community structures.

***Keywords**:* Complex network; Community structure; Spectral clustering; Quantum clustering

1. Introduction

Complex networks are attracting increasing interests of scientists from physics and other fields. In the context of network theory, the term "complex network" refers to a network by virtue of certain non-trivial topological structures [1-4], which include a heavy-tail in the degree distribution, a high clustering coefficient, assortativity or disassortativity among nodes, community structures at many scales and evidence of a hierarchical structure. One of the key problems is how to detect community structures in complex networks, which have dense internal links and a lower density of external links. Many studies have verified the community structure in various complex networks such as protein interaction [5], the worldwide web [6] and scientific collaboration [4, 7]. Clearly, the ability to detect community structure in a network has important practical applications and can help us understand the network system.

There are several empirical methods to detect community structures in complex networks. Kernighan and Lin [8] proposed a heuristic procedure to produce a dimidiate network, and traditional spectral method [9, 10] was also a bisection algorithm. Newman and Girvan [4, 11] introduced the shortest-path betweenness algorithm to split the whole network into the disconnected communities, until the network is decomposed to components consisting of one single node. These methods have been shown to be very powerful only when the number of the community as a priori knowledge is given. To overcome this limitation, lots of efficient heuristic methods have been proposed over the years. Newman and Girvan [12] devised a quantitative measure called modularity Q to evaluate the quality of dividing the nodes in networks into different communities. This method can select the optimal number of clusters by maximizing the modularity Q. Following this approach, many algorithms [13-16] have investigated different exploration to find the community structure while maximizing Q. On the other hand, computer science was also working on clustering of a particular instance of networks. A common tool used to address clustering of the complex network is spectral analysis [17-23], which is based on the analysis of the adjacency matrix and the hard clustering algorithms for exploring community structures. This kind of methods combine the power of spectral

* Corresponding author.
E-mail address: niuyanqing2005@hotmail.com; yanqing.niu@gmail.com

analysis to reveal underlying structures, and they are not constrained by the iterative bisection, and not needing a priori information about the number of communities as the extra input.

Quantum clustering proposed by D. Horn [24-26] is a novel kernel-based clustering approach using quantum mechanics. The approach introduces the Schrödinger partial differential equation to characterize a quantum system and calculate the potential function using the eigenstate of the quantum system. The approximation of eigenstate is obtained by summing up Gaussian kernel functions. The potential function derived from Schrödinger partial differential equation is assimilated with the probability density function for datasets and act as a tool to find the clusters from datasets. The local minima of the potential function interpret the centers of the data samples corresponding to different clusters. Moreover, the number of cluster depends on the appropriate selection of the Gaussian kernel scale parameter σ [25-29]. This issue was addressed by Varshavsky [31] who used a statistical approach based on the Bayesian information criterion (BIC) [30] to estimate the parameter σ.

Our proposed algorithm begins by using spectral clustering analysis to form the input datasets for further clustering procedure, and then the quantum clustering algorithm is used to make division of community structures. The kernel scale parameter σ is selected by the Bayesian information criterion. Because the lowest BIC score reflects the optimal selection of the parameter σ and determines the best clustering function, we find the optimal number of community structures by minimizing the BIC score. When applied to two real-world networks in which community structures are already known, our method appears to give excellent agreement with the expected results. We also compare the new algorithm with Newman's algorithm on these two networks datasets in terms of the modularity Q.

The remainder of the paper is organized as follows. In section 2, a general introduction to spectral clustering analysis is provided, while the nonparametric estimation approach derived from quantum mechanics is described in section 3. The proposed method for detecting community structure is detailed in Sections 4. The section 5 is contributed to experiments and discussions about our method. At last, the conclusions are drawn.

2. Spectral clustering analysis

Spectral clustering analysis focuses on the relationship between the community structure and the spectral property of the complex network. Spectral methods are based on the analysis of the adjacency matrix A of the network [21, 22], in which element A_{ij} is equal to 1 if node i points to node j and 0 otherwise. Actually, instead of using the adjacency matrix A, it is more convenient for us to study three extending matrices derived from A, respectively named the Laplacian matrix $D-A$, the normal matrix $D^{-1}A$ and the matrix $D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ [19], where D is the diagonal matrix with elements $D_{ii}=\sum_{j=1}^{N}A_{ij}$ and N is the number of nodes in the network.

Though the matrix $D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ has the same eigenvalues with the normal matrix $D^{-1}A$, they have different eigenvectors. It is easily to verify that after the rows of eigenvectors are normalized to length 1, the eigenvectors obtained from these two matrices are identical. Therefore, due to the complexity of computation, we discuss the normal matrix $D^{-1}A$ in the following part.

The normal matrix $D^{-1}A$ has always the largest eigenvalue which equals to one, associated to a trivial constant eigenvector, due to the row normalization of the adjacency matrix A. For a network with the apparent cluster structure, the normal matrix has also a certain number $k-1$ of eigenvalues close to unit length, where k is the number of well defined communities. The very similar components of eigenvectors associated to these first $k-1$ nontrivial eigenvalues correspond to nodes within the same cluster. But in most common occurrences, communities cannot be simply detected by exploring the nontrivial eigenvectors. However, in an automatic manner, the use of the clustering algorithm is a natural way to identify communities from the components of nontrivial eigenvectors corresponding to nodes.

Furthermore, we would like to know how to evaluate the quality of the clustering of nodes. This problem is addressed by Newman and Girvan [12] who defined a measure of the quality of a particular division of a network, which gives the modularity Q

$$Q = \sum_i (e_{ii} - a_i^2) = Tre - \left\| e^2 \right\| \tag{1}$$

where e is a $k \times k$ symmetric matrix whose element e_{ij} is the fraction of all edges in the network that link nodes in community i to nodes in community j, and $\left\| e^2 \right\|$ indicates the sum of the elements of the matrix e^2. The trace $Tre = \sum_i e_{ii}$ represents the fraction of edges in the network that connect the nodes in the same community, and a clear division into communities should have a high value of this trace. The row (or column) sums $a_i = \sum_j e_{ij}$ gives the fraction of edges that connect to nodes in community i.

The quantity Q measure the fraction of the inter-community edges minus the expected value of the same quantity in a network with the same community structures but random connections between the nodes. In practice, the high value of Q represents apparent community structures.

3. Quantum clustering algorithm

Quantum clustering algorithm proposed by D. Horn [24-26] is a novel kernel-based clustering approach using quantum mechanics. It focuses on the Schrödinger potential function $V(x)$ provided by the Schrödinger partial differential equation

$$H\psi(x) \equiv (-\frac{\sigma^2}{2}\nabla^2 + V(x))\psi = E\psi \tag{2}$$

where H is the Hamiltonian operator, E is an eigenvalue energy level, $\psi(x)$ corresponds to the eigenstate of the given quantum system, $V(x)$ is the Schrödinger potential and ∇^2 is the Laplacian operator. The potential is always positive i.e. $V(x) \ge 0$.

Based on the quantum mechanics principles, the Schrödinger partial differential equation characterizes a quantum system and the dataset can be seen as particles of the quantum system on orbits that obey the quantum physics laws [28, 29]. In quantum mechanics, the wave function $\psi(x)$ can be defined corresponding to the given potential and energy level, by solving the Schrödinger equation. While in the quantum clustering algorithm, the inverse problem is to be considered. By giving the wave function $\psi(x)$, the goal in quantum clustering becomes to estimate the quantum potential $V(x)$, which characterizes the probability density function for data samples.

Due to its smoothness and differentiability properties, the Gaussian kernel function

$$G(x)=\exp\left\{\frac{-(x-x_i)^T(x-x_i)}{2\sigma^2}\right\} \tag{3}$$

is assigned to each data sample $x_i, i=1,2,\cdots,N$, where σ is a kernel parameter called scale, width or bandwidth. Afterwards, the eigenstate function $\psi(x)$ is approximated by

$$\psi(x)=\sum_{i=1}^{N}\exp\left\{\frac{-(x-x_i)^T(x-x_i)}{2\sigma^2}\right\} \tag{4}$$

After replacing $\psi(x)$ from Eq. (4) into Eq. (2), we can solve the corresponding Schrödinger potential $V(x)$ assimilated with the probability distribution of the given data samples as

$$V(x)=E+\frac{\sigma^2\nabla^2\psi}{2\psi}=E-\frac{d}{2}+\frac{1}{2\sigma^2\psi}\sum_i(x-x_i)^T(x-x_i)\exp\left\{\frac{-(x-x_i)^T(x-x_i)}{2\sigma^2}\right\} \tag{5}$$

Let us furthermore require that $\min V=0$, which sets the value

$$E=-\min\frac{\sigma^2\nabla^2\psi}{2\psi} \tag{6}$$

The probability distribution approximation by Eq. (5) and constraint Eq. (6) results in a smooth potential function that fits well with the data samples. The analogies exist between data samples and quantum particles. Quantum particles that are characterized by a certain state have equal potential, while data samples that are located nearby have close potential values. Another analogy is between particles which are characterized by low potential values and the local minima of the potential function specific to cluster centers. However, the approximation of the data probability distribution function depends on the choice of the parameter σ.

4. Detecting the community structure in complex networks using quantum clustering

In this section, we use the quantum clustering approach to detect the community structure in complex networks on the basis of preceding insights. Because the estimation performance of the potential function is controlled by a scale parameter σ which can be observed from Eq. (4), the key to the approach is how to choose the appropriate kernel scale parameter σ.

4.1 Kernel scale estimation

We note that, one of the difficulties within this kernel-based approach is how to select the scale parameter σ. As expected, the kernel scale plays the role of a smoothing parameter, and there is a trade-off between sensitivity to noise at small σ and over-smoothing at large σ. Implicitly, the number of clusters in the data samples depends on the appropriate selection of the kernel scale parameter σ.

The Bayesian information criterion (BIC) was proposed by Fraley and Raftery [30] in a model-based analysis that assumed the datasets follow Gaussian probability distribution, and used by Varshavsky [31] to select the kernel scale parameter σ and assess the quality of clustering.

BIC is defined as follows

$$\text{BIC}=-2l_M(x,\Theta)+m_M\log(N)\approx-2\log p(x\mid M)+\text{constant} \tag{7}$$

where $l_M(x,\Theta)$ is the mixture log likelihood of the data x and the model Θ, which is maximized under the constraint that m_M (a function of the number of independent parameter), is minimized.

Under the assumption that the model errors or disturbances are normally distributed, BIC becomes

$$\mathrm{BIC} = N\log(\frac{\mathrm{RSS}}{N}) + K\log N \tag{8}$$

where N is the number of observation, equivalently, the sample size, K is the number of free parameters to be estimated. RSS is the residue sum of squares from the model

$$\mathrm{RSS} = \sum_{i=1}^{N} I(t_i \neq \Theta(x_i)) \tag{9}$$

$$I(t_i \neq \Theta(x_i)) = \begin{cases} 0 & t_i = \Theta(x_i) \\ 1 & t_i \neq \Theta(x_i) \end{cases} \tag{10}$$

where t_i is the target value, $i = 1, 2, \cdots, N$.

The minimum of the BIC score reflects the optimal selection of the parameter σ and the optimal number of clusters.

4.2 Algorithm

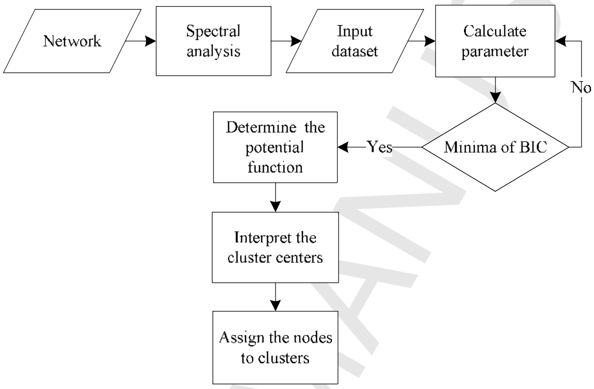

Network
Spectral analysis
Input dataset
Calculate parameter
No
Minima of BIC
Yes
Determine the potential function
Interpret the cluster centers
Assign the nodes to clusters

Fig. 1. Data flow of the proposed approach for detecting community structures

The proposed algorithm consists of the following steps:

(1) Form the input datasets according to the spectral analysis

(a) Given the input datasets. For complex networks, these datasets is described by the adjacency matrix such that $A_{ij} = 1$ if node i and j are connected by an edge and $A_{ij} = 0$ otherwise.

(b) Define D to be the diagonal matrix whose (i,i) -element is the sum of A 's i -th row which represents the degree of the node i and afterwards construct the normal matrix $D^{-1}A$.

(c) Find k nontrivial eigenvalues of the normal matrix $D^{-1}A$, where k is the number of clusters and choose $u_1, u_2, \cdots, u_k$, the k eigenvectors attributed to the k nontrivial eigenvalues (chosen to be orthogonal to each other in the case of repeated eigenvalues), then form the matrix $U = [u_1, u_2, \cdots, u_k]$ by stacking the eigenvectors in columns.

(d) Treating each of the row of U as a point in R^k , form the input samples as vectors $x_i = (U_{i1}, U_{i2}, \cdots, U_{ik})$, $i = 1, 2, \cdots, n$

(2) Select the optimal scale parameter σ by minimizing the BIC score

(3) Find the quantum potential $V(x)$ assimilated with the probability distribution of the given datasets

(a) Compute the function $\psi(x) = \sum_{i=1}^{N} \exp\left\{\frac{-(x-x_i)^T(x-x_i)}{2\sigma^2}\right\}$

(b) Compute the energy $E = -\min\frac{\sigma^2\nabla^2\psi}{2\psi}$

(c) Compute the potential $V(x)=E-\frac{d}{2}+\frac{1}{2\sigma^2\psi}\sum_i (x-x_i)^T(x-x_i)\exp\left\{\frac{-(x-x_i)^T(x-x_i)}{2\sigma^2}\right\}$

(4) The local minima of $V(x)$ interpret cluster centers.

(5) Finally, assign the original node i in the complex networks to cluster j if and only if the row i of the matrix U was assigned to cluster j.

5. Experiments and results

In this section, we test our algorithm on two real-word networks. One is the karate club [32], and the other is the American college football [33].

5.1 Zachary's karate club

The well-known karate club network studied by Zachary [12, 32] was widely used as a test network to indentify community structures. The network consists of 34 nodes demonstrating members in the karate club and 78 edges representing the friendship between club's members. Due to a disagreement between the administrator of the club and the club's instructor, the club eventually split into two smaller ones, centered round the administrator and the instructor. In Figure 2, we show the network, with the instructor and the administrator represented by nodes 1 and 34, respectively. Here we use an unweighted version of the network and apply our algorithm to it so as to extract community structures.

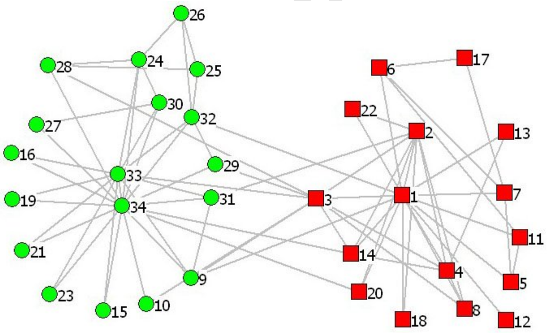
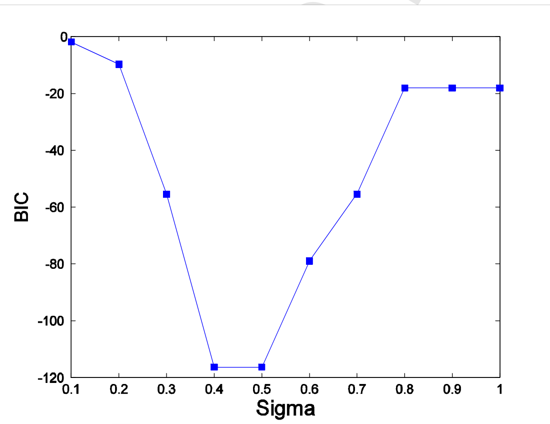


Fig. 2. The network of friendships between individuals in the karate club study of Zachary [12, 32]



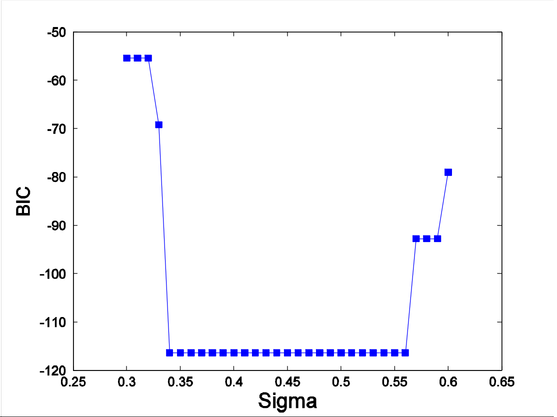


Fig. 3. Functional representation of BIC for each nonparametric method according to the arbitrary value σ from 0.1 to 1

Fig. 4. Functional representation of the BIC for each nonparametric method according to the different value σ from 0.3 to 0.6

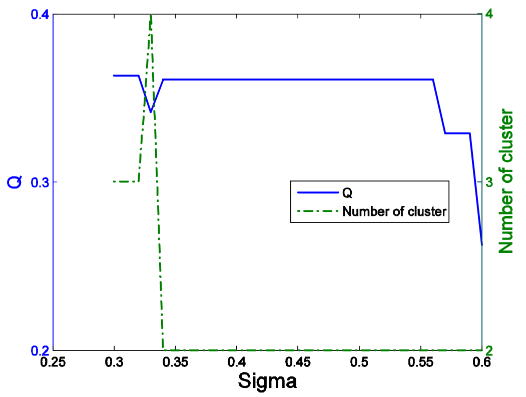


Fig. 5. Functional representation of the number of clusters and the modularity according to the different value σ from 0.3 to 0.6

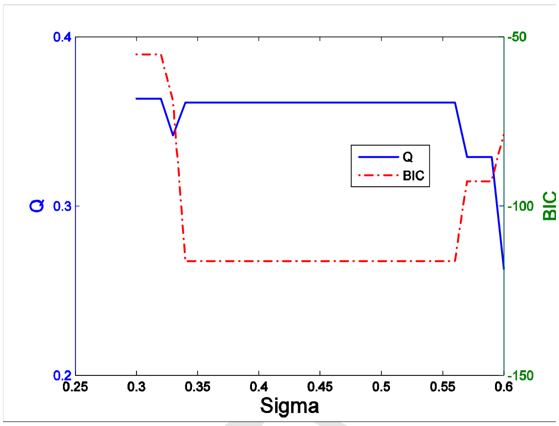


Fig. 6. Functional representation of BIC and the modularity according to the different value σ from 0.3 to 0.6

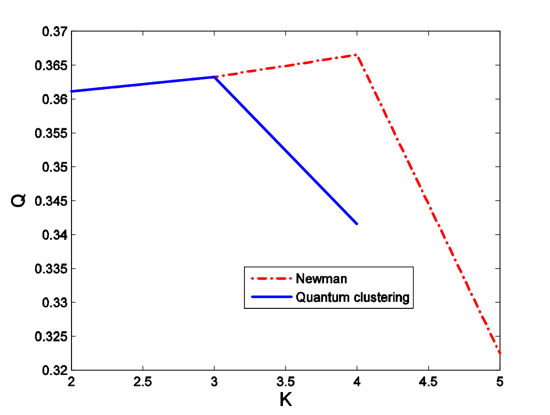


Fig. 7. Functional representation of the number of clusters k and modularity Q for Newman's algorithm [12] and our method

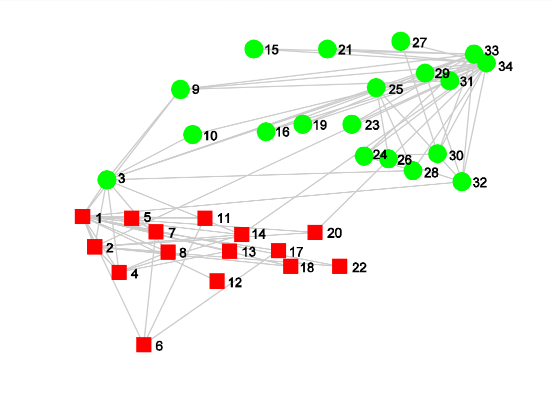


Fig. 8. The community structures of the Zachary's karate club detected by our method

Figure 7 shows how the modularity Q varies with cluster k for both of our algorithm and Newman's shortest-path betweeness algorithm [12]. As Newman pointed out, the optimal number of cluster k can be obtained by selecting the level of the resulting dendrograms for which the modularity Q is highest. For Newman's algorithm, the best clustering is $k = 4$, $Q = 0.36654$, which can not correspond precisely with the actual number of the karate club.

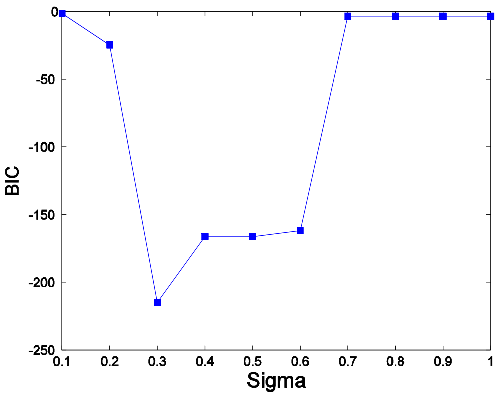
Unlike the Newman's algorithm which seeks the maximum of Q, our algorithm finds the optimal number of community structures and the best clustering by minimizing the BIC score, because the best quantum clustering function is determined by the minimum of the BIC score. The modularity Q is only used as a tool to compare the quality of the quantum clustering approach with Newman's algorithm. We note that, in the process of selecting the optimal parameter σ, when the modularity Q reaches the peak, the BIC score is not the lowest as shown in Figure 6. However, our algorithm aims to find the minimum of the BIC score. The best clustering found by our method is $k = 2$, $Q = 0.36111$.

Figure 8 shows the community structures of the karate club network detected by our method. The karate club network has been divided into two groups in which only node 3 is classified incorrectly.

5.2 American college football

We also apply our algorithm to the world of American college football [4, 33]. The unweighted network was drawn from the schedule of games played between 115 American college football teams.

The nodes in this network represent college football teams and the edges represent the fact that two teams played games together. Because the twelve conferences to which each team belongs is known and because games are more frequent between teams of the same conference than between teams of different conferences, the community structures should be findable in the college football network.



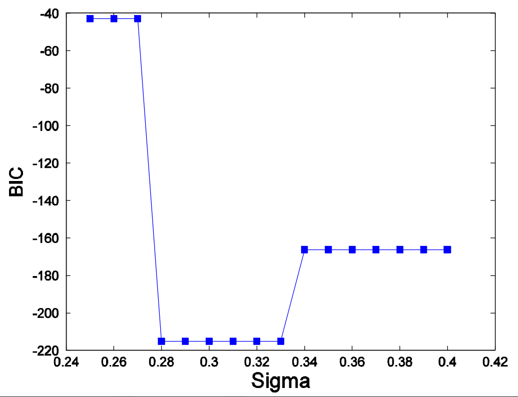
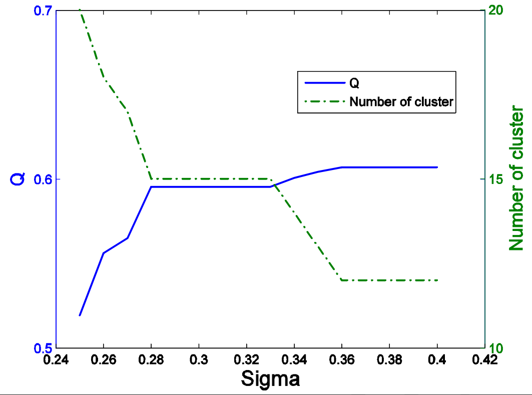


Fig. 9. Functional representation of BIC according to the arbitrary σ from 0.1 to 1

Fig. 10. Functional representation of the BIC according to the different value σ from 0.25 to 0.4



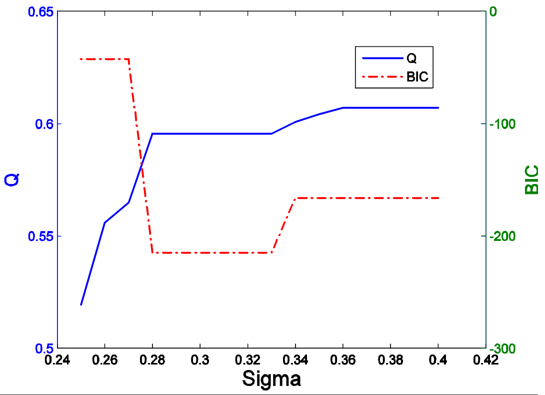


Fig. 11. Functional representation of the number of clusters and modularity according to the different value σ from 0.25 to 0.4

Fig. 12. Functional representation of BIC and modularity according to the different value σ from 0.25 to 0.4

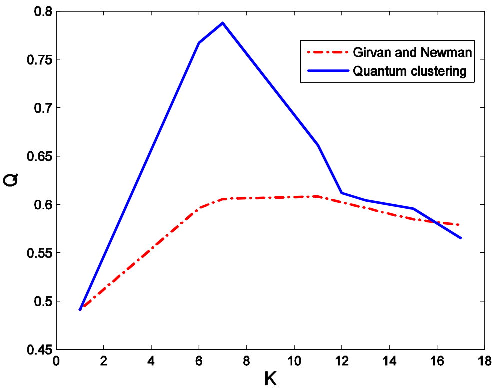


Fig. 13. Functional representation of the number of clusters k and modularity Q for Girvan and Newman [4] and our method

Figure 13 shows how the modularity Q varies with cluster k for both of our algorithm as well as Girvan and Newman's algorithm [4]. The best clustering for Newman's algorithm is $k = 6$, $Q = 0.59597$. Our method seeks to minimize the BIC score. We note that, the BIC score is not the lowest, when the modularity Q reaches the peak, as shown in Figure 12. Applying our algorithm to this network, we find that it identifies twelve community structures with $Q = 0.59553$ (see Fig. 14), which corresponds precisely with the actual number of conferences in the American college football league. But three teams of the Independent conference: Navy, Notre Dame and Connecticut do not belong to any of the twelve community structures. The other teams of the Independent conference are grouped with the Western Athletic conference and the Sunbelt conference. The Sunbelt conference is broken into two smaller conferences and grouped with teams of the Western Athletic conference and Independent conference. However, all other team assignments to community structures made by our algorithm are correct.

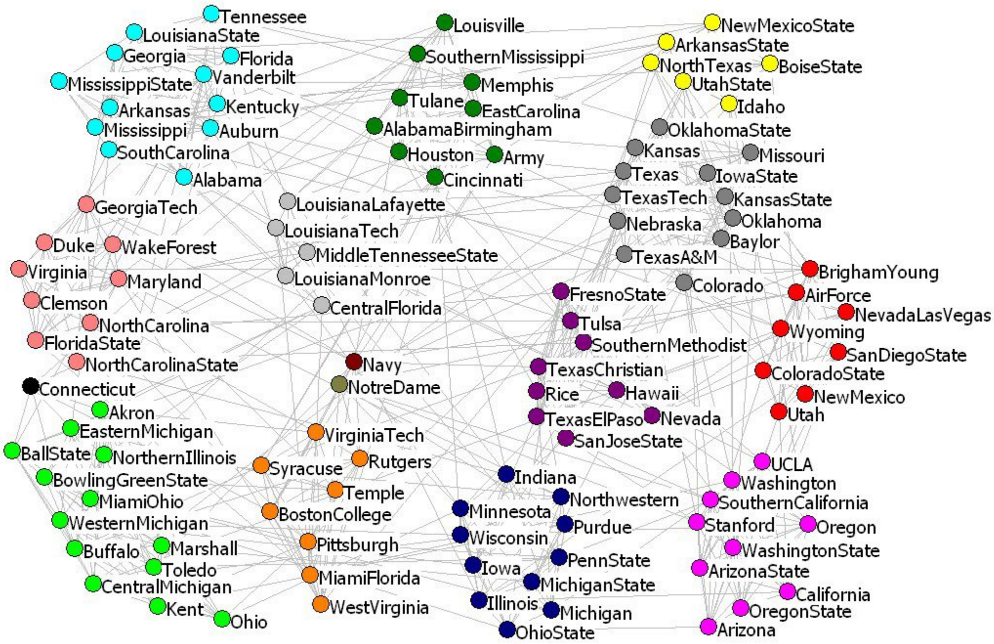


Fig. 14. Plot of community structures viewed in different colors for American college football obtained by our method

6. Conclusions

In this paper, we develop a new method to detect community structures in complex networks. The approach combines quantum clustering algorithm, the spectral clustering technique and the concept of the Bayesian information criterion. Our approach works by using quantum clustering algorithm to find the strongly connected cores of community structures, which is characterized by the local minima of the quantum potential. We apply our method to two real-world networks, and compare experimental results with the known community structures. We find that in both cases the method identifies apparent community structures. Some extensions or improvements of the proposed method can be considered further, and we hope to generalize the method to handle both weighted and directed networks in future.

Acknowledgments

The authors would like to thank Dr. Narendra Shivaji Chaudhari for useful conversations. This paper is supported by the Natural Scientific Foundation of China (Grant Nos. 70771081 and 60773167) and 973 National Basic Research Program of China (Grant No. 2007CB310804).

References

[1] R. Albert, A.L. Barabási, Reviews of Modern Physics, 74 (2002) 47-97.

[2] S.H. Strogatz, Nature, 410 (2001) 268-276.

[3] M.E.J. Newman, Phys. Rev. E 67 (2003) 026126.

[4] M. Girvan, M.E.J. Newman, Proc. Natl. Acad. Sci. USA 99 (2002) 7821-7826.

[5] H. Jeong, S. Mason, Z. Oltvai, A.L. Barabási, Nature, 411 (2001) 41-42.

[6] R. Albert, H. Jeong, A.L. Barabási, Nature, 401 (1999) 130-131.

[7] M.E.J. Newman, Phys. Rev. E 64 (2000) 016131.

[8] B.W. Kernighan, S. Lin, Bell System Technical Journal, 49 (1970) 291-307.

[9] M. Fiedler, Czech. Math J, 23 (1973) 660-670.

[10] A. Pothen, H. Simon, K.P. Liou, SIAM J Matrix Anal. App l, 11 (1990) 430-452.

[11] M.E.J. Newman, Phys. Rev. E 64 (2001) 016132.

[12] M.E.J. Newman, M. Girvan, Phys. Rev. E 69 (2004) 026113.

[13] A. Clauset, M.E.J. Newman, C. Moore, Phys. Rev. E 70 (2004) 066111.

[14] M.E.J. Newman, Phys. Rev. E 69 (2004) 066113.

[15] F. Wu, B.A. Huberman, Eur. Phys. J. B, 38 (2004) 331-338.

[16] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, D. Parisi, Proc. Natl. Acad. Sci. USA 101 (2004) 2658-2663.

[17] Y. Weiss, In Proceedings of IEEE International Conference on Computer Vision (1999) 975-982.

[18] J. Shi, J. Malik, In IEEE Transactions on Pattern Analysis and Machine Intelligence, 22 (2000) 888-905.

[19] A. Ng, M. Jordan, Y. Weiss, In Advances in Neural Information Processing Systems 14 (2002) 849-856.

[20] S. White, P. Smyth, In SIAM Data Mining Conference (2005).

[21] A. Capocci, V.D.P. Servedio, G. Caldarelli, F. Colaiori, Phys. A 352 (2005) 669-676.

[22] V.D.P. Servedio, F. Colaiori, A. Capocci, G. Caldarelli, AIP Conf. Proc. 776 (2005) 277-286.

[23] M.E.J. Newman, Phys. Rev. E 74 (2006) 036104.

[24] D. Horn, Phys. A 302 (2001) 70-79.

[25] D. Horn, A. Gottlieb, Proceedings of Advances in Neural Information Processing Systems, 14 (2001) 769-776.

[26] D. Horn, A. Gottlieb, Phys. Rev. Lett. 88 (2002) 1-4.

[27] S.J. Roberts, Pattern Recognition, 30 (1997) 261–272.

[28] N. Nasios, A.G. Bors, Proceedings of the IEEE International Conference on Image Processing, Italy, 3 (2005) 820-823.

[29] N. Nasios, A.G. Bors, Pattern Recognition, 40 (2007) 875-889.

[30] C. Fraley, A.E. Raftery, Computer Journal, 41 (1998) 578-588.

[31] R. Varshavsky, D. Horn, M. Linial, ISBRA, Lecture Notes in Bioinformatics, 4463 (2007) 85-96.

[32] W.W. Zachary, Journal of Anthropological Research 33 (1977) 452-473.

[33] Network data is available at http://www-personal.umich.edu/~mejn/netdata/.