Finding and evaluating the hierarchical structure in complex networks

2007 J. Phys. A: Math. Theor. 40 5013

(http://iopscience.iop.org/1751-8121/40/19/006)

IOP Publishing Journal of Physics A: Mathematical and Theoretical

J. Phys. A: Math. Theor. **40** (2007) 5013–5023 doi:10.1088/1751-8113/40/19/006

Finding and evaluating the hierarchical structure in complex networks

Fei Chen, Zengqiang Chen, Zhongxin Liu, Linying Xiang and Zhuzhi Yuan

Department of Automation, Nankai University, Tianjin 300071, People's Republic of China

E-mail: chernf@gmail.com

Received 28 December 2006, in final form 27 March 2007
Published 24 April 2007
Online at stacks.iop.org/JPhysA/40/5013

Abstract

A number of recent studies have focused on a statistical property of networked systems—the hierarchical structure. The problem of detecting and characterizing the hierarchical structure has recently attracted considerable attention. In this paper, it is rewritten as optimization in terms of the eigenvalues and eigenvectors. Based on that, an algorithm for reconstructing the hierarchical structure of complex networks is proposed. It is tested on some real-world graphs and is found to offer high sensitivity and reliability.

PACS numbers: 89.75.Hc, 02.50.−r

(Some figures in this article are in colour only in the electronic version)

1. Introduction

The study of networked systems has a history stretching back several centuries, but it has experienced a particular surge of interest in the last decade, partly as a result of the increasing availability of large-scale accurate data describing the topology of networks in the real world [1–6]. There are quite a wide variety of complex systems, described by networks, i.e. the metabolic network [7], the Internet [8, 9], the World-Wide Web [10], etc [11, 12]. The last few years have witnessed a tremendous activity devoted to the characterization and understanding of networked systems.

The hierarchical structure is a common feature of many networked systems and has received a considerable amount of attention in recent years. For instance, in a social network, each person may have different 'importance', or we say centrality, in the network. Different hierarchies may correspond to different clusters of people, in which people with approximately the same centrality lie on the same hierarchy. If such a hierarchical structure is found, it can be used for many purposes, i.e. control of rumour, virus spreading, etc. Moreover, it is shown that the hierarchical structure is related to some significant characteristics of complex

systems, such as the high clustering coefficient and scale-free degree distribution [7]. In [13], a one-dimensional model for diffusion on a hierarchical tree structure was proposed. It was shown numerically that this model exhibited ageing phenomena and was originated from the hierarchical structure of phase space. In [14], Variano *et al* reported the emergence of modularity and hierarchical organization in evolved networks supporting asymptotically stable linear dynamics. Numerical experiments demonstrated that linear stability benefited from the presence of a hierarchy of modules and that this architecture improved the robustness of network stability to random perturbations in network structure. In [15], Gallos extended the model of Bonabeau *et al* in the case of scale-free networks and analysed the appearance of hierarchies associated with the scaling exponent. In general, the hierarchical structure gives us many insights to both the structure and the function of complex networked systems [7, 14–17]. Thus, reconstructing the hierarchical structure from a given network is a non-trivial task.

Our work in the present paper is by no means the first to reconstruct the hierarchical structure from networks. In [18], Yang *et al* proposed an approach based on eigenvector centrality to reconstruct the hierarchical structure from a complex network. However, their approach encountered some drawbacks: The two parameters EC_{crit} and D^{c}_{left} should be prescribed, which, due to the insufficiency of knowledge, cannot be prescribed in advance.

In the present paper, we rewrite this problem as the task of optimization in terms of the eigenvalues and eigenvectors. Then a method based on spectral partitioning is presented to reconstruct the hierarchical structure from a complex network. This method is different from the one introduced in [18] and overcomes the drawbacks it raises. In section 2, we give the description of some criterion to evaluate the centrality of a node. Section 3 gives our algorithm for reconstructing the hierarchical structure from a complex network. Moreover, in section 4 we feed two real-world networks to our algorithm and show the validity of the present method. Finally, section 5 summarizes the main conclusion.

2. What is a node's centrality?

Before giving the hierarchical structure of a complex network, the first issue we should draw is to measure the centrality of each node. Depending on the context of networks, various measures of centrality are proposed. Among those four are commonly used: degree, closeness, betweenness [19] and eigenvector centrality [20, 21].

- Degree centrality of a node i is defined to be

$$d_i = \sum_j a_{ij}, \tag{1}$$

  where $a_{ij} = 1$ if there is an edge from node i to j, otherwise $a_{ij} = 0$. In an acquaintance network, if the popularity is being accessed, then the degree centrality would be appropriate for this purpose.
- Closeness centrality of a node i is defined to be

$$c_i = \sum_j d_{ij}, \tag{2}$$

  where d_{ij} is the length of the path from node i to j. Closeness is an inverse measure of centrality in that a larger value indicates a less central position, while a smaller value indicates a more central one. In the context of network diffusion, closeness can be interpreted as an index of the expected time until arrival at a given node of whatever is flowing through the network.

- Betweenness centrality [22] of a node i is defined as:

$$b_i = \sum_{jk} \frac{g_{jik}}{g_{jk}}, \tag{3}$$

  where g_{jik} is the number of paths that pass through node i from node j to node k and g_{jk} is the number of paths from node j to node k. It measures the centrality of nodes when the information diffuses along the shortest path in the network.
- Eigenvector centrality is best described by the following equation:

$$\lambda e(i) = \sum_{j} a_{ij} e(j). \tag{4}$$

  In the matrix form, it is written as

$$\lambda x = Ax. \tag{5}$$

  This type of equation is well solved as the eigenvalues and eigenvectors of the adjacency matrix A.

Eigenvector centrality is based on the simple fact that a node's centrality is determined by the centrality of the nodes that is incident to that node. As A is the adjacency matrix of a network, A is non-negative and due to the theorem of Perron–Frobenius, there exists an eigenvector of the maximal eigenvalue, called principal eigenvector, with only non-negative entries. Due to the physical meaning of the eigenvector, only the principal eigenvector is suitable for being the centrality measure. Moreover, since each eigenvalue has many corresponding eigenvectors and they are all parallel, we use $|e_{\text{principal}}|$ as our measure for the eigenvector centrality, with $e_{\text{principal}}$ the principal eigenvector of A.

In [21], Ruhnau proposed a criterion for evaluating the measure of centrality called node centrality. It was defined as follows.

Definition 1 (node centrality)**.** *Let $G = (V, E)$ be an undirected and connected graph with $|V| = n$. Let nc be a function which assigns a real value to every node of G. $nc(v_i)$ is called a node centrality of node v_i if*

(I) $nc(v_i) \in [0, 1]$ *for every* $v_i \in V$,
(II) $nc(v_i) = 1$ *if and only if* $G = S_{1,n-1}$ *and* $i = 1$,

where $S_{1,n-1}$ denotes the star-shaped networks with the centre v_1.

It was drawn from [21] that the betweenness centrality and eigenvector centrality are node centrality, while the other two are not. Hence, in this paper, we would focus mainly on eigenvector centrality and betweenness centrality as the measure to a node's centrality. However, the algorithm presented in this paper is not constrained by the definite measure being used. Hence, we can use a proper centrality measure according to the application requests which are well established.

3. Finding the hierarchical structure from a network

In this section, we propose the algorithm based on spectral partitioning for reconstructing the hierarchical structure from a network. There is a large literature within computer science on spectral partitioning, in which network properties are linked to the spectrum of the graph Laplacian matrix [27–29]. Despite its evident success in the graph partitioning arena, spectral partitioning is a poor approach for detecting the hierarchical structure in real-world networks, which is the primary topic of this paper. The condition for graph partitioning method to

be valid is that the sizes of the groups into which the networks are divided should be fixed. However, in most cases, we do not know the sizes in advance. We would like to let the sizes of hierarchies be free and in this case try to make the best division. But the graph partitioning method will break down in this case. How to solve this problem? This is the main issue of this section.

First we should consider the simplest case that dividing the network into two hierarchies. We begin by defining the eigenvector centrality matrix, EC, to be the matrix with elements:

$$EC_{ij} = |EC(i) - EC(j)|, \tag{6}$$

with $EC(i)$ the centrality of node i depending on the centrality measure used. It is obvious that the matrix EC is symmetric.

Therefore, we get the following criterion:

$$D = \frac{1}{2}\sum_{i,j}(EC_{ij} - Avg_{ij})(1 - \delta(h(i), h(j))), \tag{7}$$

where $h(i)$ denotes the hierarchy to which node i belongs and

$$\delta(i, j) = \begin{cases} 1 & i = j \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

The factor $\frac{1}{2}$ compensates for our calculation of each node pair twice.

The physical meaning of Avg_{ij} is the expectation of the difference of centrality of nodes i and j. It is defined as follows:

$$Avg_{ij} = \begin{cases} \dfrac{1}{n(n-1)}\displaystyle\sum_{i,j} EC_{ij} & i \neq j \\ 0 & i = j. \end{cases} \tag{9}$$

The expectation of the difference of centrality between node i and node j is the expectation of the difference of centrality in the whole network. Since nodes i and j do not tell us anything valuable to the expectation of the difference of centrality between them, equation (9) is reasonable. Moreover, we note that the expectation is dependent on the particular network, being consistent with our intuition that different networks have different expectations of the difference of centrality.

In the following, we define the index vector $s = [s_1, s_2, \ldots, s_n]^T$:

$$s_i = \begin{cases} +1 & \text{if node } i \text{ belongs to hierarchy 1} \\ -1 & \text{if node } i \text{ belongs to hierarchy 2.} \end{cases} \tag{10}$$

Note that s satisfies the normalization condition

$$s^T s = n. \tag{11}$$

Then

$$\frac{1}{2}(1 - s_i s_j) = \begin{cases} 0 & \text{if } i \text{ and } j \text{ belong to the same hierarchy} \\ 1 & \text{otherwise.} \end{cases} \tag{12}$$

Thus, equation (7) could be rewritten as

$$D = \frac{1}{4}\sum_{i,j}(EC_{ij} - Avg_{ij})(1 - s_i s_j). \tag{13}$$

Since

$$\sum_{i,j}(EC_{ij} - Avg_{ij}) = 0, \tag{14}$$

we rewrite equation (13) as

$$D = \frac{1}{4}\sum_{i,j}(Avg_{ij} - EC_{ij})s_i s_j. \tag{15}$$

In the following, we define the matrix B to be

$$B_{ij} = Avg_{ij} - EC_{ij}. \tag{16}$$

It is obvious that the matrix B is also symmetric. So equation (15) could be rewritten in the matrix form

$$D = \tfrac{1}{4}s^T Bs. \tag{17}$$

Lemma 1 (SymMtrx)**.** *Let A be a real symmetric matrix, then the following properties hold.*

(1) All eigenvalues of A are real.
(2) We can take each eigenvector to have only real entries.
(3) If u and v are eigenvectors of A associated with different eigenvalues, then u and v are orthogonal.
(4) One can always construct eigenvectors $v_1, v_2, \ldots, v_n$ that are orthogonal and of unit norm

$$v_i^T v_j = \begin{cases} 0 & i \neq j \\ 1 & i = j. \end{cases} \tag{18}$$

Since B is a real symmetric matrix, according to Lemma 1, all eigenvalues of B are real. Moreover, it has a set of normalized orthogonal eigenvectors, denoted by v_i. We let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be the corresponding eigenvalues. Without loss of generality, assume that $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n$. Therefore, we may write s as the linear combination of v_i:

$$s = \sum_{i=1}^{n} a_i v_i. \tag{19}$$

Written in a matrix form $s = Va$, where $a = [a_1, a_2, \ldots, a_n]^T$ and $V = [v_1, v_2, \ldots, v_n]$. Thus,

$$a = V^{-1}s = V^T s. \tag{20}$$

Then,

$$a_i = v_i^T s. \tag{21}$$

Since $s^T s = n$ and $\sum_{i=1}^{n} a_i^2 = n$, we have

$$D = \frac{1}{4}\sum_i a_i v_i^T B \sum_j a_j v_j \tag{22}$$

and

$$D = \frac{1}{4}\sum_{i,j} a_i a_j v_i^T B v_j. \tag{23}$$

Since $Bv_j = \lambda_j v_j$,

$$D = \frac{1}{4}\sum_{i,j} a_i a_j v_i^T \lambda_j v_j. \tag{24}$$

Moreover,

$$v_i^T v_j = \begin{cases} 0 & i \neq j \\ 1 & \text{otherwise.} \end{cases} \tag{25}$$

Hence, equation (24) can be rewritten as

$$D = \frac{1}{4} \sum_{i=1}^{n} a_i^2 \lambda_i. \tag{26}$$

Therefore, the issue of maximizing D is equal to the task of choosing the proper a_i such that the weight put on the term corresponding to the largest eigenvalue in equation (26) is as much as possible.

If there is no constraint of s, we would choose s proportional to the eigenvector corresponding to the largest eigenvalue. However, the world is not perfect. As described before, s could only be 1 or -1, which means that, in most cases, s could not be chosen proportional to v_1. However, a good approximate solution exists by choosing s to be as close to parallel with v_1 as possible, which is achieved by setting

$$s_i = \begin{cases} 1 & \text{if } v_{1i} \geqslant 0 \\ -1 & \text{otherwise,} \end{cases} \tag{27}$$

where v_{1i} is the ith element of the vector v_1.

Here comes the algorithm for dividing the networks into two hierarchies:

Algorithm 1.

(1) Compute each node's centrality according to the specific measure proposed in section 2.
(2) Using the centralities, create the matrix B, described in section 3.
(3) Compute the eigenvalues and eigenvectors of B, then choose s according to equation (27).
(4) If the division contains one empty group, return 'false', otherwise 'true'.

To the end, we have dealt with the simplest case of dividing the networks into two hierarchies. When extending to the case of more than two hierarchies, at first sight, we may first call algorithm 1 to divide the networks into two hierarchies and then feed each hierarchy to algorithm 1 again and so forth. However, this method does not work very well. The fundamental problem is that we should divide the subgraph according to the centrality of nodes in the whole network, not in the subgraph. Therefore, when applying algorithm 1 to the subgraph, we use the centrality vector of nodes, $EC(i)$, in the original graph but not in the subgraph. In practice, the method is much better than the former one.

The present method has several advantages which are as follows.

(1) It does not depend on the definite node centrality.
(2) It does not give any constraint of the graph itself. Any kinds of graphs, such as directed graph and weighted graph, can use our algorithm.

The above features of the present method make it a good choice in practical application.

4. Simulations

In this section, we present a number of tests of our algorithm on real-world networks. In each case, we find that our algorithm reliably detects the hierarchical structure.

The first example is drawn from a network describing the football player market. Figure 1 describes the 22 soccer teams participating in the World Championship in Paris, 1998 [30]. Players of the national team often have contracts in other countries. This constitutes a player

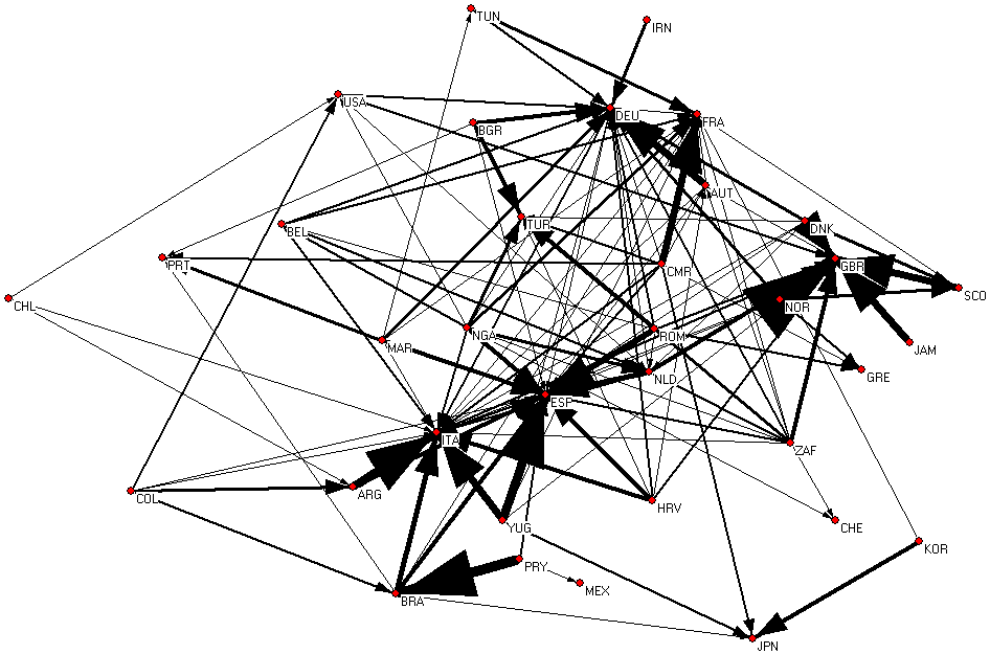

Figure 1. The network example describes the 22 soccer teams which participated in the World Championship in Paris, 1998. If there is one arc from node i to node j, it means that there are players exported from country i to country j. The thickness of line represents the number of players.

market where national teams export players to other countries. Members of the 22 teams had contracts in altogether 35 countries. Counting which team exports how many players to which country can be described with a valued and asymmetric graph. The graph is highly asymmetric: some countries only export players, while some countries are only importers. Figure 2 illustrates the adjacency matrix of the example describing the 22 soccer teams, participating in the World Championship in Paris, 1998.

At first, we use the betweenness centrality as our measure for centrality. The result, described by figure 3, is as follows.

First, we focus our attention on the first division generated by our algorithm. The first division is denoted as $R1$ and $R2$. The set of $R1$ contains only five entries DEU, ITA, ESP, FRA, GBR which represents the countries Germany, Italy, Spain, France, England, respectively. The division sounds reasonable since all the five countries have the top football leagues: Germany Bundesliga, Italy Seria A, Spain Primera divisiń de Liga, France Le Chanpionnat and England Premiership. More interesting, it is worth noting that Italy, Germany and France took part in the semi-final of the World Cup 2006 and Italy and France were the two teams appearing in the final. Even England and Spain attended the quarterfinal. It seems that the hierarchical structure of the football players' market can be treated partly as a representative of football levels in countries.

Now let us turn to other divisions. R_{2111} = (BGR, CHL, CMR, COL, DNK, HRV, IRN, JAM, KOR, MAR, NGA, NOR, PRY, ROM, YUG). We are interested in this set since most of the Afro-Asian countries are in it. The main reason for this division is that the football levels in these countries are relatively lower compared with the other countries which participated in the World Cup 1998 as a whole. Among them, Japan is the only exception. Due to the

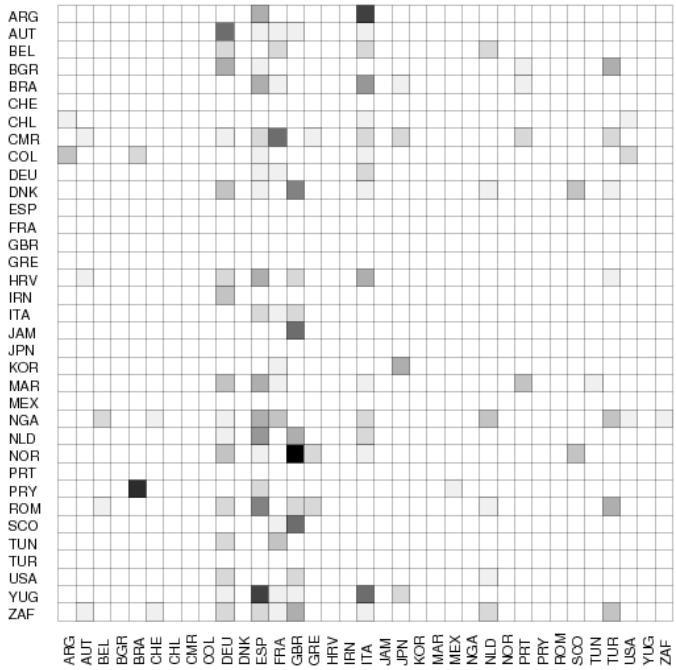

Figure 2. The adjacency matrix of network example describing the 22 soccer teams which participated in the World Championship in Paris, 1998.

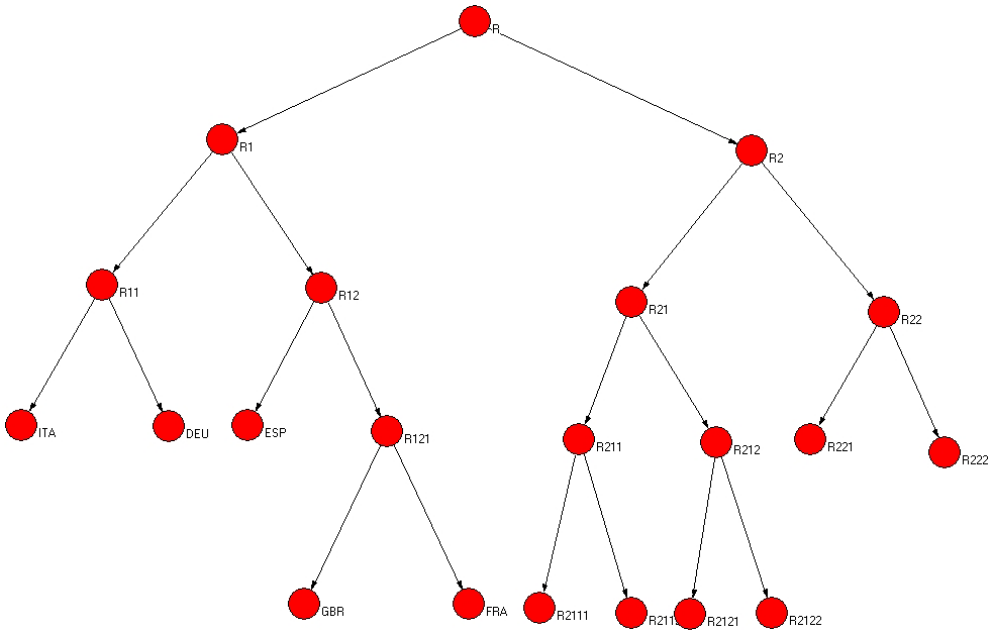

Figure 3. The hierarchical structure of network described in figure 1. Each division is described by a layer of the tree.

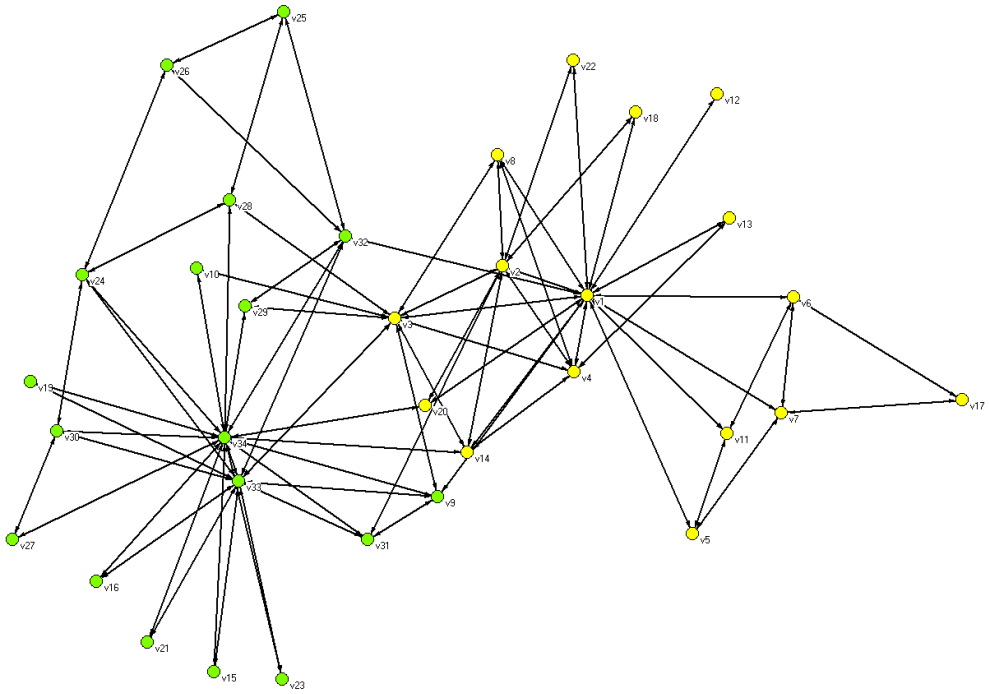

Figure 4. The network from Zacharys karate club study as described in the text. The nodes associated with the club administrator's faction are drawn in green and those associated with the instructor's faction are drawn in yellow.

financial support and good policies, Japan football has improved a lot recently. And the high reward has attracted many famous players to take part in the J-league, Japanese top football league. Hence, Japan plays a much more central role in the football players' market than any other Afro-Asian countries, leading Japan to appear in the set $R_{221} =$ (AUT, JPN, PRT, TUR) with some football-developed countries such as Portugal and Turkey.

When mentioning football, the world champion Brazil should not be omitted. In our divisions, Brazil appears in the set $R_{222} =$ (BRA, NLD, USA). It is not surprising that Brazil does not appear in the set R_1 because Brazilian football league is not that attractive compared to the five famous European leagues listed above. Most of the excellent players in Brazil join the teams of these leagues instead of native clubs. The examples are Ronaldo working for Real Marid in Spain, Kaka for AC Milan in Italy and so forth. Thus, Brazil does not gain much centrality in the football market. However, anybody admits that they are in the centre of the football world even when they failed to get the championship of World Cup 2006 in Germany.

However, when using the eigenvector centrality as the measure of centrality, the result is rather poor. The main reason lies in that the eigenvector centrality is not suitable for being the measure in this case. The eigenvector centrality of the football player market network is [0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0], which says that Columbia, the ninth entry, is in the centre of the football players' market. Obviously, it conflicts with our intuition. The reason for the inefficacy of the eigenvector centrality is described in [31].

In the following, we illustrate the possibility of using other measures in our algorithm, other than centrality measures. The example is drawn from the well-known karate club study

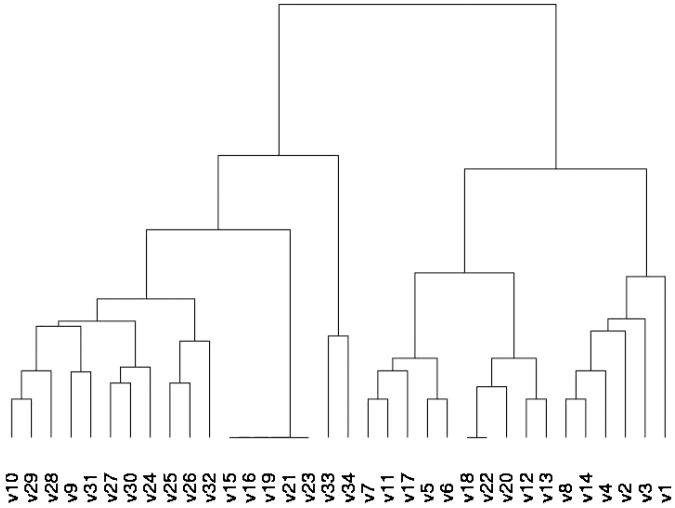

Figure 5. Hierarchical tree calculated by using the algorithm presented in this paper.

of Zachary [32]. In this study, Zachary observed 34 members of a karate club over a period of 2 years. During the course of the study, a disagreement developed between the administrator of the club and the club's instructor, which ultimately resulted in the instructor leaving and starting a new club, taking about a half of the original clubs members with him. Figure 4 shows the network with the instructor and administrator represented by nodes 1 and 34, respectively. The nodes associated with the club instructor's faction are drawn in yellow and those associated with the administrator are drawn in green. In this case, EC_{ij} is defined as follows:

$$EC_{ij} = \begin{cases} 1 & \text{if there is an edge between nodes } i \text{ and } j \\ 0 & \text{otherwise.} \end{cases} \tag{28}$$

Figure 5 shows the result of hierarchy generated by the current algorithm. As indicated by the first division, our result reveals actual divergence, marked by different colours in figure 4. Furthermore, when using the eigenvector centrality as the measure, unfortunately it fails to reveal the actual division. The reason for its failure is similar to the example of the football player market.

Finally, we should mention that the method presented in this paper is approximate but not exact. However, as the simulations show, the present method is accurate in reconstructing the hierarchical structure. Moreover, since the computation complexity lies in the cost of the calculation of the principal eigenvector of matrices, it is rather fast. Hence, our method is good for practical application.

5. Conclusions

In this paper, we have investigated the hierarchical structure in various kinds of networks, by introducing a method for detecting such a structure. We have tested it on two real-world networks with a well-documented structure and found the results to be in excellent agreement with expectations. We hope that the method presented here will be useful in real application in the future.

Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable suggestions and comments and thank Professor Fiedler for his kindness to send a copy of his paper [27]. This work was partly supported by the National Natural Science Foundation of People's Republic of China under grant 60574036, by the Specialized Research Fund for the Doctoral Program of High Education of China under grant 20050055013 and by the Program for New Century Excellent Talents of High Education of China (NCET).

References

[1] Bollobas B 1985 *Random Graphs* (London: Academic)
[2] Watts D J and Stogatz S H 1998 Collective dynamics of 'small' world *Nature* **393** 440–2
[3] Barabasi A L and Albert R 1999 *Science* **286** 509–12
[4] Newman M E J 2003 *SIAM Rev.* **45** 167–256
[5] Albert R Z and Barabaśi A L 2002 *Rev. Mod. Phys.* **74** 47–97
[6] Li C G and Maini P K 2005 An evolving network model with community structure *J. Phys. A: Math. Gen.* **38** 9741–9
[7] Ravasz E, Somera A L, Mongru D A, Oltvai Z and Barabaśi A L 2002 *Science* **297** 1551–5
[8] Broida A and Claffy K C 2001 *Proc. SPIE, Bellingham (Bellingham, WA, 2001)* ed S Fahmy and K Park, pp 172–87
[9] Chen Q, Chang H, Govindan R, Jamin S, Shenker S J and Willinger W 2002 *Proc. 21st Annual Joint Conference of the IEEE Computer and Communications Societies, New York, 2002* (Los Alamitos, CA: IEEE Computer Society)
[10] Albert R, Jeong H and Barabaśi A L 1999 *Nature* **401** 130–1
[11] Amaral L A N and Ottino J M 2004 *Eur. Phys. J.* B **38** 147–62
[12] Moreno Y, Pastor-Satorras R and Vespignani A 2002 *Eur. Phys. J.* B **26** 521–9
[13] Geppertyk U, Riegerz H and Schreckenberg M 1997 *J. Phys. A: Math. Gen.* **30** 393–400
[14] Variano E A, McCoy J H and Lipson H 2004 *Phys. Rev. Lett.* **92** 188701
[15] Gallos L K 2005 *Preprint* physics/0503004
[16] Lazaros K G 2004 *Preprint* cond-mat/0503004
[17] Bolton C and Lowe G 2005 *Theor. Comput. Sci.* **330** 407–38
[18] Yang H J, Zhao F C, Wang W X, Zhou T and Wang B H *Preprint* physics/0508026
[19] Freeman L C 1978 *Soc. Netw.* **1** 215–39
[20] Bonacich P 1972 *J. Math. Sociol.* **2** 113–20
[21] Ruhnau B 2000 *Soc. Netw.* **22** 357–65
[22] Barthelemy M 2004 *Eur. Phys. J.* B **38** 163–8
[23] Rodgers G J, Austin K, Kahng B and Kim D 2005 Eigenvalue spectra of complex networks *J. Phys. A: Math. Gen.* **38** 9432–7
[24] Sousa A O 2004 *Preprint* cond-mat/0406390
[25] Newman M E J 2001 *Phys. Rev.* E **64** 016131–2
[26] Cvetković D M, Doob M and Sachs H 1995 *Spectra of Graphs* (Leipzig: Barth)
[27] Fiedler M 1973 Algebraic connectivity of graphs *Czech. Math. J.* **23** 298–305
[28] Pothen A, Simon H and Liou K P 1990 *SIAM J. Matrix Anal. Appl.* **11** 430–52
[29] Newman M E J 2006 *Proc. Natl Acad. Sci.* **103** 8577–82
[30] Batagelj V and Mrvar A 2006 Pajek datasets. URL: http://vlado.fmf.uni-lj.si/pub/networks/data/
[31] Bonacich P 2001 *Soc. Netw.* **23** 191–201
[32] Zachary W W 1977 *J. Anthropol. Res.* **33** 452–73