Eur. Phys. J. B (2012) 85: 106
DOI: 10.1140/epjb/e2012-21019-2

THE EUROPEAN
PHYSICAL JOURNAL B

Regular Article

Uncovering evolutionary ages of nodes in complex networks

G.-M. Zhu[1,2,a], H.J. Yang[2,3,b], R. Yang[4], J. Ren[1,2], B. Li[1,2], and Y.-C. Lai[2,4,5,c]

[1] NUS Graduate School for Integrative Sciences and Engineering, 117456 Singapore, Republic of Singapore
[2] Department of Physics and Center for Computational Science and Engineering, National University of Singapore, 117546 Singapore, Singapore
[3] School of Business, Shanghai University for Science and Technology, Shanghai 200092, P.R. China
[4] School of Electrical, Computer, and Energy Engineering, Arizona State University, 85287 Tempe, AZ, USA
[5] Department of Physics, Arizona State University, 85287 Tempe, AZ, USA

Received 12 December 2011 / Received in final form 16 February 2012
Published online 26 March 2012 – © EDP Sciences, Società Italiana di Fisica, Springer-Verlag 2012

Abstract. In a complex network, different groups of nodes may have existed for different amounts of time. To detect the evolutionary history of a network is of great importance. We present a spectral-analysis based method to address this fundamental question in network science. In particular, we find that there are complex networks in the real-world for which there is a positive correlation between the eigenvalue magnitude and node age. In situations where the network topology is unknown but short time series measured from nodes are available, we suggest to uncover the network topology at the present (or any given time of interest) by using compressive sensing and then perform the spectral analysis. Knowledge of ages of various groups of nodes can provide significant insights into the evolutionary process underpinning the network. It should be noted, however, that at the present the applicability of our method is limited to the networks for which information about the node age has been encoded gradually in the eigen-properties through evolution.

1 Introduction

Many large, complex networks in existence today are the results of some evolutionary processes such as growth [1]. The Internet is one best example, which has undergone tremendous expansion in the past two decades. Growth in a decentralized manner also appears to be the hallmark of other types of networks such as various biological, social and economical networks (e.g., Facebook). Given a complex network but without any knowledge of its evolutionary history, one might be interested in the distribution of the "ages" of various nodes or subgroups of nodes in the network. Information about the node ages can provide deep insights into the organization and structure of the underlying network, and may have significant applications. For example, in a social network, the lifetimes of certain subgroups of nodes may be closely related to the network backbone structure in terms of the roles that these subgroups play in the function of the network, e.g., leadership roles. In a biological network, nodes of longer lifetimes can be more critical to the various functions of the network. It is thus of considerable interest to develop a systematic method to uncover the evolutionary ages of subgroups of nodes in complex networks.

Two situations arise when addressing the age-detection problem in complex networks: (1) network topology is known and (2) the topology is unknown but only time series measured or observed from various nodes are available. In the first case we shall establish that the spectrum of the network connectivity matrix, or the Laplacian matrix, is directly related to the evolutionary ages of various subgroups of nodes in the network. In the second case, we make use of a recently developed method of time-series based reverse engineering of complex networks [2,3] to uncover the network topology, and then could analyze the spectrum of the predicted Laplacian matrix to obtain estimates of the age distribution of nodes. Our approach thus defines a framework in which the problem of evolutionary-age detection of nodes in complex networks can be addressed in systematic way. While our method does not require a positive correlation between the node degree and age, a correlation between the eigenvalue and the node age is necessary.

It is useful to point out that for the class of scale-free networks that are generated according to the preferential-attachment rule [4], the problem of evolutionary-age estimation may be trivial. In particular, this growth rule stipulates that the probability for an existing node to acquire new links is proportional to its degree, implying a strong correlation between the node degree and its lifetime. Thus, for a scale-free network evolved predominantly according to the preferential-attachment rule, the ages of various nodes can be predicted simply by examining the

[a] e-mail: zhugm07@gmail.com
[b] e-mail: hjyang@usst.edu.cn
[c] e-mail: Ying-Cheng.Lai@asu.edu

degrees. However, many real-world networks deviate significantly from the scale-free topology [1] and, for them the problem of detecting node evolutionary ages is nontrivial. Nonetheless, scale-free networks provide an ideal testbed to validate our spectrum-analysis method.

We emphasize that, although our method is suitable even for networks for which there is no positive correlation between node degree and age, its applicability is limited to networks for which there is a positive correlation between the properties of the eigenmodes and the node age. For networks with which no evolutionary process can be affiliated, such as various citation networks and twitter-type of social networks where the importance of a node may not be related with its age, our method is not applicable.

In Section 2, we describe the main idea underlying our method. In Section 3, we validate the method by using scale-free networks generated by the standard preferential-attachment rule and by the duplication/divergence mechanism, which are especially relevant to social and biological systems, respectively. In Section 4, we consider a realistic biological network, the protein-protein interaction network for which the age distribution of nodes is available, to further validate our method. In Section 5, we address the situation where the network topology is not known a priori but only time series are available, make use of the reverse-engineering approach [2,3] to map out the network topology, and demonstrate that the approach yields correctly and accurately the spectrum of the Laplacian matrix. A brief conclusion is presented in Section 6.

2 Method

For a complex network of N nodes, its topological structure can be described by the Laplacian matrix L [5–9], where the off-diagonal elements of L are $L_{i\neq j} = L_{j\neq i} = 1(0)$ if the nodes i and j are connected (disconnected), respectively. The diagonal elements are $L_{ii} = -\sum_{j\neq i} L_{ij} = -k_i$, where k_i is the number of the nodes connected directly with the node i (node degree). The eigenvalues of L are nonnegative and can be ranked as $0 = \lambda_1 \leq \lambda_2 \ldots \leq \lambda_N$. The corresponding eigenvectors are $X_1, X_2, \ldots, X_N$, whose wavelengths are sorted in a descending order. Each eigenvector contains components concentrated on various nodes in the network.

For a regular or a small-world network [10], the eigenvectors typically exhibit some wave patterns with certain wavelengths [11,12]. When a perturbation is applied to the network, the affected eigenvectors are those whose wavelengths match the size of the perturbation (i.e., the number of nodes that it affects). In this case, some localized structure in the affected eigenvectors can emerge. Eigenvectors associated with small eigenvalues usually have large wavelengths, and so they are sensitive to perturbation on a global scale. In contrast, eigenvectors associated with large eigenvalues are most sensitive to localized perturbations that are applied to a small set of nodes in the network. The responses of the eigenvectors to perturbations thus reflect the structure of the network at different scales. An example is given in Figure 1 for a one-dimensional regular lattice of $N = 100$ nodes with periodic boundary condition, where each node is connected with 2 neighbors on either side so that the node has 4 nearest neighbors. Shown in Figure 1a are representative eigenvectors, where the values of $N \times X_i^2(s)$ are plotted and $X_i^2(s)$ is the sth component of the eigenvector X_i. We see that the eigenvectors represent periodic waves of wavelengths ranging from N to 2. To observe the effect of local structural perturbation on the eigenvectors, we add two more links to each node in the group of nodes whose indices are between 40 and 60 so that each node in this perturbed group now has six nearest neighbors. Let λ'_i $(i = 1, \ldots, N)$ be the eigenvalues in the perturbed network. Figure 1b shows some representative eigenvectors. We observe that the eigenvectors associated with small eigenvalues, e.g., λ'_1, λ'_{20}, λ'_{40}, λ'_{60}, and λ'_{80}, are basically unchanged. However, eigenvectors associated with relatively large eigenvalues, such as λ'_{100}, are strongly altered by the perturbation but the changes are focused on the perturbed group of

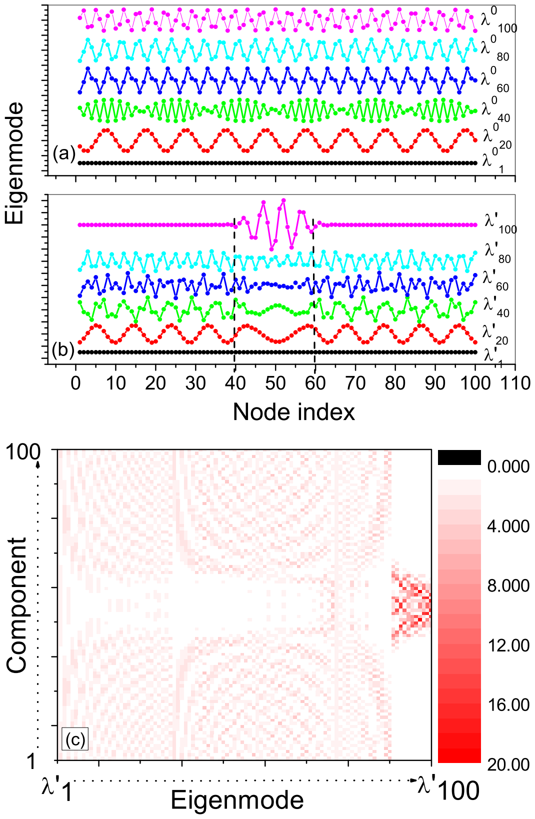

Fig. 1. (Color online) For a regular ring network of 100 nodes where each node has four neighbors, (a) examples of typical, periodic-wave like eigenvectors, (b) typical eigenvectors when each node in the group of indices between 40 and 60 acquires two additional links, one on each side. We observe significant distortions from the periodic-wave pattern, which are localized between the 40th and 60th components of eigenvectors associated with relatively large eigenvalues. (c) Representation of all eigenvectors, where those associated with eigenvalues from λ'_{90} to λ'_{100} are significantly more sensitive to the structural perturbation to the network.

nodes. Figure 1c shows the distribution of the magnitudes of all eigenvectors on nodes in the network, where we see that those associated with eigenvalues λ'_{90} to λ'_{100} are sensitive to the perturbation with large variations appearing on the perturbed nodes.

For complex networks that do not possess a regular backbone, such as random [13] and scale-free [4] networks, the eigenvectors in general do not exhibit any periodic wave structure. Nonetheless, the observation that the eigenvectors associated with larger eigenvalues are more sensitive to structural perturbations can be used to infer the evolutionary age of nodes. To see this, consider a scale-free network evolved according to the preferential-attachment rule [4], for which there is a positive correlation between the node degree and lifetime. That is, nodes of "old" ages tend to have more links and they are thus more susceptible to perturbations applied randomly to the network during the evolutionary process. Since the eigenvectors of large eigenvalues are quite sensitive to perturbations (cf., Fig. 1), we expect the large-degree nodes to dominate these eigenvectors. As a result, large eigenvalues tend to correspond to nodes of long lifetime. This argument suggests that, nodes having the most significant components of the eigenvectors associated with the largest eigenvalues are likely to possess the longest lifetime in the network.

3 Validation using scale-free networks

To exemplify the relation between eigenvalues and node ages, we consider standard scale-free networks [4]. Each network has $N = 2000$ nodes, which is evolved following the preferential-attachment rule so that the age of the ith node is $N - i + 1$. For a given eigenvalue, the lifetime of the associated eigenvector is the average age of all nodes contained in the vector, weighted by the respective components of the eigenvector. Figures 2a–2c show the ages of the eigenvectors X_i versus the index i for three networks of different edge density w. The significant feature common to all three cases is that the average age of the nodes dominating some eigenvector increases on average with the eigenvalue. The average degree of each eigenvector, i.e., the weighted average of the degrees of all nodes associated with the vector, shows the same tendency, as shown in Figures 2d–2f, where the average degree is presented on a logarithmic scale. For each network, the sizes of the eigenvectors are shown in Figures 2g–2i, where the size of an eigenvector is defined to be the number of nodes on which the vector component is larger than a small threshold value. For sufficiently dense network, e.g., Figure 2i, the size tends to decrease on average with the eigenvalue, indicating that a small group of nodes have extraordinarily long lifetimes in the network and their relative ages can be identified simply by examining the associated eigenvalues. Figures 2j–2l show, for $W = 2, 4$ and 8, respectively, the average evolution age versus the node degree. We observe an approximately monotonic relation for small degree. However, when the node degree is larger than 10, the relation deteriorates quickly and the relations approach a constant.

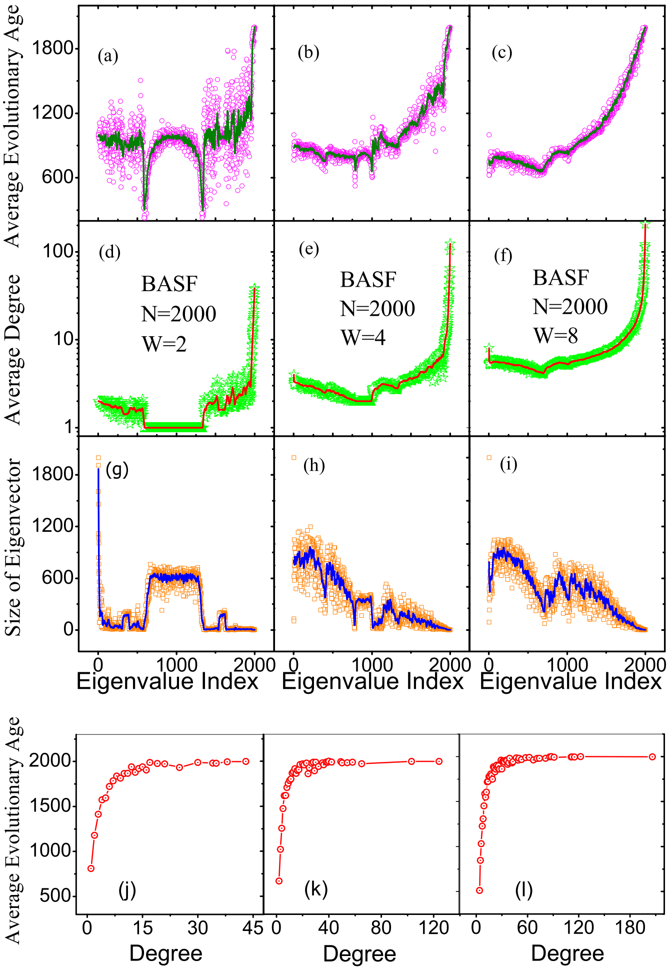

Fig. 2. (Color online) For three scale-free networks generated according to the standard preferential-attachment rule with edge density $w = 2, 4, 8$ (corresponding to the left, middle, and right column, respectively), (a–c) average ages, (d–f) average degree (on a logarithmic scale), and (g–i) size of eigenvector versus the eigenvalue index i. Eigenvectors associated with large eigenvalues generally have small sizes, but their ages are "older" in the network. (j–l) Average age versus degree. We see that, while small degree is related with the average age, information about node age deteriorates quickly as the degree is increased.

To further demonstrate our method, we have analyzed a scale-free cellular network generated by mechanism different than that of the preferential-attachment rule, namely the protein-protein interaction (PPI) networks. In such a network, duplication and divergence are believed to be responsible for the topological structure [14]. We start from a small, connected graph as a seed and duplicate a randomly selected existing protein at each step. The new comer duplicates exactly the connection pattern of its generator in the network. Due to mutations, some of the duplicated edges are broken with probability p, while new edges are generated with probability q between the new comer and other existing nodes. To compare with the PPI network of the Baker's Yeast (to be described in the next Section), we generate networks with

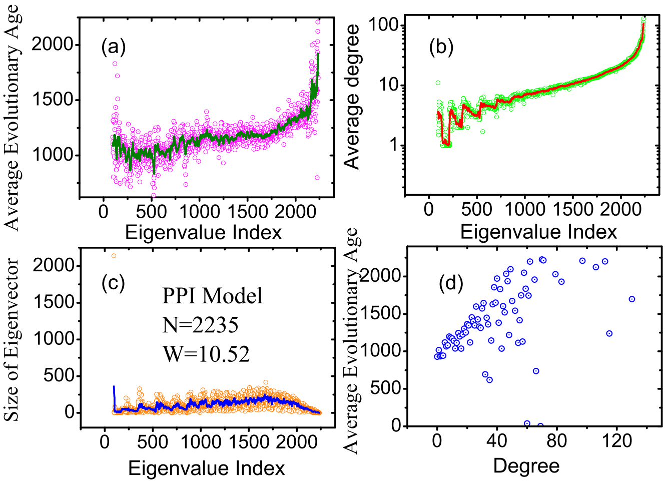

Fig. 3. (Color online) For scale-free networks generated by duplication/divergence-based mechanism from PPI network of the Baker's Yeast, (a) average age versus eigenvalue index, (b) average degree versus eigenvalue index, and (c) size of eigenvector versus the eigenvalue index. Eigenvectors associated with large eigenvalues generally have small sizes, but their ages are "older" in the network. (d) Average age versus degree. Because of large fluctuation, the degree cannot give age-related information, but the eigenvalues can.

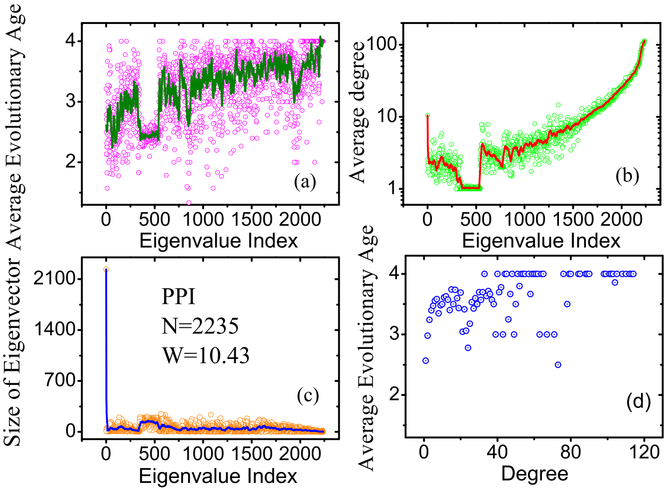

Fig. 4. (Color online) For the largest connected component of the PPI network of the baker's yeast with 2235 nodes, (a) the evolutionary age, (b) average degree (on a logarithmic scale), and (c) size of eigenvector versus the eigenvalue index i. These results further indicate that the evolutionary ages of various nodes in the network can be inferred from the eigenvalue spectrum of the Laplacian matrix. (d) Average age versus degree. We see that degree does not reveal age-related information.

comparable parameters. In particular, a typical network has 2235 nodes and average degree of 10.52, and degree distribution follows power-law with exponent 2.3. In a wide range of eigenvalues there exists a strong correlation between the eigenvalue and average age, as shown in Figure 3a. We observe that, the curve of average age versus degree exhibits large fluctuations, as shown in Figure 3d. It is thus not possible to obtain information about node age from degree. However, behaviors of the eigenmodes can reveal the age information, as will be demonstrated in Section 4.

4 Evolution ages of nodes in a protein-protein interaction network

To lend more credence to our proposition that the evolutionary ages of nodes can be inferred from the eigenvalues, we now consider a class of networks in systems biology, protein-protein interaction (PPI) networks. These networks are the result of a number of evolutionary mechanisms such as duplications of genes and reattachments of links between the proteins. Specifically, we analyze the PPI network of the baker's yeast (*Saccharomyces cerevisiae*) [15,16]. Von Mering et al. [17] analyzed a total of 80 000 interactions among 5400 yeast proteins reported previously and assigned each interaction a confidence value. In order to reduce the effect of false positives, we focus on 11 855 interactions with high and medium confidence values among 2617 yeast proteins. In a PPI network, each protein is a node and each pairwise interaction represents a link between two nodes. Since our goal is to assess, through the eigenvalues, the evolutionary ages of the nodes, we neglect the directions of the edges. The largest connected component of the PPI network contains 2235 nodes. In systems biology, the evolutionary processes of the proteins are classified into four iso-temporal groups [18]: prokaryotes, eukarya, fungi, and yeast, to which numbers 4, 3, 2 and 1 are assigned according to their evolutionary process from ancient to modern times, respectively. The evolutionary age of a protein is the largest number from the groups it presents. For example, the protein YHR037w occurs in the groups prokaryotes(4), eukarya(3), fungi(2), which means that it can be found from the ancient prokaryotes, so that its age is 4. Figure 4a shows the average evolutionary age of nodes in eigenvector versus the eigenvalue index, which is similar to the behavior in Figures 2a–2c. This suggests that for a realistic biological network, there is indeed a positive correlation between the eigenvalues of the Laplacian matrix and the evolutionary ages of groups of nodes. Since PPIs typically possess a scale-free structure [19], we expect the average degree of groups of nodes to exhibit similar behaviors as in Figures 2d–2f. This is indeed the case, as shown in Figure 4b. The sizes of various eigenvectors are shown in Figure 4c. Again the behavior is similar to those in Figures 2g–2i. From Figure 4d, relation of average age versus degree, we see that the degree contains no information about the node age.

5 Time-series based detection of evolutionary ages of nodes

We now address the situation where the network topology is unknown but only time series measured or observed from various nodes are available. We shall apply

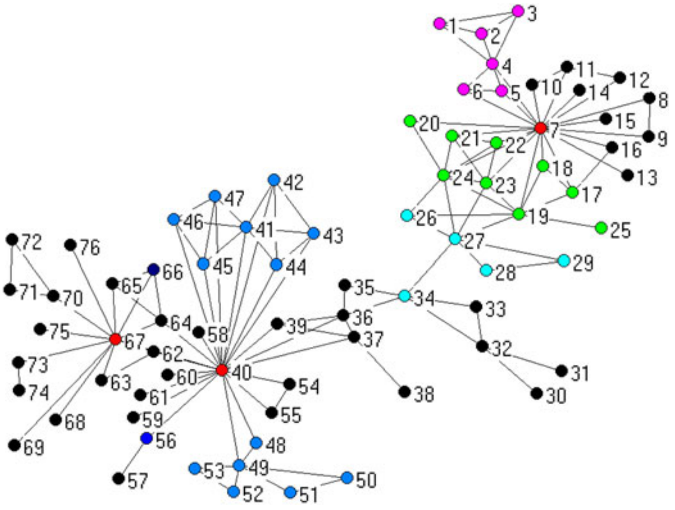

Fig. 5. (Color online) Schematic illustration of the largest component of the SFI collaboration network and the clustered structure revealed by an eigenvalue/eigenvector analysis.

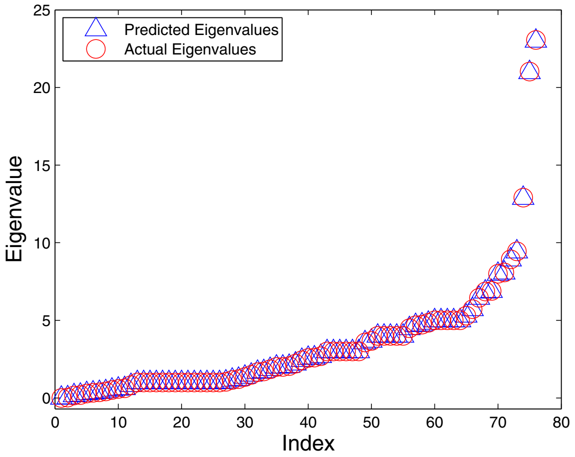

Fig. 6. (Color online) Sorted eigenvalues of the predicted and actual Laplacian matrix of the SFI collaboration network. The number of data points used in uncovering the network structure is about 40% of the number of total unknown coefficients in the power-series expansion.

a recently developed approach [2,3] based on compressive sensing [20–26] to uncover the complex-network topology and then could analyze the spectrum of the predicted Laplacian matrix to estimate the evolutionary ages of nodes. The unique feature of compressive sensing lies in its extremely low data requirement: very little observation is needed to obtain a target sparse signal. In general, the problem of compressive sensing can be described as to reconstruct a sparse vector $\mathbf{a} \in R^N$ from linear measurements $\mathbf{X}$ about $\mathbf{a}$ in the form: $\mathbf{X} = \mathbf{G} \cdot \mathbf{a}$, where $\mathbf{X} \in R^M$ and $\mathbf{G}$ is an $M \times N$ matrix. Accurate reconstruction can be achieved by solving the following convex optimization problem [20,21]

$$\min \|\mathbf{a}\|_1 \quad \text{subject to} \quad \mathbf{G} \cdot \mathbf{a} = \mathbf{X}, \tag{1}$$

where $\|\mathbf{a}\|_1 = \sum_{i=1}^{N} |\mathbf{a}_i|$ is the L_1 norm of vector $\mathbf{a}$ and $M \ll N$, i.e., the number of measurements can be much less than the number of components of the unknown signal. Various solutions of the convex optimization problem (1) have been worked out in the applied-mathematics literature [20–26].

To uncover network topology based on data, it is necessary to cast the problem in the form (1). The basic hypothesis is that a complex networked system can be viewed as a large dynamical system that generates oscillatory time series at various nodes. Under this hypothesis, it is straightforward to formulate the problem under the compressive-sensing paradigm, details of which can be found in reference [2,3].

To give a concrete example, we consider a real-world network, the Santa Fe Institute (SFI) collaboration network [27]. There are $N = 76$ nodes in the largest connected component of the network and the average degree is about 3. A schematic illustration of the network is shown in Figure 5. A spectral analysis reveals that the eigenvectors associated with λ_{76}, λ_{75} and λ_{74} characterize the three hubs: 40, 7 and 67, all marked by red. The eigenvector associated with λ_{73} involves a group of nodes numbered between 17 and 25 (marked by green). For λ_{72}, the corresponding eigenvector covers nodes 26 to 29, and node 34 (marked by cyan). The three clusters: nodes 41 to 47 (blue), 1 to 6 (magenta), and 48 to 53 (violet), are represented by eigenvectors λ_{70}, λ_{69}, and λ_{68}, respectively. In fact, clusters of larger scales can be identified for smaller eigenvalues.

Now assume that the network topology is unknown but an oscillatory time series from each node is available. To simulate the situation, we assume that the dynamics of each node is described by the chaotic Rössler oscillator [28]. Applying the compressive-sensing based method to uncover the network topology, we can then perform a spectral analysis to estimate the ages of various nodes in the network. Figure 6 shows the eigenvalues of the predicted and the actual Laplacian matrix. We observe an excellent agreement.

6 Conclusions

In summary, we have developed a procedure to estimate the evolutionary ages of nodes in complex networks. The basic observation is that eigenvectors associated with different eigenvalues of the Laplacian matrix can typically represent highly localized groups of nodes in the network. A qualitative argument can then be made for the existence of positive correlation between the node ages and the magnitudes of the eigenvalues. This means that, when the network topology is known, a simple eigenvalue analysis can lead to reliable information about the age distribution of nodes in the network. For situations where the network topology is unknown but time series from nodes are available, it is necessary to uncover the topology in order to estimate the node ages, and we have demonstrated that this can be done efficiently using compressive sensing.

Examples from model and real-world networks, including a PPI network, are used to validate our approach. We hope our method to find applications in fields such as systems biology, the propagation of a rumor, a fashion, a joke, or a flu, where estimating node ages can be of significant value.

The network-reconstruction technique used in our work is based on compressive sensing, which works for situations where the types of mathematical forms of the nodal dynamical systems and coupling functions are known (although details of these functions are not required) and can be represented by series expansion. So far the method has not been applied to gene-regulatory networks due to difficulty to find suitable series expansions. The recent method by Hempel et al. [29] is based on extracting statistical information and has been demonstrated to work well for gene-regulatory networks.

While many real-world systems such as gene regulatory and supply chain networks are directed, our present work focused on undirected networks. The main consideration is that many networks generated by some kind of evolutionary processes or constructed through experiments tend to undirected. For example, the Baker Yeast obtained through the approach of prey and predator contains no information about the directionality of the nodal interactions. Our method is based on the observation that local structures, e.g., densely connected clusters, can induce large components in the eigenvectors. Hubs or clusters of hubs can then be detected by the eigenvectors corresponding to largest eigenvalues, while clusters of larger sizes can be uncovered by eigenvectors of smaller eigenvalues. Different eigenmodes can be used to detect clusters of varying scales, providing a correlation with the evolutionary ages in situations where hubs or clusters of hubs are formed by history. The principle on which our method is based thus does not take into account directionality in the node-to-node interactions. To develop a method to uncover the evolutionary ages for directed complex networks remains to be an interesting but open question at the present.

HJY was supported by the National Science Foundation of China under Grants No. 10975099 and 10635040, the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning, and the Shanghai leading discipline project under grant No. S30501. YCL was supported by AFOSR under Grant No. FA9550-10-1-0083. We also acknowledge Mr. Liu Sha and Mr. Tao Lin for useful discussions.

References

1. M.J. Newman, *Networks, An Introduction* (Oxford University Press, New York, 2010)
2. W.-X. Wang, R. Yang, Y.-C. Lai, V. Kovanis, M.A.F. Harrison, Europhys. Lett. **94**, 48006 (2011)
3. W.-X. Wang, R. Yang, Y.-C. Lai, V. Kovanis, C. Grebogi, Phys. Rev. Lett. **106**, 154101 (2011)
4. A.-L. Barabási, R. Albert, Science **286**, 509 (1999)
5. G.-M. Zhu, H.J. Yang, C.Y. Yin, B. Li, Phys. Rev. E **77**, 066113 (2008)
6. H.J. Yang, F.C. Zhao, B.H. Wang, Chaos **16** (2006)
7. J. Ren, B. Li, Phys. Rev. E **79**, 051922 (2009)
8. J. Ren, W.-X. Wang, B. Li, Y.-C. Lai, Phys. Rev. Lett. **104**, 058701 (2010)
9. S. Jalan, G.-M. Zhu, B. Li, Phys. Rev. E **84**, 046107 (2011)
10. D.J. Watts, S.H. Strogatz, Nature **393**, 440 (1998)
11. J.G. Restrepo, E. Ott, B.R. Hunt, Phys. Rev. Lett. **93**, 114101 (2004)
12. K. Park, L. Huang, Y.-C. Lai, Phys. Rev. E **75**, 026211 (2007)
13. P. Erdös, A. Rényi, Publ. Math. Inst. Hung. Acad. Sci. **5**, 17 (1960)
14. A.V. Vazquez, A. Flammini, A. Maritan, A. Vespignani, Complexus **1**, 38 (2003)
15. A. Wagner, Mol. Biol. Evol. **18**, 1283 (2001)
16. A. Wagner, Proc. R. Soc. Lond. B Biol. Sci. **270**, 457 (2003)
17. C. von Mering, R. Krause, B. Snel, M. Cornell, S.G. Oliver, S. Fields, P. Bork, Nature **417**, 399 (2002)
18. C.R. Woese, Microbiol. Rev. **51**, 221 (1987)
19. E. Ravasz, A.L. Somera, D.A. Mongru, Z. Oltvai, A.-L. Barabási, Science **297**, 1551 (2002)
20. E. Candès, J. Romberg, T. Tao, IEEE Trans. Inf. Theory **52**, 489 (2006)
21. E. Candès, J. Romberg, T. Tao, Commun. Pure Appl. Math. **59**, 1207 (2006)
22. E. Candès, in *Proceedings of the International Congress of Mathematicians* (Madrid, Spain, 2006)
23. D. Donoho, IEEE Trans. Inf. Theory **52**, 1289 (2006)
24. R.G. Baraniuk, IEEE Signal Process. Mag. **24**, 118 (2007)
25. E. Candès, M. Wakin, IEEE Signal Processing Mag. **25**, 21 (2008)
26. E. Candès, T. Tao, IEEE Trans. Inf. Theory **51**, 4203 (2005)
27. M. Girvan, M.E.J. Newman, Proc. Natl. Acad. Sci. USA **99**, 7821 (2002)
28. O.E. Rössler, Phys. Lett. A **57**, 397 (1976)
29. S. Hempel, A. Koseska, J. Kurths, Z. Nikoloski, Phys. Rev. Lett. **107**, 054101 (2011)