International Journal of Bifurcation and Chaos, Vol. 22, No. 10 (2012) 1250236 (12 pages)

DOI: 10.1142/S0218127412502367

PROBING COMPLEX NETWORKS FROM MEASURED TIME SERIES

LIANG HUANG
Institute of Computational Physics and Complex Systems, Lanzhou University, Lanzhou, Gansu 730000, P. R. China
School of Electrical, Computer, and Energy Engineering, Arizona State University, Tempe, AZ 85287, USA

YING-CHENG LAI
School of Electrical, Computer, and Energy Engineering, Department of Physics, Arizona State University, Tempe, AZ 85287, USA

MARY ANN F. HARRISON
Physical Sciences and Technology, Scientific Research Group, WVHTC Foundation, Fairmont, WV 26554, USA

Received February 9, 2011

We propose a method to detect nodes of relative importance, e.g. hubs, in an unknown network based on a set of measured time series. The idea is to construct a matrix characterizing the synchronization probabilities between various pairs of time series and examine the components of the principal eigenvector. We provide a heuristic argument indicating the existence of an approximate one-to-one correspondence between the components and the degrees of the nodes from which measurements are obtained. The striking finding is that such a correspondence appears to be quite robust, which holds regardless of the detailed node dynamics and of the network topology. Our computationally efficient method thus provides a general means to address the important problem of network detection, with potential applications in a number of fields.

Keywords: Network detection; topological reconstruction; time series analysis.

1. Introduction

The ubiquity of complex networks in nature and in technological systems, and the studies of dynamical processes on networks have motivated the "inverse" problem of network detection. That is, given that some information about the network dynamics is available, can the intrinsic structure of the network be inferred? This "inverse" problem of network detection is somewhat opposite to the current trend in research on complex networks, where the focus has been on what kinds of interesting dynamics can occur on networks whose topologies and node-to-node interactions are fully known. Comparing with such "direct" network-dynamics problem, the inverse problem, in spite of its importance, has received relatively little attention. In particular, novel approaches have been proposed recently [Yu *et al.*, 2006; Timme, 2007; Napoletani & Sauer, 2008; Wang *et al.*, 2009; Ren *et al.*, 2010]. Depending on the objective, the extent of the structural information needed can be quite different, leading to a diverse range of dynamical information required for solving the inverse problem. For example, in [Yu *et al.*, 2006], the detailed dynamical processes on each node of a network are assumed to be known. A replica of this network, or a computational model

model of this "target" network, can then be constructed, with the exception that the interaction strengths among the nodes are chosen randomly. It has been demonstrated that in situations where a Lyapunov function for the network dynamics exists, the connectivity of the model network converges to that of the target network [Yu *et al.*, 2006]. In [Timme, 2007], a Kuramoto-type of phase dynamics [Kuramoto, 1984] on a network to be detected is assumed, where a steady-state solution exists. For this system, by linearizing the network dynamics about the steady-state solution, the associated Jacobian matrix can be obtained, which reflects the network topology and connectivity under proper perturbations [Timme, 2007]. In both approaches, *complete* information about the dynamical processes, i.e. equations governing the evolution of *all* nodes, is required. Another feature of these approaches is that the amount of computations required tends to increase significantly with the size of the network. For example, suppose the node dynamics is described by a set of differential equations. For a network of size N, in order for its structure to be detected, the number of differential equations to be solved typically increases with N as N^2. For small sparse networks which are typical in metabolic and genomic pathways, Napoletani and Sauer [2008] found that constrained optimization techniques based on the $L1$ vector norm is superior for inference of the network connections. For dynamical systems in a noisy environment, the correlation of the noisy output can be employed to unveil the noise pathways which reveal the connection topologies [Wang *et al.*, 2009; Ren *et al.*, 2010]. In most cases of previous works, the objective is to detect every link in the network, then one would need all the dynamical details, i.e. the parameters of the oscillator at each node, and to measure the dynamical signals from all the oscillators. Consequently, the computation is demanding, rendering the method applicable only to small networks. However, in many cases, such complete structural information is not needed, e.g. knowing which nodes are the hubs in many cases is sufficient to protect the network. Thus a pressing issue is then to develop a *computationally efficient* methodology to detect only certain important characteristics of the network *based solely on a limited number of measured time series.* The aim of this paper is to address this issue. In this paper, we propose a method employing synchronization probability between the nodes to detect the hubs in the network. The dynamical signals required to construct the synchronization probabilities can have broad forms. When the signal can only be extracted from a subset of the nodes, the method can detect the hub nodes in the subset.

The basic hypotheses of our approach are the following: (1) detailed node dynamics are unknown and only a set of measured signals (time series) is available, (2) the network to be probed can be large but the set of accessible "probing points" might be much smaller than the network size. Since the network, except for this set of points, is assumed to be completely inaccessible, we shall regard these points effectively as "nodes," and our goal is to estimate their relative degrees (number of links). This would allow a set of *importance* nodes, i.e. nodes with relatively large numbers of connections, to be detected among these accessible probing points. We shall first show that, if the underlying network is large (larger than only a few nodes), the largest eigenvalue of the network connection matrix (e.g. the adjacency matrix) is often isolated and its associated eigenvector typically possesses the interesting property that its components have an approximately one-to-one correspondence with the node degrees. The issue then becomes that of obtaining an approximate connection matrix from the set of measured time series. We find that a matrix based on the degrees of synchronization among the time series can serve the purpose effectively which, as we will see, is extremely computationally efficient. The surprising finding is that the methodology turns out to be quite general, as it is effective regardless of the nature of the node dynamics. We envision, among others, the following outstanding problems for which our method can find potential use. In neurobiology, a set of electrodes placed at various probing points of a brain region yields a set of electrical signals. It is of interest to know which point is the most important one in terms of the particular brain activity, e.g. epileptic seizures, so as to guide treatment including surgical operations. In social science applications, one can obtain information about the appearance of certain individuals from a set of cameras placed in public areas, and it is of interest to know which individuals are the most important ones in terms of their connections with other individuals. In a communication network, it is important to be able to identify points where the traffic is likely to get bottlenecked. Accurate detection of these bottleneck points in advance

can result in strategies to effectively prevent traffic congestions on the network.

2. Methods

2.1. *One-to-one correspondence between principal components of adjacency matrix and node degrees*

To gain insight, we consider a large complex network with adjacency matrix $\mathbf{A}$, where $A_{ij} = 1$ if there is a connection between nodes i and j, $A_{ij} = 0$ otherwise, and $A_{ii} = 0$. The degree k_i of node i is given by $k_i = \sum_{j=1}^{N} A_{ij}$. Since $\mathbf{A}$ is symmetric, it has a set of real eigenvalues, thus it can be decomposed as [Wall *et al.*, 2003]: $\mathbf{A} = \sum_i \lambda_i \mathbf{e}_i \mathbf{e}_i^T$, where $|\lambda_1| \geq |\lambda_2| \ldots |\lambda_N|$ are the eigenvalues and $\mathbf{e}_i$'s $(i = 1, \ldots, N)$ are the corresponding eigenvectors. Note that for asymmetric matrix, or even rectangular matrix, such a decomposition also exists, known as the singular value decomposition, with now the eigenvalues being called singular values [Wall *et al.*, 2003]. For a large matrix, the largest eigenvalue is typically much greater in magnitude than the remaining eigenvalues [Farkas *et al.*, 2001]. We thus have

$$\mathbf{A} \approx \lambda_1 \mathbf{e}_1 \mathbf{e}_1^T.$$

Summing over the rows on both sides gives

$$k_j \approx C \lambda_1 e_{1,j} \sim e_{1,j}, \tag{1}$$

where $C = \sum_l e_{1,l}$ is a constant. Our simple reasoning thus suggests a direct, proportional relation between the node degree and the component of the eigenvector associated with the largest eigenvalue of the adjacency matrix. For complex networks occurring in nature and technological systems, such as random networks [Erdős & Rényi, 1959], small-world networks [Watts & Strogatz, 1998], and scale-free networks [Barabási & Albert, 1999], they are typically large. Thus we expect Eq. (1) to hold for complex networks arising in applications. Numerical verification of Eq. (1) for two different types of networks, random and scale-free networks, is provided in Figs. 1(a)–1(d). The random network is the classical Erdős–Rényi network [Erdős & Rényi, 1959] where any pair of

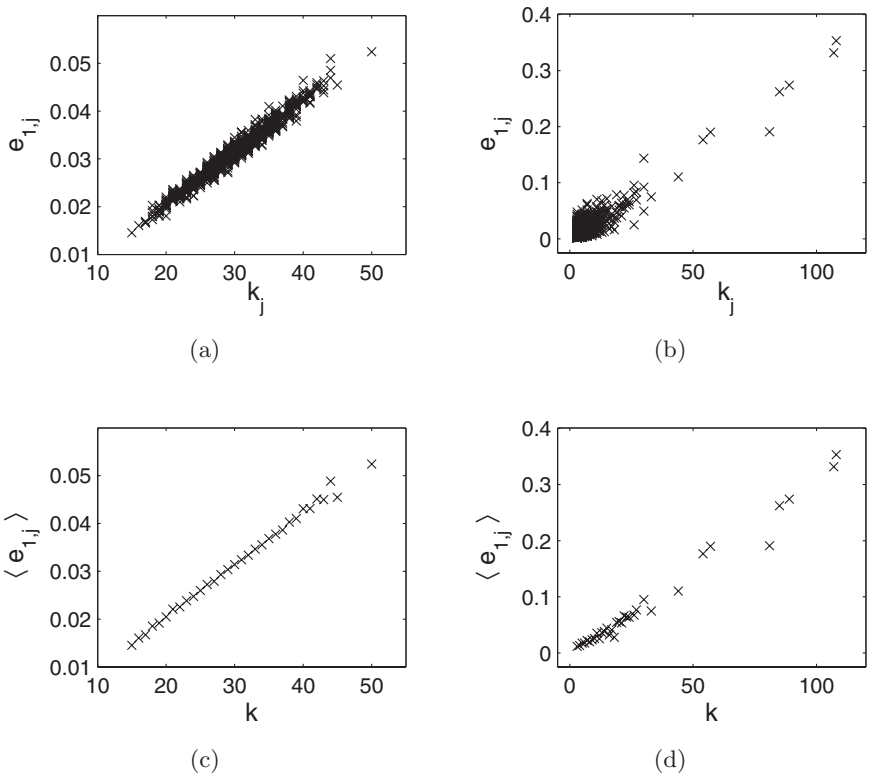


Fig. 1. Components $e_{1,j}$ of the principal eigenvector $\mathbf{e}_1$ of the adjacency matrix versus node degree for (a) a random network and (b) a scale-free network, where the corresponding behaviors of the degree-averaged components $\langle e_{1,j} \rangle$ are shown in (c) and (d), respectively. Both networks have the size $N = 1000$. The average degree is $\langle k \rangle = 30$ for the random network and $\langle k \rangle = 6$ for the scale-free network.

nodes are connected with probability p. The scale-free network is constructed by using the standard preferential-attachment rule with the degree distribution $P(k) \sim k^{-3}$ [Barabási & Albert, 1999].

2.2. *General synchronization probability matrix*

The problem of interest is to probe network structure based on a limited set of measurements. The nature of this problem stipulates that the adjacency matrix is not known *a priori*. A key issue is thus to search for some proper matrix that exhibits a similar behavior to that given by Eq. (1). A criterion is that such a matrix can be constructed based on measured time series *only*. There can be many choices for the matrix. Here, we choose the *general synchronization matrix* as proposed in [Zhao *et al.*, 2005] and demonstrate that it exhibits characteristics of the adjacency matrix in the sense of Eq. (1). In particular, say N points of the network to be probed are externally accessible and N time series $\{x_i(t)\}$ $(i = 1, \ldots, N)$, one from each accessible point, have been measured. For any pair of time series $x_i(t)$ and $x_j(t)$, and a long time interval T_0, the synchronization-probability Φ_{ij} is defined to be the *fraction* of time during which $|x_i(t) - x_j(t)| < \delta$, where δ is a small threshold comparing with the range of the time series [Marwan *et al.*, 2007]. By definition, $\mathbf{\Phi}$ is symmetric and the computation of Φ_{ij} is efficient. The general synchronization-probability matrix $\mathbf{\Phi}$ can be obtained after the synchronization probabilities from all distinct pairs of time series are calculated. In order for the matrix $\mathbf{\Phi}$ to reflect the connectivity of the network, a requirement is that there should be some coherence among the node dynamics. If there is no correlation among the node dynamics, the time series will be random with respect to each other. In this case, the various synchronizabilities will assume small values that are approximately identical for different pairs of nodes. As a result, the matrix $\mathbf{\Phi}$ can be badly conditioned, preventing the node degrees to be detected. In the opposite case where there is a complete synchronization among the nodes, the synchronization probabilities will be close to unity and the matrix $\mathbf{\Phi}$ will be singular. Thus, our method is expected to work in the "weakly-coherent" regime where node dynamics are somewhat correlated but not completely synchronized. Also note that, when the signals are contaminated with noise, the threshold needs to be adapted accordingly [Marwan *et al.*, 2007].

Note that, the proportional relation Eq. (1) holds only for the adjacency matrix $\mathbf{A}$. For the general synchronization-probability matrix, it can be regarded as a randomized adjacency matrix: $\mathbf{\Phi} = c_1\mathbf{A} + c_2\mathbf{\Delta}$, where c_1 and c_2 are positive coefficients that $c_1 + c_2$ is in the order of 1, $\mathbf{\Delta}$ is a symmetric random matrix whose elements are bounded in $[0, 1]$ with mean value μ and standard deviation σ, and $\mathrm{diag}(\mathbf{\Delta}) = 0$. Approximating $\mathbf{\Phi}$ by its first eigenvalue and eigenvector, we have

$$\lambda_1 \mathbf{e}_1 \mathbf{e}_1^T \approx c_1\mathbf{A} + c_2\mathbf{\Delta}.$$

Summing over the rows on both sides yields

$$e_{1,j} \approx \frac{1}{C\lambda_1}(c_1 k_j + c_2 \mu N), \tag{2}$$

where $C = \sum_l e_{1,l}$. The coefficients c_1 and c_2 depend on both topological and dynamical parameters and can be quite different for different systems. Therefore, the component of the eigenvector associated with the largest eigenvalue of the general synchronization-probability matrix has a linear dependence on the node degree, which is superimposed on a plateau (the second term). Since the eigenvector is normalized, $\sum_j e_{1,j}^2 = 1$, the larger the value of the plateau, the smaller the span in the variation of $e_{1,j}$ [i.e. $c_1(k_{\max} - k_{\min})/C\lambda_1$]. Thus when the noise term $c_2\mathbf{\Delta}$ contributes equally to each component $e_{1,j}$, it only reduces the span of the value of the principal eigenvector components while keeping the linear relation untouched (although the slope could be different). In general, due to fluctuation, such an ideal linear relation between principal eigenvector component and node degree cannot be expected. However, insofar as there is an approximate one-to-one correspondence between the eigenvector components and the degrees, the goal of detecting hub nodes — those with large $e_{1,j}$s — can be achieved. Since the range of $e_{1,j}$ can be quite different for different networks or even different dynamical systems, the absolute value of $e_{1,j}$ for one case cannot be used to judge other cases. Due to the approximate nature of the method and other random factors such as noise, such a one-to-one correspondence should be interpreted statistically: a relatively large value of the principal eigenvector component only means a higher probability for the corresponding node to have a large degree. In a practical sense, in order to detect, say, a small set of

m hub nodes, it is necessary to use $n \gg m$ observational points to obtain an $n \times n$ matrix of synchronization probability. Independent measurements are also necessary so that an ensemble of matrices can be obtained, enabling meaningful average values of the principal-eigenvector components to be calculated so as to make the detection of node degrees reliable.

3. Results

To demonstrate the generality of our method, we have carried out extensive computations using continuous-time and discrete-time, regular and chaotic dynamics on random and scale-free networks. Our computations reveal that the base of our method, an approximate one-to-one correspondence apparently holds, regardless of the specific node dynamics and network topology. Here, we shall present five representative examples.

3.1. *Example 1 — Networks of coupled phase oscillators*

This is the classical Kuramoto model [Kuramoto, 1984], where the individual node dynamics is that of a regular phase oscillator:

$$\frac{d\theta_i}{dt} = \omega_i - \varepsilon \sum_{j=1}^{N} A_{ij} \sin(\theta_i - \theta_j),$$

where θ is a phase variable, the frequency ω_i of unconstrained oscillation at node i is random and uniformly distributed in $[-1, 1]$, and A_{ij} is the element of the adjacency matrix. Assume that all these are unknown, and only a set of time series $x_i(t) = \sin\theta_i(t)$ $(i = 1, \ldots, N)$ can be measured. To calculate the general synchronization-probability matrix, we set $T_0 = 3000$ and $\delta = 0.2$ (somewhat arbitrarily, and the result does not depend sensitively on the choice of these parameters). Note that in general, people use a small δ value (e.g. 0.001) to reveal complete synchronization. However, for a less coherent system, a larger value is desirable to mitigate the possible decoherent effects and, in the meanwhile, to characterize certain amount of synchronization. Thus for each pair (i, j), we calculate the time T_{ij} where $|x_i(t) - x_j(t)| < \delta = 0.2$ during the time interval T_0. Then the synchronization probability for pair (i, j) is given by $\Phi_{ij} = T_{ij}/T_0$, from which the general synchronization probability matrix $\mathbf{\Phi}$ is constructed. The eigenvector $\mathbf{e}_1$ associated with the largest eigenvalue of matrix $\mathbf{\Phi}$ can then be obtained. Note that there is a one-to-one correspondence between the component, say, $e_{1,j}$, of the eigenvector $\mathbf{e}_1$ and the index j of the measured time series. Also, index j means that the time series is measured from node j, thus it has a one-to-one correspondence with k_j, the degree of node j. Suppose the mth component of $\mathbf{e}_1$ has the largest value, among all components. If the relation Eq. (1) holds, the mth node would then have the largest possible degree among N accessible nodes. There can, of course, be more than one node with the same degree, thus it is useful to compute the average $\langle e_{1,j} \rangle$ of the eigenvector components over the nodes with the same degrees. Figures 2(a) and 2(b) show the results for the two types of complex networks. We observe an approximate one-to-one correspondence between $\langle e_{1,j} \rangle$ and the node degree, as in Eq. (2).

3.2. *Example 2 — Networks of chaotic Rössler oscillators*

We now consider the coupled nonidentical chaotic Rössler oscillators. The system equations are:

$$\frac{dx_i}{dt} = -(y_i + z_i) + \epsilon \sum_{j=1}^{N} G_{ij} x_j,$$

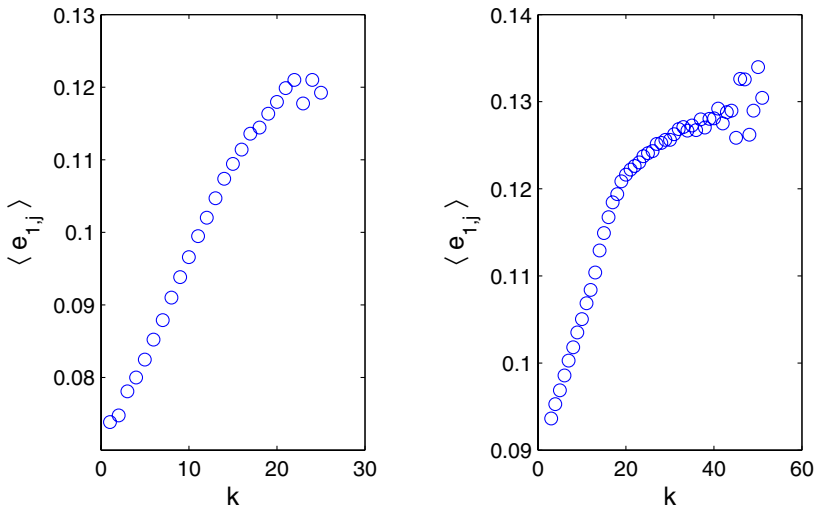
$$\frac{dy_i}{dt} = x_i + a_i y_i,$$

$$\frac{dz_i}{dt} = 0.2 + z_i(x_i - 9),$$

where ϵ is the coupling constant, $\mathbf{G}$ is the coupling matrix such that $G_{ii} = 1$ and $G_{ij} = -A_{ij}/k_i$ for $j \neq i$. The parameter a_i is random but node-specific and it is assumed to distribute uniformly in the interval $(0.16, 0.24)$. After a short transient, we measure the dynamical variable x_i and calculate the synchronization-probability between the nodes. The first eigenvector of the general synchronization-probability matrix is calculated subsequently. The one-to-one relation between the components and the degrees of their corresponding nodes is shown in Fig. 3.

3.3. *Example 3 — Networks of double-scroll Lorenz oscillators*

The above two types of oscillators are bounded and a tiny coupling will lead to nonzero synchronization-probability [Zhao *et al.*, 2005]. Here we study



(a) (b)

Fig. 2. For the classical Kuramoto model, $\langle e_{1,j}\rangle$ versus the node degree for (a) a random network and (b) a scale-free network obtained by using the general synchronization-probability matrix. Both networks have the size $N = 100$. The average degree is $\langle k\rangle = 10$ for the random network and $\langle k\rangle = 6$ for the scale-free network. The coupling parameter is (a) $\varepsilon = 0.25$ and (b) $\varepsilon = 0.3$. Each data point is the average of results from 1000 measurements.

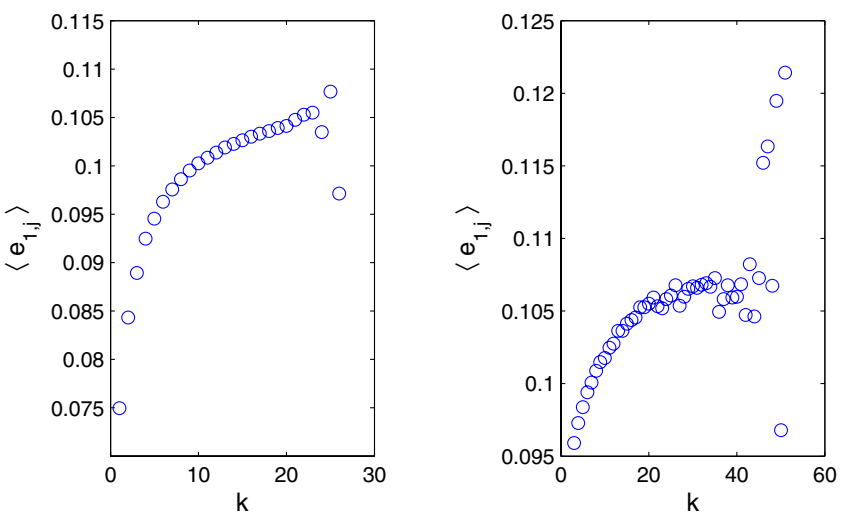
Lorenz oscillator, which has a double-scroll geometry. This geometry requires the coupling to be strong enough to induce certain synchronization. The system equation is given by:

$$\frac{dx_i}{dt} = -\sigma_i(y_i - x_i) + \epsilon\sum_{j=1}^{N} G_{ij}x_j,$$

$$\frac{dy_i}{dt} = x_i(28 - z_i) - y_i,$$

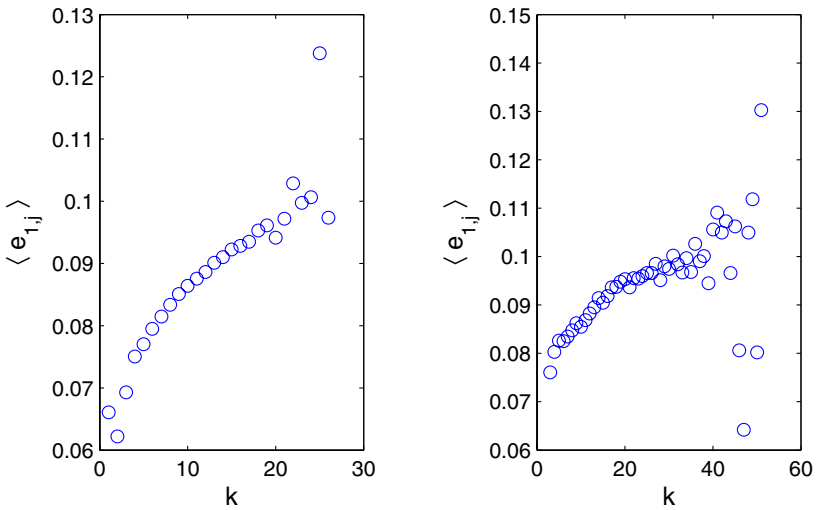
$$\frac{dz_i}{dt} = x_i y_i - \frac{8}{3}z_i,$$

where ϵ is the coupling constant, $G_{ii} = 1$ and $G_{ij} = -A_{ij}/k_i$ for $j \neq i$, σ_i is the Prandtl number and usually takes a value of 10. Here, since



(a) (b)

Fig. 3. For the coupled chaotic Rössler model, $\langle e_{1,j}\rangle$ versus the node degree for (a) a random network and (b) a scale-free network obtained from the general synchronization-probability matrix. The network parameters are the same as in Fig. 2. The coupling parameter is (a) $\varepsilon = 0.5$ and (b) $\varepsilon = 0.05$. $T_0 = 3000$ and $\delta = 1$. Each data point is the average of results from 1000 independent measurements.



(a) (b)

Fig. 4. For the coupled Lorenz oscillators, $\langle e_{1,j}\rangle$ versus node degree k for (a) random network with $\epsilon = 10$, and (b) scale-free network with $\epsilon = 10$. Other parameters are the same as for Fig. 3.

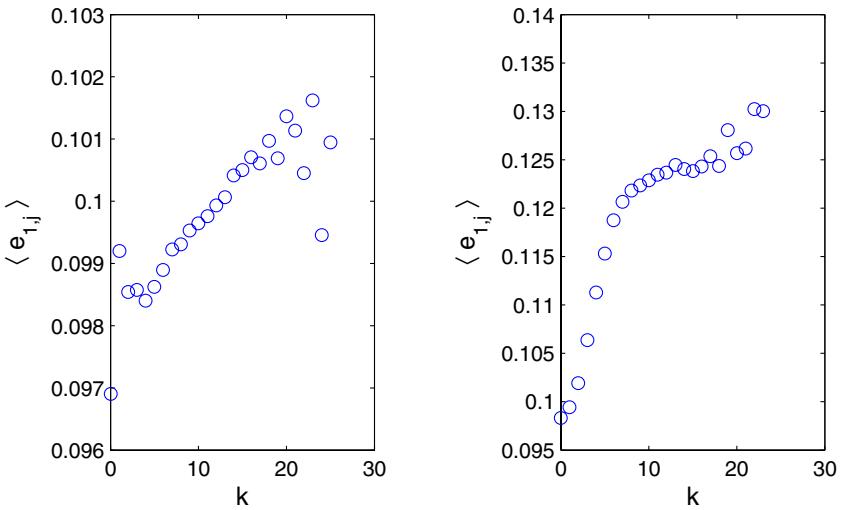
our aim is to study the correlation between non-synchronized oscillators, we introduce a small discrepancy between different oscillators, by taking σ_i as a random number uniformly distributed in the interval $(8, 12)$, where the mean value is 10. Assume x_i is the measurement signals. The relation between the averaged components and degree k for this coupled double-scroll chaotic system is shown in Fig. 4. The one-to-one relation is clear. Thus for either phase oscillator, one-scroll chaotic oscillator, or double-scroll chaotic oscillator, the measurement based component $e_{1,j}$ of the general synchronization-probability matrix can be used as an indicator of the relative magnitude of the degree.

3.4. *Example 4 — Subnetworks of coupled phase oscillators*

In this case, the node dynamics is the same as in Example 1. However, we assume the accessible points are much smaller than the network size N. In particular, for a network of size $N = 1000$, we assume only $n = 100$ nodes are accessible. Correspondingly, the general synchronization probability matrix $\mathbf{\Phi}$ for this subnetwork is constructed and the one-to-one relation between $\langle e_{1,j}\rangle$ and the node degree is demonstrated in Figs. 5(a) and 5(b). Note that here the degree is the number of links a node has in the subnetwork. As the nodes of the subset can be scattered randomly in the original network, then on average, the degree of a node in the subnetwork $k \simeq k_0 n/N$, where k_0 is the degree of the node in the original network. Thus, the hub nodes in the subnetwork are more likely to be the hub nodes in the original network. Viewing the subnetwork as a separated system, the rest of the original network can be regarded as the environment, which sends persistent perturbations into the subsystem (through the couplings). Therefore, the support for the ability to detect hubs in the subnetwork also indirectly substantiates the resistance of the method to external noise.

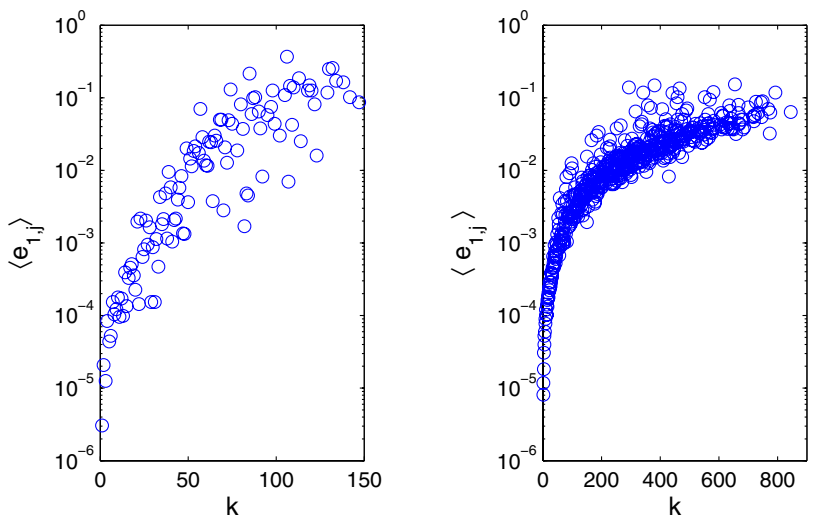
3.5. *Example 5 — Application to real networks*

Here, we consider the actor network where two actors have a link if they ever performed in a movie [Barabási & Albert, 1999]. The data consists of 392 340 actors and $M = 127\,823$ movies. For such a large networked system, it is usually impossible or unnecessary to probe all the nodes. Thus, we focus on a small set of access points, i.e. the first $n = 1000$ actors, which are chosen rather arbitrarily. Since there are no typical dynamical processes on the network, we construct the observable "time series" as follows. First, we assign each movie a number, m, from 1 to M, as the discrete "time" step, and $x_i(m) = 1$ if actor i appeared in movie m, otherwise $x_i(m) = -i$. Then the characteristic matrix $\mathbf{\Phi}$ is constructed from the "time series" x_i and the one-to-one relation between $\langle e_{1,j}\rangle$ and the node degree is demonstrated in Fig. 6.



(a) (b)

Fig. 5. For the classical Kuramoto model with $N = 1000$ oscillators and originally (a) a random network and (b) a scale-free network, $\langle e_{1,j}\rangle$ versus the node degree for a subnetwork of size $n = 100$. The average degree is (a) $\langle k\rangle = 100$ and (b) $\langle k\rangle = 6$ for the original network. The coupling parameter is (a) $\varepsilon = 0.1$ and (b) $\varepsilon = 0.3$. Other parameters are the same as for Fig. 2.



(a) (b)

Fig. 6. For a subset of (a) $n = 1000$ and (b) $n = 10\,000$ of the actor network, $\langle e_{1,j}\rangle$ versus the node degree k in the subnetwork. (a) $\langle k\rangle = 23$ and (b) $\langle k\rangle = 59$. $\delta = 0.1$.

To test the ubiquity of the method, we have additionally carried out simulation on coupled identical and nonidentical logistic maps, and avalanche processes for several distinct network topologies, they all show the one-to-one relation between principal component and the node degree. Therefore, the proposed method to detect hubs can have wide applications independent of the specific dynamic models and connecting topologies.

4. Remarks

Since the method is rather statistical than deterministic, one may expect to select more nodes in order to include a small subset of hub nodes. To be specific, let us consider the actor network as in Sec. 3.5. Figure 7(a) shows $e_{1,j}$ of the general synchronization probability matrix versus k_j without doing average. The one-to-one relation between $e_{1,j}$ and k_j holds only approximately. As the relation is

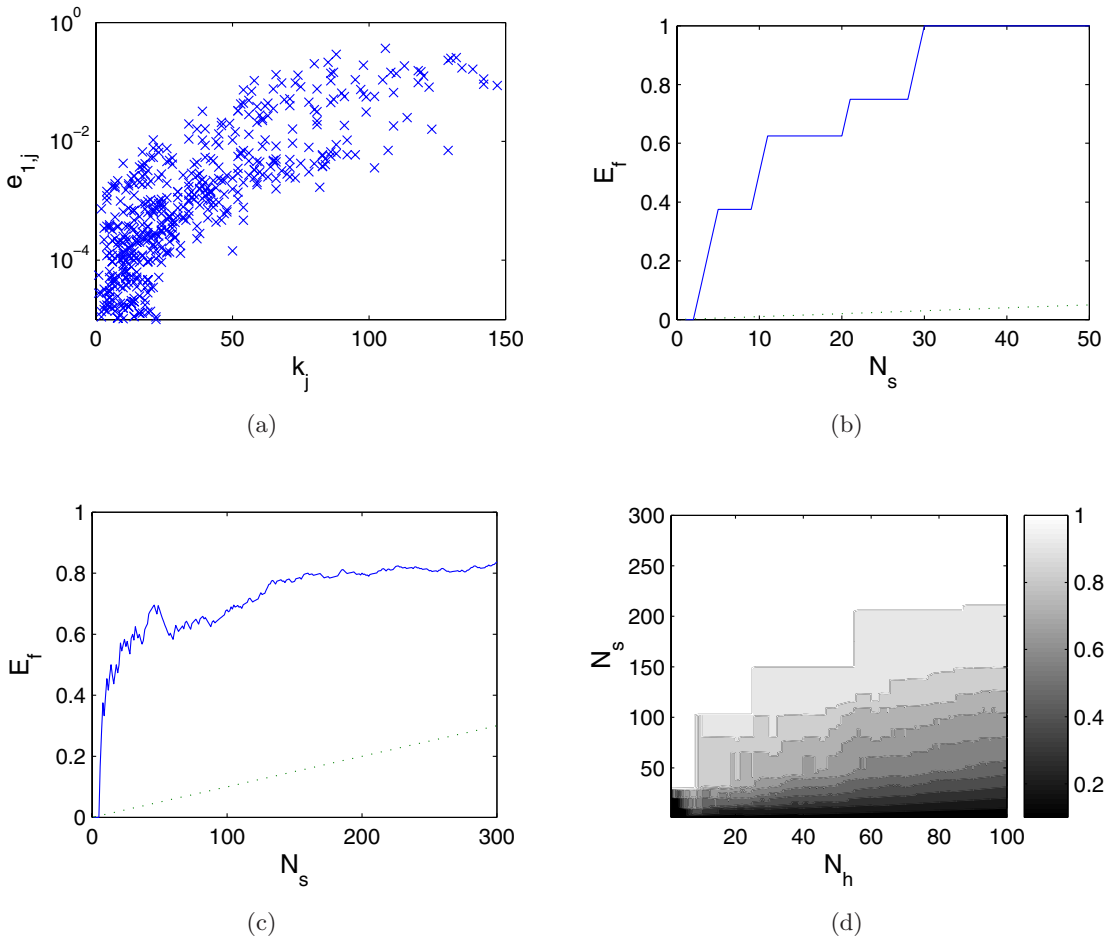


Fig. 7. For a subset of $n = 1000$ nodes of the actor network as in Sec. 3.5, (a) $e_{1,j}$ versus the node degree k_j without doing average; (b) in order to include eight hubs, the fraction E_f of the eight hubs included in the N_s subset of nodes detected as hubs by the method (solid line), the dotted line is the result if the nodes are chosen randomly; (c) E_f versus N_s (solid line) where the number of hubs N_h is the same as the number of nodes N_s in the detected subset, the dotted line is the result if the nodes are chosen randomly; (d) contour plot of E_f versus N_h and N_s.

not perfect, using $e_{1,j}$ to detect hubs will contain uncertainties. For example, if we want to determine N_h hub nodes, we may locate the first N_s nodes with the largest values of $e_{1,j}$. To characterize the efficiency of the method, we introduce E_f as the fraction of the N_h hub nodes found in the subset of N_s nodes from $e_{1,j}$. Figure 7(b) demonstrates an example of E_f versus N_s with $N_h = 8$. It can be seen that five nodes with the largest values of $e_{1,j}$ contain three of the eight hub nodes. In order to include all eight hubs, 30 nodes with the largest $e_{1,j}$ values need to be selected. Although the number of selected nodes by the method to include all target hubs is substantially larger than the number of actual hubs, it still outperforms the random selection method significantly (picking up N_s nodes randomly, as indicated by the dotted line). Figure 7(c) shows E_f versus N_s for $N_h = N_s$. Insofar as the number of hubs to be detected is more than a few, E_f can be as high as 0.6. That is, by the method from the same number of nodes with high $e_{1,j}$ values, one can locate approximately 60% of the hub nodes. A typical variation of E_f over N_h and N_s is shown in Fig. 7(d). Again, while the method cannot determine all hub nodes precisely, it can predict 60% or even 80% of the target hubs. A potential application is that when only a limited number of measured time series are available for a networked system, our method could provide a list of potential hubs with high probability.

The synchronization probability defined here is the same as the recurrence probability that can be computed, e.g. from cross recurrence plots [Marwan *et al.*, 2007], which detects complete

synchronization between two variables. The method will fail if the two variables have a phase shift, a lag, or any additional variation of the amplitude. To be more general and in order to detect lag synchronization, phase synchronization or even generalized synchronization, similar but more complicated approaches, e.g. joint recurrence plot, τ-recurrence rate etc., can be used [Marwan *et al.*, 2007]. Nevertheless, an appropriate pretreatment, such transferring the phase angle to oscillation variables and normalizing the amplitude in case of oscillations, etc., can broaden the applicability of the method. There are cases where other measures can work better than the synchronization probability, such as the cross correlation matrix, τ-recurrence etc. However, there are also cases when the synchronization probability measure works better. In particular, for the actor network, the synchronization probability is far better for the constructed time series than, say, the cross correlation. This is because the variables only take on two values, 1 or $-i$ for node i. The synchronization probability counts values 1, which reflects the probabilities for connections among nodes, and neglects $-i$, but the correlation focuses on $-i$ since a node is more likely to assume this value than 1. Therefore, although we start our procedure from dynamical networks, the method performs better for networks with variables taking on discrete values.

The key to our method is that the largest eigenvalue of the adjacency matrix should be well separated from all other eigenvalues. This property actually holds regardless of whether the network is sparse or dense, insofar as the size of the network is large. In particular, for a random network of N nodes and connecting probability p, the average number of links at a node is Np. The eigenvalues of the adjacency matrix are distributed according to the Wigner's semicircle law [Wigner, 1955]: the largest eigenvalue is of the order of Np, while all other eigenvalues are approximately bounded by $|\lambda_i| \leq 2\sqrt{Np(1-p)}$. We thus have $|\lambda_2| \sim 2\sqrt{Np(1-p)}$ and

$$\left|\frac{\lambda_1}{\lambda_2}\right| \sim 0.5\sqrt{\frac{Np}{(1-p)}}, \tag{3}$$

which becomes larger and larger as N and/or p are increased, making more accurate the linear relation between the components of the largest eigenvector and the node degree. Similar arguments hold for scale-free networks, based on the results of, e.g. [Farkas *et al.*, 2001]. All these have been verified by numerical computations. For example, we have examined random networks of 100 nodes with p ranging from 0.05 to 0.95 and scale-free networks of the same size with average degree ranging from 4 to 90. The results are shown in Fig. 8. Note that the networks considered contain both sparse and dense networks.

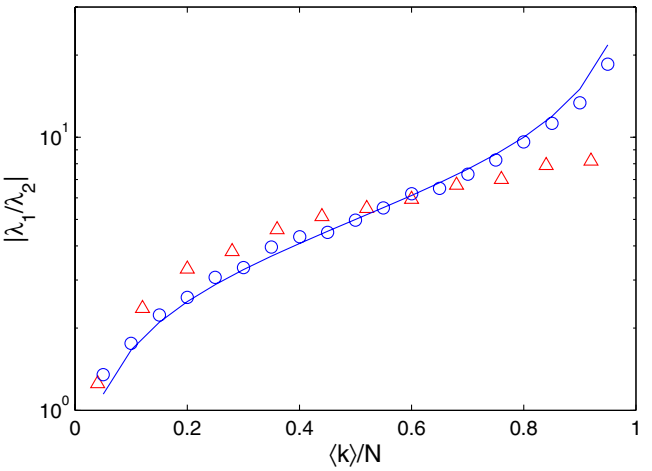


Fig. 8. The ratio $|\lambda_1/\lambda_2|$ versus normalized average degree $\langle k\rangle/N$ for random networks (circles) and scale-free networks [Barabási & Albert, 1999]. Network size $N = 100$. Symbols are simulation results from one realization of the network, curve is the theoretical result Eq. (3).

There exist various approaches to infer system structures from applying bivariate time series analysis techniques. Examples include applications to financial systems [Plerou *et al.*, 1999], neuronal functional networks [Jia *et al.*, 2004], brain functional networks [Eguiluz *et al.*, 2005], human EEG data [Seba, 2003], Internet traffic [Barthelemy *et al.*, 2002], and atmospheric motion [Santhanam & Patra, 2001], etc. The treatments mainly employ correlation matrix and uncover the structural information of the specific system by comparing the results with those from random matrices. For our purpose to detect hub nodes, we found that using the general synchronization-probability matrix yields better results and is practically easier to be implemented. Depending on the output signal, the synchronization probability may characterize complete synchronization of the dynamical systems, or it may describe the coincidence among the dynamical nodes (can be human beings) on some specific events. Therefore, the general synchronization-probability matrix is

more flexible and can be easily adapted for different tasks. And insofar as the recorded data bear interaction information, the constructed general synchronization-probability matrix can be used to locate the hub nodes.

The procedure we employed is similar to the eigenvector centrality method to assign node scores [Bonacich, 1972; Newman, 2004; Estrada & Rodríguez-Velázquez, 2005]. The difference is that for eigenvector centrality, the adjacency matrix is assumed to be known and the goal is to determine the relative importance of the nodes. While in our procedure, the goal is the same but only the measured time series are accessible and the matrix employed is the general synchronization-probability matrix which mimics the role of the adjacency matrix. The same holds for weighted networks [Newman, 2004].

There are benefits of using the components of the principal eigenvector as an indicator of node importance. First, as previously discussed and numerically demonstrated in the paper, the components have an approximate one-to-one relation with the node degrees thus they can be used to identify the hub nodes. Second, the components are in fact the principal components treating the general synchronization-probability matrix of size N as N time series (channels) [Pearson, 1901; Gorban *et al.*, 2007]. Thus these components give the relative strength of the channels, and effectively bypass the noises. The general synchronization-probability matrix, by construction, is contaminated with a huge amount of noise. Particularly, since the information interested is the degree of "synchronization," revealing the interaction thus possible existence of a link, the useful "signal" is small compared with the "noise" which is the variation of the time series observed from the nodes. Thus from the general synchronization-probability matrix it is difficult to infer the adjacency matrix directly. Applying a noise reduction procedure such as principal component analysis, the calculated principal components are able to indicate the hub nodes. Note that principal component analysis may suffer from nonunicity issues. For our application, since we use only the first component, e.g. the eigenvector of the largest eigenvalue of the constructed generalized synchronization matrix, insofar as the largest eigenvalue is not degenerated (as in most cases), the eigenvector is unique. Thus the unicity for our case is preserved.

5. Conclusion

We have developed a general and computationally efficient method to detect selected but important topological information of an unknown complex network based only on measured time series. A heuristic argument suggests that, when the network to be probed is large, the degrees of the nodes from which measurements are taken can be uniquely determined from the components of the principal eigenvectors of the underlying coupling matrix. By assuming that a set of time series can be measured, which bear information about the intrinsic node-to-node interactions in the network, and by defining a general synchronization-probability matrix that can be computed straightforwardly from the time series, we demonstrate the existence of a robust one-to-one correspondence between the principal-eigenvector components and the node degrees, regardless of the details of the node dynamics and the network topology. Our method can find broad applications. For instance, in order to find a small set of nodes of great importance from an unknown network, one can simply identify those components of the principal eigenvectors with relatively large values, and this can be done using measured time series only.

Acknowledgments

This work was supported by AFOSR under Grant No. FA9550-10-1-0083, and by NSF under Grants No. BECS-1023101 and No. CDI-1026710. L. Huang thanks L. Yang for valuable discussions and acknowledge the support from NSF of China through Grants No. 11005053 and No. 11135001.

References

Barabási, A.-L. & Albert, R. [1999] "Emergence of scaling in random networks," *Science* **286**, 509–512.

Barthelemy, M., Gondran, B. & Guichard, E. [2002] "Large scale cross-correlations in Internet traffic," *Phys. Rev. E* **66**, 056110.

Bonacich, P. [1972] "Factoring and weighting approaches to status scores and clique identification," *J. Math. Soc.* **2**, 113–120.

Eguiluz, V. M., Chialvo, D. R., Cecchi, G. A., Baliki, M. & Apkarian, A. V. [2005] "Scale-free brain functional networks," *Phys. Rev. Lett.* **94**, 018102.

Erdős, P. & Rényi, A. [1959] "On random graphs. I," *Publ. Math. Debrecen* **6**, 290–297.

Estrada, E. & Rodríguez-Velázquez, J. A. [2005] "Subgraph centrality in complex networks," *Phys. Rev. E* **71**, 056103.

Farkas, I. J., Derenyi, I., Barabási, A.-L. & Vicsek, T. [2001] "Spectra of 'real-world' graphs: Beyond the semicircle law," *Phys. Rev. E* **64**, 026704.

Gorban, A., Kegl, B., Wunsch, D. & Zinovyev, A. (eds.) [2007] *Principal Manifolds for Data Visualisation and Dimension Reduction* (Springer, NY).

Jia, L. C., Sano, M., Lai, P.-Y. & Chan, C. K. [2004] "Connectivities and synchronous firing in cortical neuronal networks," *Phys. Rev. Lett.* **93**, 088101.

Kuramoto, Y. [1984] *Chemical Oscillations, Waves and Turbulence* (Springer-Verlag, Berlin).

Marwan, N., Romano, M. C., Thiel, M. & Kurths, J. [2007] "Recurrence plots for the analysis of complex systems," *Phys. Rep.* **438**, 237.

Napoletani, D. & Sauer, T. D. [2008] "Reconstructing the topology of sparsely connected dynamical networks," *Phys. Rev. E* **77**, 026103.

Newman, M. E. J. [2004] "Analysis of weighted networks," *Phys. Rev. E* **70**, 056131.

Pearson, K. [1901] "On lines and planes of closest fit to systems of points in space," *Philos. Mag.* **2**, 559–572.

Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L. A. N. & Stanley, H. E. [1999] "Universal and nonuniversal properties of cross correlations in financial time series," *Phys. Rev. Lett.* **83**, 1471–1474.

Ren, J., Wang, W.-X., Li, B. & Lai, Y.-C. [2010] "Noise bridges dynamical correlation and topology in complex oscillator networks," *Phys. Rev. Lett.* **104**, 058701.

Santhanam, M. S. & Patra, P. K. [2001] "Statistics of atmospheric correlations," *Phys. Rev E* **64**, 016102.

Seba, P. [2003] "Random matrix analysis of human EEG data," *Phys. Rev. Lett.* **91**, 198104.

Timme, M. [2007] "Revealing network connectivity from response dynamics," *Phys. Rev. Lett.* **98**, 224101.

Wall, M. E., Rechtsteiner, A. & Rocha, L. M. [2003] "Singular value decomposition and principal component analysis," in *A Practical Approach to Microarray Data Analysis*, eds. Berrar, D. P., Dubitzky, W. & Granzow, M. (Kluwer, Norwell, MA), pp. 91–109.

Wang, W.-X., Chen, Q.-F., Huang, L., Lai, Y.-C. & Harrison, M. A. F. [2009] "Scaling of noisy fluctuations in complex networks and applications to network detection," *Phys. Rev. E* **80**, 016116.

Watts, D. J. & Strogatz, S. H. [1998] "Collective dynamics of 'small-world' networks," *Nature* **393**, 440–442.

Wigner, E. P. [1955] "Characteristic vectors of bordered matrices with infinite dimensions," *Ann. Math.* **62**, 548–564.

Yu, D., Righero, M. & Kocarev, L. [2006] "Estimating topology of networks," *Phys. Rev. Lett.* **97**, 188701.

Zhao, L., Lai, Y.-C. & Shih, C.-W. [2005] "Transition to intermittent chaotic synchronization," *Phys. Rev. E* **72**, 036212.