Inferring topologies of complex networks with hidden variables

Xiaoqun Wu[*]
School of Mathematics and Statistics, Wuhan University, Hubei 430072, China and School of Computing, Engineering and Mathematics, University of Western Sydney, Penrith NSW 2751, Australia

Weihan Wang
School of Mathematics and Statistics, Wuhan University, Hubei 430072, China and School of Mathematical Sciences, Fudan University, Shanghai 200433, China

Wei Xing Zheng[†]
School of Computing, Engineering and Mathematics, University of Western Sydney, Penrith NSW 2751, Australia

(Received 29 September 2011; revised manuscript received 11 September 2012; published 12 October 2012)

Network topology plays a crucial role in determining a network's intrinsic dynamics and function, thus understanding and modeling the topology of a complex network will lead to greater knowledge of its evolutionary mechanisms and to a better understanding of its behaviors. In the past few years, topology identification of complex networks has received increasing interest and wide attention. Many approaches have been developed for this purpose, including synchronization-based identification, information-theoretic methods, and intelligent optimization algorithms. However, inferring interaction patterns from observed dynamical time series is still challenging, especially in the absence of knowledge of nodal dynamics and in the presence of system noise. The purpose of this work is to present a simple and efficient approach to inferring the topologies of such complex networks. The proposed approach is called "piecewise partial Granger causality." It measures the cause-effect connections of nonlinear time series influenced by hidden variables. One commonly used testing network, two regular networks with a few additional links, and small-world networks are used to evaluate the performance and illustrate the influence of network parameters on the proposed approach. Application to experimental data further demonstrates the validity and robustness of our method.

DOI: 10.1103/PhysRevE.86.046106 PACS number(s): 89.75.Fb, 05.45.Tp, 05.45.Xt

I. INTRODUCTION

Networks are all around us, and the study of complex networks is pervasive in almost all scientific and technological fields. The past decade has seen many exciting developments and important achievements in the research of complex networks, which have provided tremendous insight into the topological properties of complex networks with interacting dynamical systems. It is believed that the interaction topology has important consequences on the robustness of the network function and the responses to external perturbations, such as random failures or targeted attacks. It is also well known that network topology plays a crucial role in determining the emergence of collective behaviors, such as synchronization, or in governing the main features of relevant processes that take place in complex networks, such as the spreading of epidemics, information, and rumors [1]. Therefore, to understand a complex network, particularly its evolutionary mechanisms and collective behaviors, it is necessary to first gain knowledge of the intrinsic network topology, which is usually unknown or uncertain in many practical situations.

The importance of inferring the interaction pattern among dynamical systems of a complex network, although very clear and already noted, was realized by researchers fairly slowly, partly because it involves the challenging inverse problem, especially with the lack of data and in the presence of noise. Nevertheless, there has been a steady growth of approaches regarding this topic in the past few years [2–13].

Methods have been proposed for topology identification of networks consisting of interacting deterministic systems based on synchronization [2–7], where some adaptive controllers are designed so that an auxiliary response network can achieve synchronization with the network under study (the drive network) and topological parameters can thus be estimated. However, in these methods, prior to successful topology identification, the nodal dynamics in a network has to be known and all the nodes have to be measurable and noise-free, which is not very practical.

For simultaneously observed time series, there are also some techniques such as measuring the cross correlation or partial correlation to recover their interaction patterns. Recently, a new method based on the noise-induced relationship between dynamical correlation and network topology was proposed to identify links among nodes [10]. However, these correlation-based techniques are incapable of distinguishing between direct and indirect interactions, so in many situations they do not provide very satisfactory results. Information-theoretic-based approaches have been widely employed for detecting links from purely observed data, such as the technique based on partial mutual information [8] and the technique of using conditional mutual sorting information [9]. Some straightforward methods based on the theory of recurrences have also been proposed [11,12]. Very recently, a permutation-based asymmetric association measure called

[*] xqwu@whu.edu.cn
[†] w.zheng@uws.edu.au

inner composition alignment was presented to infer directed networks from short time series [13], which is robust to noise but only performs well on sparse networks.

The present study focuses on an extension and application of a multivariate data-driven statistical technique known as Granger causality to infer the directed connections among multiple observed time series. The basic idea of Granger causality can be traced back to Wiener [14], who proposed a new way to measure the causal influence of one time series on another by conceiving the concept that if the prediction of one time series could be improved by incorporating the information of a second one, then the second series is said to have a causal influence on the first. Granger later formalized this concept in the context of linear regression models [15]. In the past few years, Granger causality has been widely employed in neuroscience and economics. Meanwhile, many extensions have been made, such as conditional Granger causality [16,17], blockwise Granger causality [18], and so on. In real situations, most, if not all, obtained time series are nonlinear. As Granger causality was originally formulated for linear models, it might not be sufficient for detecting effective connectivity for nonlinear time series. Thus, some extensions have been made to generalize the linear Granger causality to the nonlinear case [19–22].

However, the performance of all the above-mentioned approaches depends on being able to measure all relevant variables in a network, which is usually impractical in real-world situations, especially for large-scale networks in which it is impossible to observe or monitor all the nodes and there might also be undetectable disturbance. In 2008, Guo *et al.* introduced the so-called partial Granger causality to recover the interactions among elements in a network in the presence of exogenous inputs and latent variables [23]. They also extended the linear partial Granger causality to the nonlinear case using the radial basis function approach. In contrast to this mathematically complex technique, piecewise Granger causality, which was proposed by Wu *et al.* based on the idea of piecewise approximation [24], is much simpler and more straightforward for dealing with both linear and nonlinear time series.

In this paper, the piecewise approximation technique is employed in the partial Granger causality test to detect interactions among nonlinear time series in the presence of latent variables. The basic ideas of Granger causality, partial Granger causality, and our proposed approach are introduced in Sec. II. In Sec. III, the nonlinear Kuramoto model as well as the FitzHugh-Nagumo (FHN) model is used as nodal dynamics, and several types of network models are employed to test the performance of the proposed method. Additionally, the method is applied to gene expression data and gives convincing results. A brief conclusion is drawn in Sec. IV.

II. METHODS

In this section, the basic methods of Granger causality and partial Granger causality to evaluate causalities for linear systems are reviewed. Our extension of Granger's idea to nonlinear time series based on the piecewise linear approximation is then proposed.

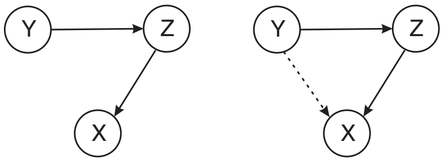

FIG. 1. Left: the true topology; right: a false link from Y to X is given.

A. Granger causality and partial Granger causality

The major approach to causality analysis is to examine if the prediction of one time series could be improved by incorporating information from the other, as proposed by Granger [15]. Specifically, given two time series X_t and Y_t which are jointly stationary, consider the autoregressive prediction of the current value of X_t based on its past measurements, described by

$$X_t = \sum_{i=1}^{\infty} a_{1i} X_{t-i} + \varepsilon_{1t}, \tag{1}$$

and the prediction using information of past measurements of both processes X_t and Y_t, given by

$$X_t = \sum_{i=1}^{\infty} a_{2i} X_{t-i} + \sum_{i=1}^{\infty} c_{2i} Y_{t-i} + \varepsilon_{2t}, \tag{2}$$

where ε_{it} $(i = 1,2)$ represents the prediction error. According to the definition of Granger causality [15], if $\mathrm{var}(\varepsilon_{2t}) < \mathrm{var}(\varepsilon_{1t})$, then Y_t influences X_t. The causal influence is quantified by $F_{Y\to X} = \ln(\frac{\mathrm{var}(\varepsilon_{1t})}{\mathrm{var}(\varepsilon_{2t})})$. Obviously, $F_{Y\to X} = 0$ indicates that there is no causal connection from Y_t to X_t, and $F_{Y\to X} > 0$ suggests that there is. The causal connection from X_t to Y_t can be defined similarly.

For a network having numerous nodes, various possibilities for causal connections among nodes arise. From the above pairwise Granger causality analysis for more than two time series, some false connections might be given due to the influence of observable or hidden variables in the network. For example, consider three variables X, Y, and Z, whose connection pattern is shown in the left panel of Fig. 1. However, the false link from Y to X is likely to be incorrectly inferred by a pairwise Granger causality test due to the mediation of Z. Another possible causal connection is shown in Fig. 2, where X and Y are simultaneously driven by Z. If the driving signal Z is powerful enough, then X, Y, and Z might evolve into generalized synchronization and it is very likely that one will get a false causal link between X and Y, such as the dashed line from Y to X. In 1984, Geweke [17] introduced conditional Granger causality, which has the ability to resolve whether the interaction between two time series is direct or mediated by

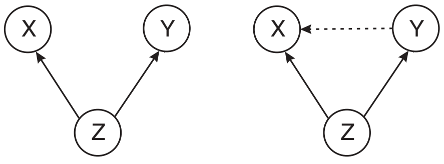

FIG. 2. Left: the true topology; right: a false link from Y to X is inferred.

another recorded time series and whether the causal influence is simply due to different time delays in their respective driving input. Critically, conditional Granger causality is effective only when all relevant variables in a network are observable. This is practically impossible, since both environmental inputs and unmeasured hidden variables can obscure accurate causal connections. In 2008, Guo *et al.* introduced partial Granger causality to detect causal connections, which is said to be capable of eliminating the influence of exogenous inputs and latent variables [23].

According to Guo *et al.* [23], partial Granger causality can be explained in the following way. Given two processes X_t and Z_t, the joint autoregressive representation for X_t and Z_t can be written as

$$\begin{aligned} X_t &= \sum_{i=1}^{\infty} a_{2i} X_{t-i} + \sum_{i=1}^{\infty} c_{2i} Z_{t-i} + u_{1t}, \\ Z_t &= \sum_{i=1}^{\infty} b_{1i} X_{t-i} + \sum_{i=1}^{\infty} d_{1i} Z_{t-i} + u_{2t}. \end{aligned} \tag{3}$$

The noise covariance matrix for the model can be represented by

$$\begin{bmatrix} \mathrm{var}(u_{1t}) & \mathrm{cov}(u_{1t},u_{2t}) \\ \mathrm{cov}(u_{1t},u_{2t}) & \mathrm{var}(u_{2t}) \end{bmatrix}, \tag{4}$$

with var and cov representing variance and covariance, respectively. Based on partial correlation in statistics, the value of $\mathrm{var}(u_{1t}) - \mathrm{cov}(u_{1t},u_{2t})\mathrm{var}(u_{2t})^{-1}\mathrm{cov}(u_{2t},u_{1t})$ measures the accuracy of the autoregressive prediction of X_t based on its previous values conditioned on Z_t by eliminating the influence of all other variables present in the network, such as common exogenous input and hidden variables.

Extending the concept further, the vector autoregressive representation for a system involving three time series X_t, Y_t, and Z_t can be written as follows:

$$X_t = \sum_{i=1}^{\infty} a_{2i} X_{t-i} + \sum_{i=1}^{\infty} b_{2i} Y_{t-i} + \sum_{i=1}^{\infty} c_{2i} Z_{t-i} + u_{3t}, \tag{5}$$

$$Y_t = \sum_{i=1}^{\infty} d_{2i} X_{t-i} + \sum_{i=1}^{\infty} e_{2i} Y_{t-i} + \sum_{i=1}^{\infty} f_{2i} Z_{t-i} + u_{4t}, \tag{6}$$

$$Z_t = \sum_{i=1}^{\infty} g_{2i} X_{t-i} + \sum_{i=1}^{\infty} h_{2i} Y_{t-i} + \sum_{i=1}^{\infty} k_{2i} Z_{t-i} + u_{5t}. \tag{7}$$

The noise covariance matrix for the above model can be represented by

$$\begin{bmatrix} \mathrm{var}(u_{3t}) & \mathrm{cov}(u_{3t},u_{4t}) & \mathrm{cov}(u_{3t},u_{5t}) \\ \mathrm{cov}(u_{4t},u_{3t}) & \mathrm{var}(u_{4t}) & \mathrm{cov}(u_{4t},u_{5t}) \\ \mathrm{cov}(u_{5t},u_{3t}) & \mathrm{cov}(u_{5t},u_{4t}) & \mathrm{var}(u_{5t}) \end{bmatrix}. \tag{8}$$

Similarly, the value of $\mathrm{var}(u_{3t}) - \mathrm{cov}(u_{3t},u_{5t})\mathrm{var}(u_{5t})^{-1}$ $\mathrm{cov}(u_{5t},u_{3t})$ represents the accuracy of predicting the present value of X_t based on the previous information of both X_t and Y_t conditioned on Z_t by eliminating the effect of other variables in the network. According to Guo *et al.* [23], the partial Granger causality from Y_t to X_t conditioned on Z_t by eliminating the effect of the common exogenous inputs and hidden variables present in the network can be expressed as

$$F_{Y\to X} = \ln\left(\frac{\mathrm{var}(u_{1t}) - \mathrm{cov}(u_{1t},u_{2t})\mathrm{var}(u_{2t})^{-1}\mathrm{cov}(u_{2t},u_{1t})}{\mathrm{var}(u_{3t}) - \mathrm{cov}(u_{3t},u_{5t})\mathrm{var}(u_{5t})^{-1}\mathrm{cov}(u_{5t},u_{3t})}\right). \tag{9}$$

$F_{Y\to X} = 0$ indicates that there is no direct casual influence from Y_t to X_t, and $F_{Y\to X} > 0$ means that there is.

B. Piecewise partial Granger causality

In this paper, to infer the topology of a complex network with stochastic perturbations from partially observed data, the famous Kuramoto model consisting of N nonlinearly coupled oscillators is mainly considered. In the presence of system noise, the model is described by the following governing equations:

$$\dot{\theta}_i = \omega_i + c\sum_{j=1}^{N} a_{ij}\sin(\theta_j - \theta_i) + \eta_i, \quad i = 1,2,\ldots,N, \tag{10}$$

where θ_i and ω_i are, respectively, the phase and the natural frequency of the ith oscillator, η_i is the independent Gaussian white noise with zero mean and intensity δ that represents the noisy background, and c represents the coupling strength. The topological information of the network is contained in the binary adjacency matrix $A = (a_{ij})_{N\times N}$, where $a_{ij} = 1$ if there is a directed link from the jth oscillator to the ith one (otherwise $a_{ij} = 0$) for $i \neq j$, and $a_{ii} = 0, i,j = 1,2,\ldots,N$.

As previously stated, time series obtained from real applications are mostly nonlinear. To deal with nonlinear time series in a statistically simple and operationally easy way, the idea of a piecewise linear approximation [24] is incorporated into the partial Granger causality. In practical situations, collecting numerous realizations of data will be time-consuming and sometimes impossible. Thus, in the procedure to be presented, for a given set of model parameters, only one realization of time series contaminated with Gaussian white noise from the Kuramoto model (10) is used. To have a confidence level for every link in the network, the 3σ rule is adopted, which states that for a normal distribution, nearly 99.73% of the data values lie within three standard deviations of the arithmetic mean μ; that is, $(\mu - 3\sigma, \mu + 3\sigma)$ is approximately a 99.73% confidence interval. To infer the underlying network topology from partially observed time series, the following procedure is

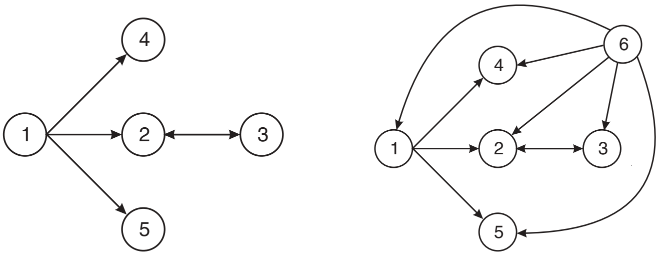

FIG. 3. Left: The standard testing network; right: the standard testing network with an immeasurable hidden node 6.

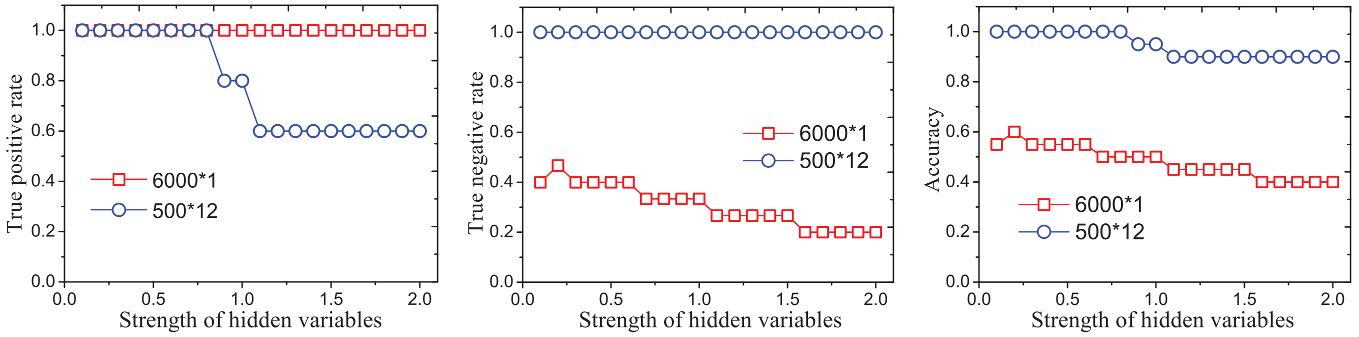

FIG. 4. (Color online) True positive rate (left), true negative rate (middle), and accuracy (right) vs the strength of hidden variables for the traditional case 6000 × 1 and the piecewise case 500 × 12, where $c = 10$ and $\delta = 0.4$.

proposed to recover their causal connections eliminating the influence of other variables.

Step (i). Partition each observed noisy time series of appropriate length into K consecutive parts of identical length N_0. Then, for the ith partition $(i = 1,2,\ldots,K)$ of two time series X_t and Y_t, add Gaussian white observation noise (noise intensity 0.05) and fit the linear regression models (1) and (2), and calculate $F_{Y\rightarrow X}$. Perform the noise addition and time series fitting for m runs. Then, calculate the 3σ confidence interval for $F_{Y\rightarrow X}$, whose lower bound is denoted as α_i $(i = 1,2,\ldots,K)$. According to the 3σ rule, if α_i $(i = 1,2,\ldots,K)$ is greater than zero, one infers a causal connection $Y \rightarrow X$ from the ith partition of data. Calculate the piecewise partial Granger causality index (PPGCI) defined as

$$\mathrm{PPGCI} = \frac{1}{K}\sum_{i=1}^{K}\alpha_i. \qquad (11)$$

If PPGCI > 0, the directed connection $Y \rightarrow X$ is accepted to appear in the network. Otherwise, the link is supposed to be absent.

Step (ii). Perform step (i) for all the pairwise time series, and hence obtain an initial topological structure of the network.

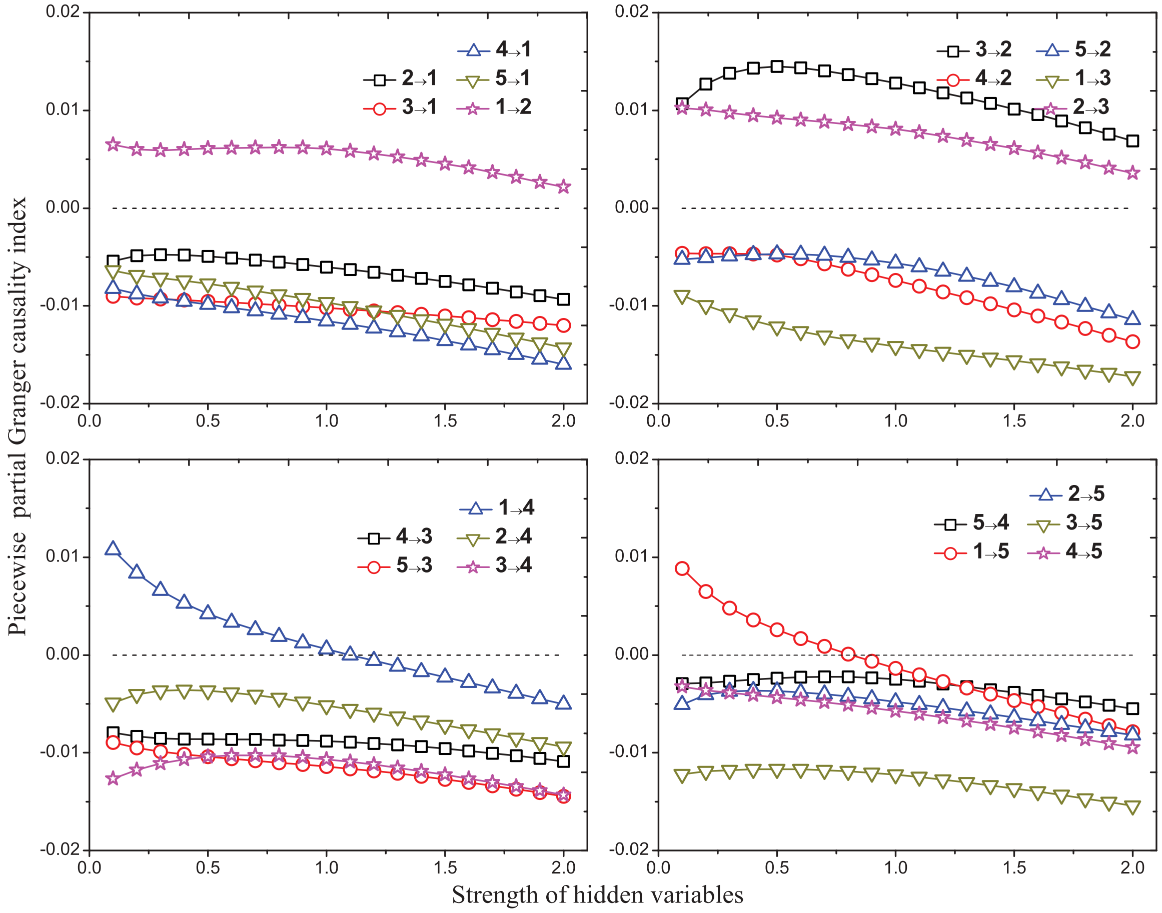

FIG. 5. (Color online) PPGCI vs the strength ε of hidden variables for the 500 × 12 partition case, where $c = 10$ and $\delta = 0.4$.

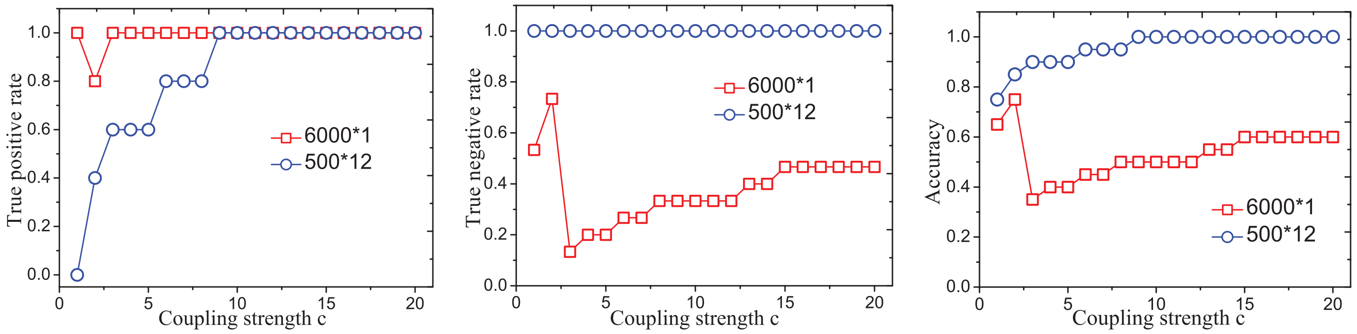

FIG. 6. (Color online) True positive rate (left), true negative rate (middle), and accuracy (right) vs the coupling strength for the traditional case 6000 × 1 and the piecewise case 500 × 12, where $\varepsilon = 0.7$ and $\delta = 0.4$.

Note that it is very likely that there are incorrectly inferred connections.

Step (iii). Check the initially obtained network structure. If there is any subgraph as shown in the right panel of Figs. 1 or 2, then take Z_t as the condition and apply the linear regression models (3), (5), and (7) to the K respective parts of related time series and similarly calculate the PPGCI as described in step (i). Those false causal connections, which are indirect or mediated by other observed variables Z_t, are thus eliminated. According to the definition of partial Granger causality, this can also eliminate the mediation of latent variables.

Step (iv). For the resulting network structure, consider those existent links which are not previously considered in step (iii), such as a link between some two processes X_t and Y_t. Take a third process other than X_t or Y_t in the network, say Z_t, as the condition, and calculate the PPGCI as described in step (i). If PPGCI $\leqslant 0$, then stop and set the considered link to be absent; otherwise, take another process in the network as the condition and calculate the PPGCI again. If the values of the PPGCI conditioned on all the other processes are larger than 0, then take the minimum of them as the causality index for the considered link. Keep the resulting network structure, and stop.

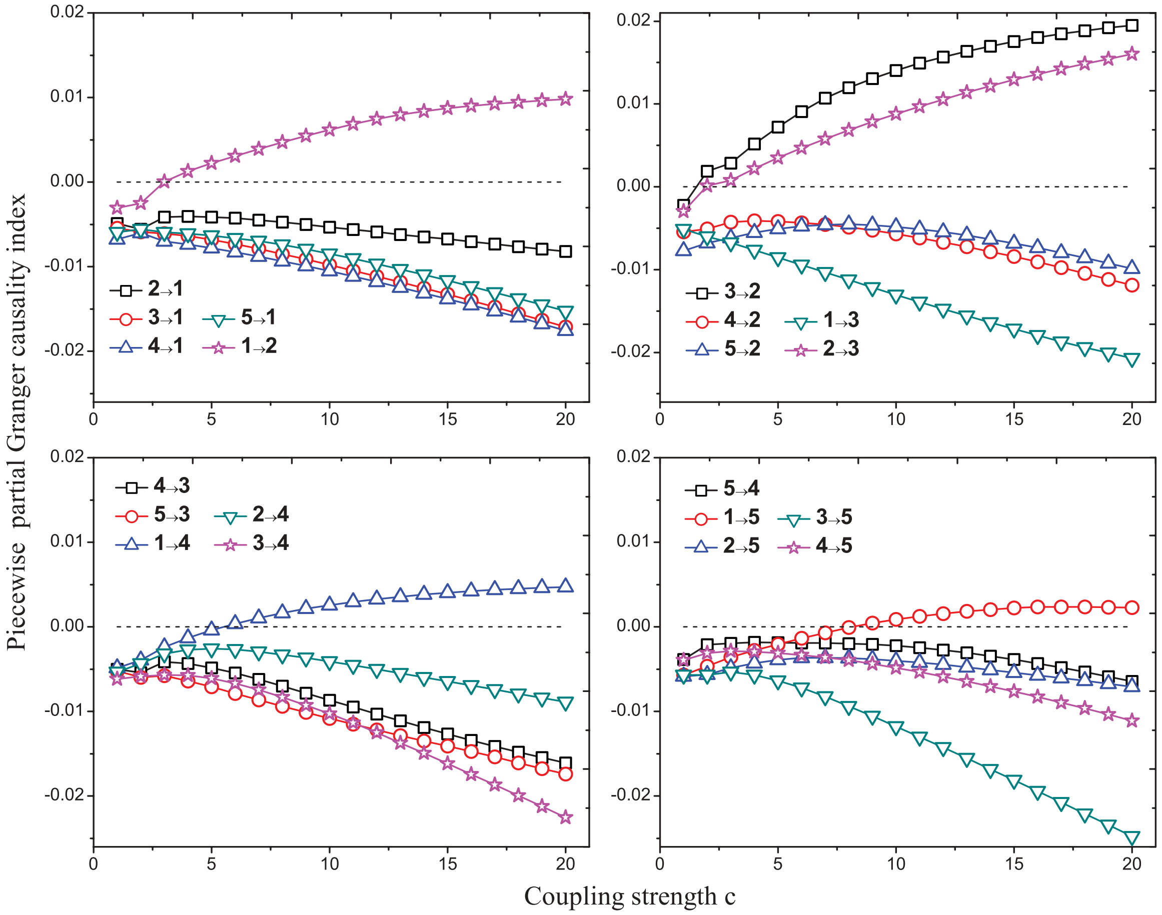

FIG. 7. (Color online) PPGCI vs the coupling strength c for the 500 × 12 partition case, where $\varepsilon = 0.7$ and $\delta = 0.4$.

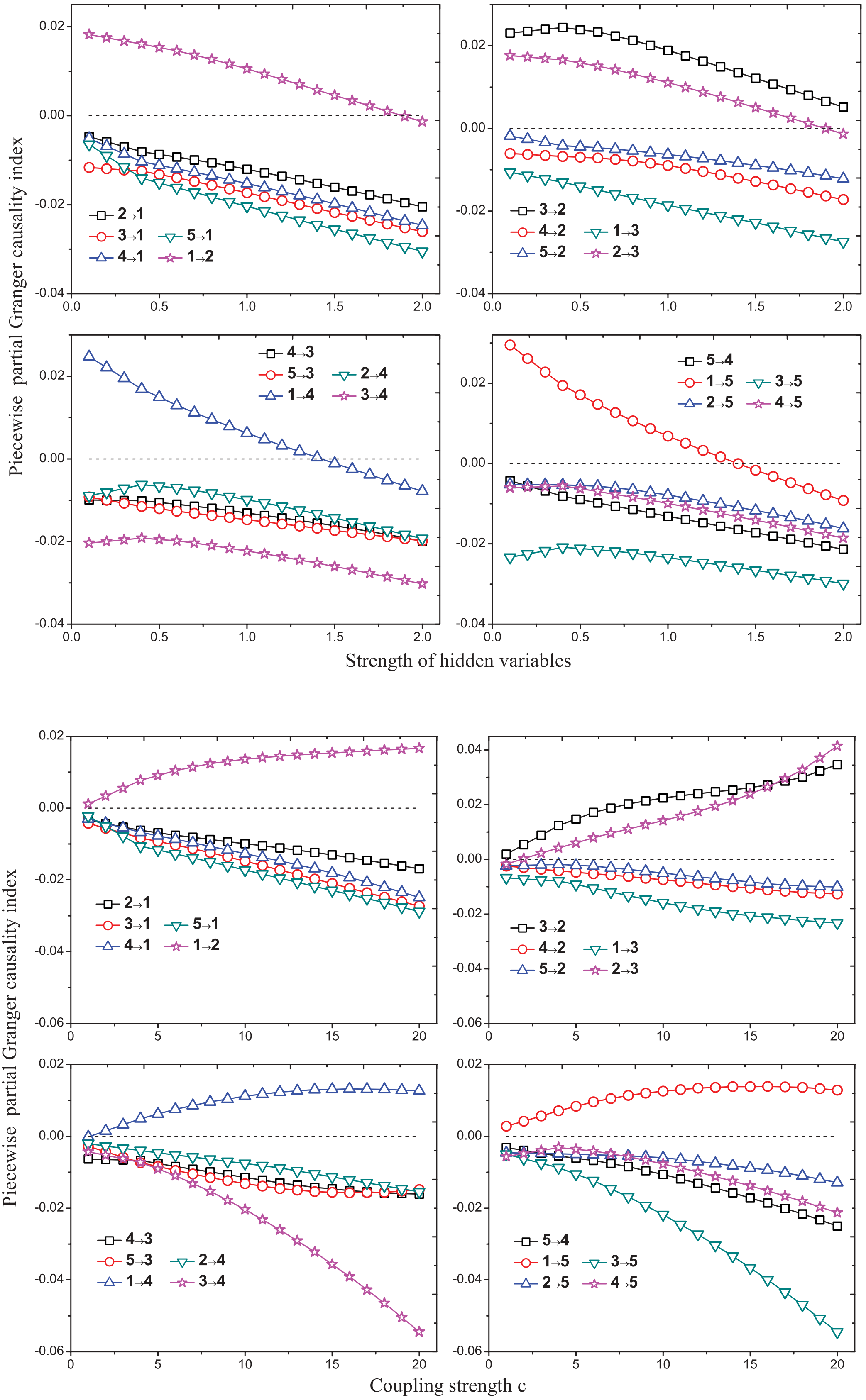

FIG. 8. (Color online) Topology identification for the network composed of FHN oscillators. Top: PPGCI vs the strength ε of hidden variables for the 500×12 partition case, where $c = 10$ and $\delta = 0.4$; bottom: PPGCI vs the coupling strength c for the 500×12 partition case, where $\varepsilon = 0.7$ and $\delta = 0.4$.

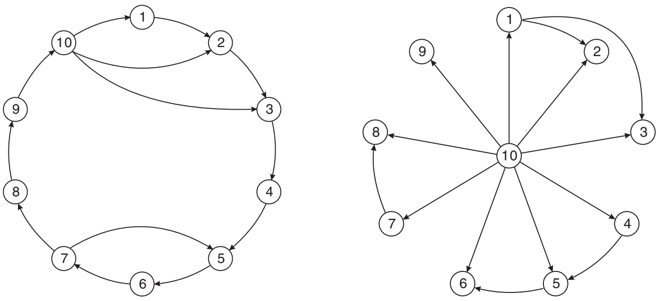

FIG. 9. Left: a directed ring network model with three additional links; right: a directed star network model with five additional links.

Remark. Basically one can skip step (iii) and calculate the PPGCI of a link conditioned on all the other considered processes one by one, as described in step (iv). However, that may be computationally expensive and requires a huge amount of memory and time for computation. By finding some specific conditional process, step (iii) can greatly reduce the testing expense. It is also noted that if the initial network inferred from step (ii) is densely connected, then all the existent links might probably be considered in step (iii), which are thus verified by eliminating the influence of both observable and latent variables using the partial Granger causality. In this way, step (iv) may be skipped.

In general, the partition length N_0 is crucial for detecting the topology of a network. As stated in Ref. [24], as long as the data length in each partition is large enough for good statistics, denser partitions always lead to better results, which is in accordance with the basic idea of piecewise linear approximation.

III. NUMERICAL SIMULATIONS

In simulation study, the Euler-Maruyama method is employed to generate time series from the above stochastic differential equations (10) with an equal time step 0.01, where ω_i is assumed to be 1 for $i = 1,2,\ldots,N$. The time series generated are of length 6000. The number of runs used throughout the paper is $m = 30$. Besides the piecewise partial Granger causality index as defined in Eq. (11), the true positive rate (the fraction of correctly inferred links out of all existent links), the true negative rate (the fraction of correctly inferred nonexistent links out of all nonexistent links), and the accuracy (the fraction of correctly inferred links out of all links) are also computed as performance measures to evaluate the identification ability of the proposed method.

A. A standard testing network

In this subsection, to compare the difference between the partial Granger causality and the piecewise partial Granger causality introduced here, an extensively used standard testing network for Granger causality is considered, as shown in the left panel of Fig. 3. This model is modified by adding a common exogenous input to each node, as shown by node 6 in the right panel of Fig. 3. Here, node 6 is assumed to be an immeasurable hidden variable. To analyze how the hidden variable impacts the piecewise partial Granger causality, the influence strength of node 6 on all the observable nodes is supposed to be ε, which is termed the "strength of hidden variables." The greater the value of ε, the stronger the influence of the hidden variable on the observable nodes. The adjacency matrix of the underlying network can be modified as

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & \varepsilon \\ 1 & 0 & 1 & 0 & 0 & \varepsilon \\ 0 & 1 & 0 & 0 & 0 & \varepsilon \\ 1 & 0 & 0 & 0 & 0 & \varepsilon \\ 1 & 0 & 0 & 0 & 0 & \varepsilon \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad (12)$$

which is employed in Eq. (10) to generate the time series, but only the data set of the first five time series is used to recover the interaction patterns among them.

To compare the piecewise partial Granger causality with the traditional one, we apply the approach to time series of length 6000 with two cases: One is the traditional partial Granger causality without partition (6000 × 1) and the other is the piecewise case with partition 500 × 12. Figure 4 displays the true positive rate, the true negative rate, and the accuracy varying with the strength of hidden variables for the two cases, where the network coupling strength $c = 10$ and the system noise intensity $\delta = 0.4$. From the left and the middle panels of Fig. 4, it can be seen that for the traditional case 6000 × 1, the true positive rate stays at 1 while the true negative rate

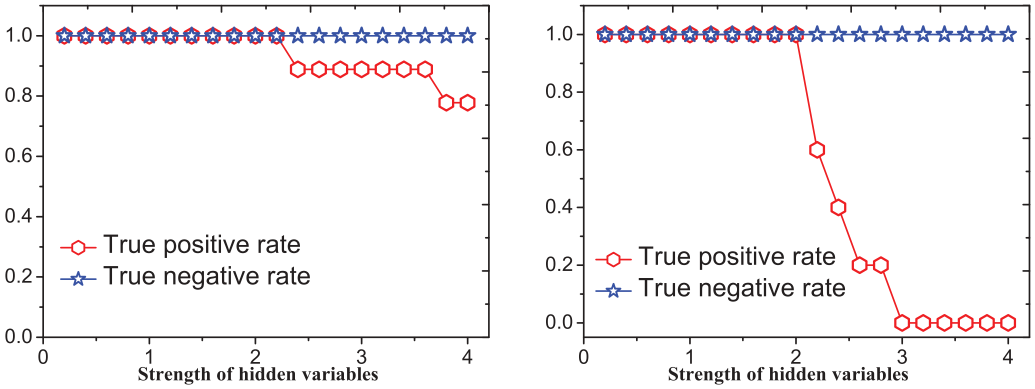

FIG. 10. (Color online) True positive rate and true negative rate vs the strength of hidden variables for the directed ring network (left) and the directed star network (right), where $c = 10$, $\delta = 0.4$.

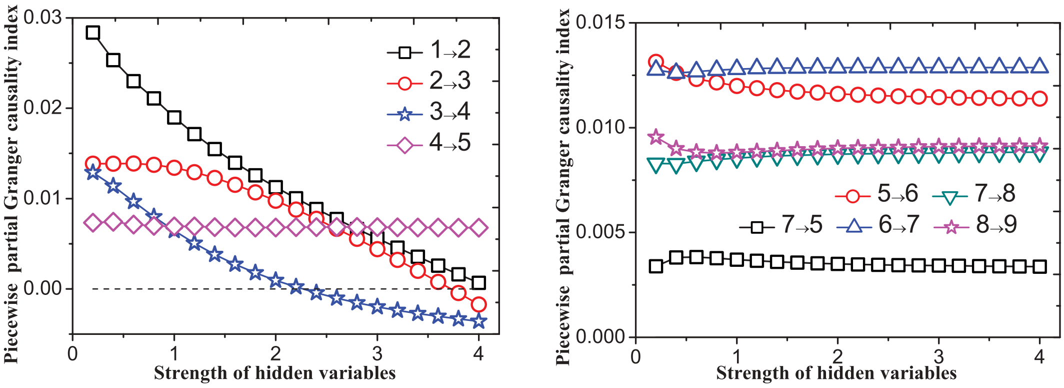

FIG. 11. (Color online) PPGCI vs the strength of hidden variables for the existent links in the directed ring network, where $c = 10$, $\delta = 0.4$.

stays below 0.5 and keeps decreasing with increasing ε, which means many nonexistent links are incorrectly predicted to be existent in the network. For the piecewise case 500×12, the true negative rate stabilizes at 1 while the true positive rate stays at 1 when the strength ε of hidden variables is smaller than 0.9, then it experiences a slight drop to 0.8 and decreases further to 0.6 as ε increases, which means one or two existent links are missed in the topology detection. The right panel of Fig. 4 displays the identification accuracy, from which it is observed that the accuracy decreases as the influence of hidden variables gets stronger for both cases. That is to say, the performance of partial Granger causality is actually affected by the influence of latent variables. Nevertheless, the piecewise partial Granger causality significantly outperforms the traditional one by yielding a much higher accuracy, which is maintained at above 0.9.

To take a detailed look at the identification performance of the piecewise partial Granger causality test varying with the influence of latent variables, Fig. 5 presents the piecewise causality index for all the links, both existent and nonexistent. It can be seen from the figure that when the strength ε is small, all the links are correctly inferred. However, when ε reaches 0.9, the existent link $1 \rightarrow 5$ is missed; when ε surpasses 1.1, another link $1 \rightarrow 4$ is further missed by the causality test. From the declining trend of all the PPGCI curves, it can be observed that more and more existent links will be missed with the increasingly stronger influence of hidden variables. That is to say, the piecewise partial Granger causality will no longer succeed in recovering the interactions among simultaneously obtained time series if the hidden variables get too powerful.

To see the influence of the network coupling strength c on the identification, Fig. 6 shows the true positive rate, the true negative rate, and the accuracy varying with respect to the network coupling strength c for the traditional and piecewise cases, where the strength of latent variables $\varepsilon = 0.7$ and the system noise intensity $\delta = 0.4$. For the traditional case 6000×1, the true positive rate basically stabilizes at 1 while the true negative rate, which is mostly below 0.5, basically increases with c. For the piecewise case 500×12, the true negative rate stays at 1 while the true positive rate increases rapidly from 0 at $c = 1$ to 1 with the coupling strength c reaching 9. The right panel of Fig. 6 displays the accuracy of identification, from which it can be seen that the accuracy increases as the coupling strength gets stronger. The results displayed for both the traditional and the proposed piecewise cases show that strong coupling favors correct identification, which is easy to understand. Moreover, it is obvious that the piecewise partial Granger causality is still superior to the traditional one for various values of coupling strength.

Figure 7 displays the piecewise partial Granger causality index for all the links with respect to varying coupling strength c. When c is as small as 1, all the existent links are missed. As c increases, more and more existent links are recovered. When c reaches 9, the original pattern of connection is correctly inferred. Furthermore, the PPGCI for existent links increases with c and that for nonexistent links decreases. However, the rising rate of the PPGCI for existent links slows down as c increases. This can be explained from the theory of synchronization-based method [25–27], which states that topology identification is likely to fail if some nodes of a network get into generalized synchronization. Analogously, with a strong coupling strength, a certain number of nodes might get into approximate generalized synchronization, which makes it more difficult to differentiate one time series from another. Generally speaking, strong couplings contribute to the synchronization, while proper system noise helps desynchronize the network and the nodes can still be distinguished from each other. In this sense, the presence of minor noise facilitates topology inference.

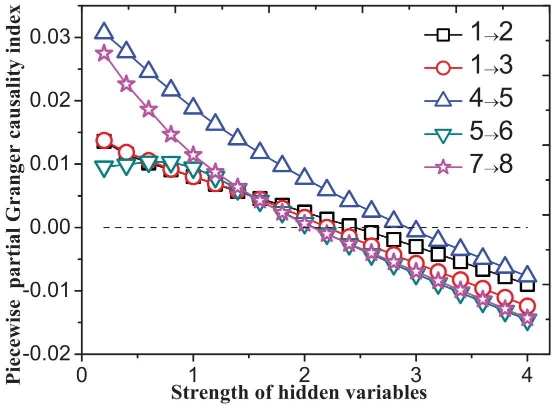

FIG. 12. (Color online) PPGCI vs the strength of hidden variables for the existent links in the directed star network, where $c = 10$, $\delta = 0.4$.

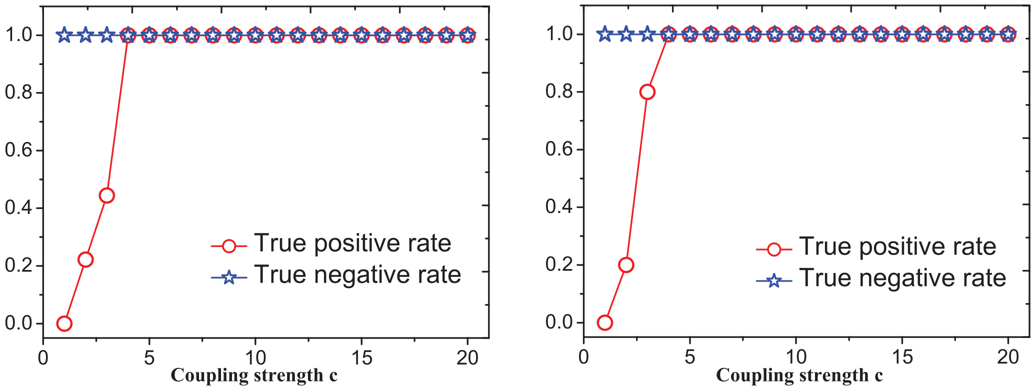

FIG. 13. (Color online) True positive rate and true negative rate vs the coupling strength for the ring network (left) and the star network (right), where $\varepsilon = 1$, $\delta = 0.4$.

To show the validity of the proposed technique, we further test it with a network of coupled neural systems. The FitzHugh-Nagumo (FHN) model is considered [28,29], with the stochastically perturbed network being described by

$$\dot{V}_i = V_i - \frac{1}{3}V_i^3 - W_i + I_{\text{ex}} + c\sum_{j=1}^{N} a_{ij}(V_j - V_i) + \eta_i, \qquad (13)$$
$$\dot{W}_i = 0.08(V_i + 0.7 - 0.8W_i),$$

where V_i is the membrane potential, W_i is the recovery variable, and I_{ex} is the external stimulus current, which is supposed to be $\cos\frac{t}{50}$ here. The standard testing network shown in Fig. 3 is employed as the underlying topology and node 6 is still assumed to be the hidden variable with strength ε. The identification results varying with ε and the coupling strength c are shown in the top and bottom panels of Fig. 8, respectively. Similar observations to those for the Kuramoto model can be obtained.

B. Two regular networks

To illustrate the ability of the proposed method, two regular network models—a directed ring network and a directed star network—are further considered, as shown in Fig. 9. Node 10 is assumed to be the hidden variable in both models, and only the evolutionary time series of the other nine nodes are collected for topology identification. To show the impact of hidden variables on those nodes with different connection patterns, a few additional links are added. Similarly, denote by ε the influence strength of node 10 on its outgoing nodes. That is, only those nonzero elements for existent links in the 10th column of the 10×10 adjacency matrix are modified to ε.

Figure 10 presents the true positive rate and the true negative rate with respect to ε for the directed ring network (left panel) and the directed star network (right panel), where the network coupling strength $c = 10$ and noise intensity $\delta = 0.4$. It is clearly observed that the underlying topologies are correctly recovered when ε is relatively small for both models. Moreover, the true negative rate is always maintained at 1. Specifically, for the ring network, when ε arrives at 2.4, the true positive rate drops to 0.89 and decreases further to 0.78 when ε gets to 3.8. For the star network, it can be seen that the true positive rate dramatically drops from 1 to 0 when ε increases gradually from 2 to 3. The different decline rates of the true positive rate in the two networks should be due to the fact that for the ring network, the hidden node only directly influences nodes 1, 2, and 3, whereas for the star network, the hidden node directly impacts all the observable nodes. It is thus verified again that the partial Granger causality works well only if the latent variables are not too influential on other concerned nodes.

Figure 11 displays the PPGCI of the existent links for the directed ring network with three additional links. The PPGCI for the existent link 3 → 4 declines below 0 (the dashed line) when ε surpasses 2.4, meaning that this link is currently missed by the causality test. Another link 2 → 3 is further missed when ε reaches 3.8. The two critical values of ε correspond to

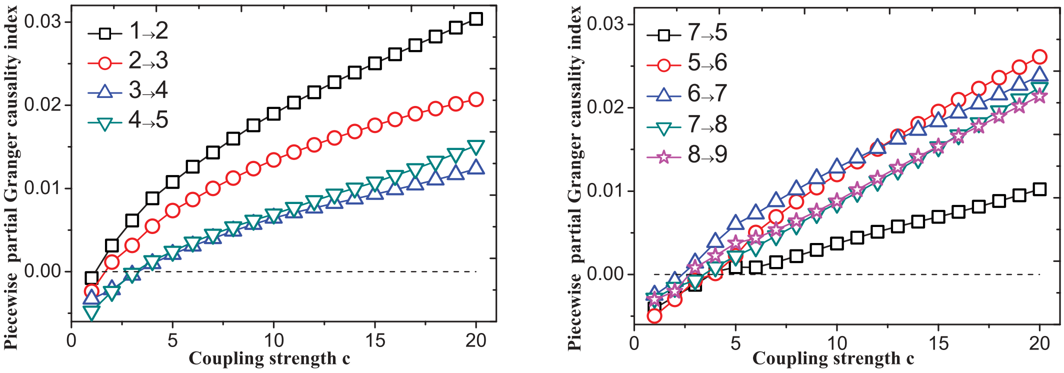

FIG. 14. (Color online) PPGCI vs the coupling strength for the existent links in the directed ring network, where $\varepsilon = 1$, $\delta = 0.4$.

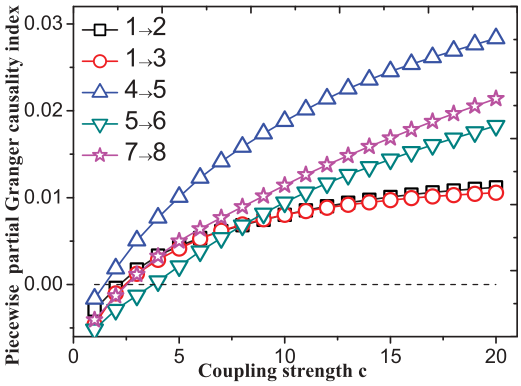

FIG. 15. (Color online) PPGCI vs the coupling strength for the existent links in the directed star network, where $\varepsilon = 1$, $\delta = 0.4$.

the two drops in the true positive rate curve for the ring network in Fig. 10. Moreover, the PPGCI curves for $1 \to 2$, $2 \to 3$, and $3 \to 4$ are obviously decreasing with increasing ε, while the others are basically leveled, which should be due to the fact that nodes 1, 2, and 3 are directly influenced by the hidden node 10. That is to say, the strength of hidden variables mainly influences the interaction identification of those observed time series which are directly impacted by the hidden nodes.

Figure 12 displays the PPGCI of all the existent links for the directed star network containing five additional links. Apparently, since all the nodes are immediately impacted by the latent hub node, the PPGCI for all the existent links gradually declines below 0 (the dashed line) with increasing ε. When ε reaches 3, all the existent links are missed by the causality prediction, corresponding to the zero value of the true positive rate in the right panel of Fig. 10. In a word, for the star network where the hub node is unobservable, the connection pattern of the leaf nodes can be correctly recovered unless the hidden node is too influential.

Figure 13 presents the true positive rate and the true negative rate with respect to varying c for the directed ring network (left panel) and the directed star network (right panel) with additional links, where the strength of hidden variables $\varepsilon = 1$ and the noise intensity $\delta = 0.4$. From both panels, it is clearly seen that the topologies are correctly identified when the network coupling strength reaches 4.

Figures 14 and 15 display the PPGCI of the existent links for the directed ring network and the star network, respectively. For both networks, all the existent links are missed when $c = 1$, which is due to the fact that weak coupling makes the underlying topologies obscured. When c is increasing, the PPGCI for all existent links rises too. As c reaches 4, all the links are correctly recovered.

C. Small-world networks

The commonly used testing model and two regular networks with a couple of additional links demonstrate the effectiveness of the proposed piecewise partial Granger causality and its applicability range with respect to the strength of hidden variables and the network coupling strength. In this subsection, the proposed method is tested on small-world networks, which can well model many real-world systems, such as social influence networks, road maps, food chains, gene networks, and so on.

The Newman-Watts (NW) algorithm is employed here to construct small-world networks [30,31]. First, consider a network with 50 nodes, with each node connecting to its nearest neighbor, that is, an undirected ring network. Then, directed "shortcuts" are added with probability p to the ring network between randomly chosen pairs of nodes (except those already connected pairs). Obviously, the larger the probability p of adding links, the denser the network. The generated NW network is adopted in the Kuramoto model (10) to generate time series. Then the first five time series are adopted to infer the connections among them, assuming that the other 90% nodes are latent or unconcerned. The true positive rate and the true negative rate are calculated. Note that the directed links are added to the ring network randomly, thus the resulting adjacency matrixes can be various. To be more representative, the above procedure is repeated 30 times. Then the average values of the true positive rate and the true negative rate are recorded.

The left panel of Fig. 16 displays the true positive rate and the true negative rate varying along with different probabilities of adding links, where the network coupling strength $c = 20$ and the system noise intensity $\delta = 4$. It can be seen from the panel that the true negative rate experiences a sharp drop when p varies from 0 to 0.01 and then remains in a very slight decline, while the true positive rate keeps decreasing when p

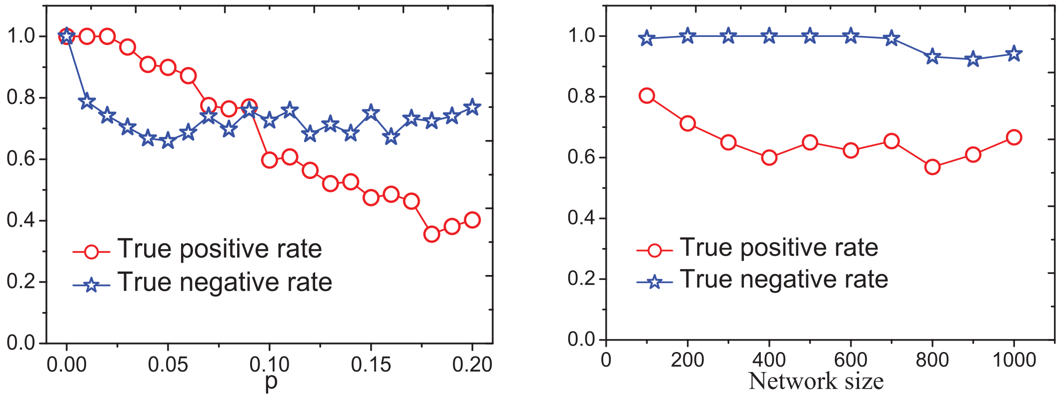

FIG. 16. (Color online) Left: true positive rate and true negative rate vs the probability of adding edges, where $c = 20$ and $\delta = 4$; right: true positive rate and true negative rate vs the network size, where $c = 20$, $\delta = 4$, and $p = 0.01$.

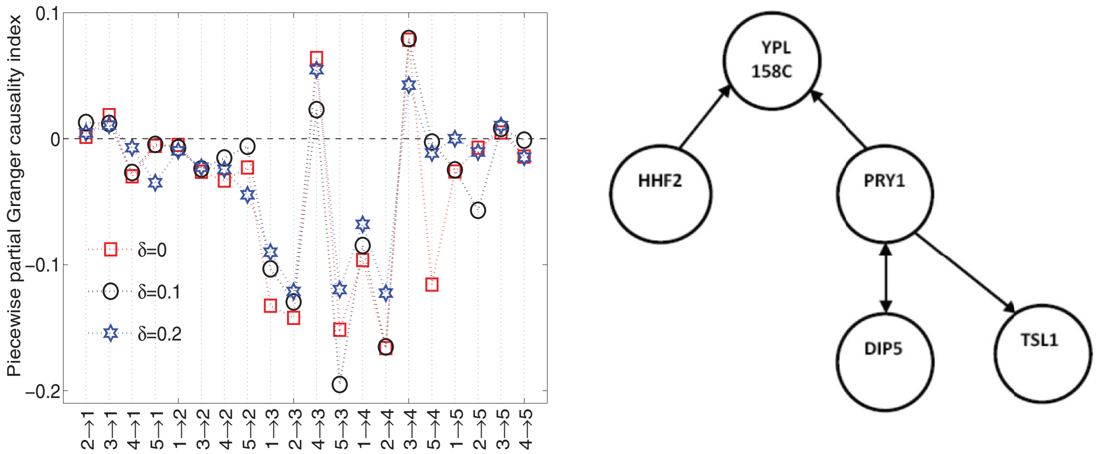

FIG. 17. (Color online) Left: PPGCI for the 20 causal connections with different levels of observation noise; right: the inferred network topology of the five genes.

increases gradually from 0 to 0.2 with a step size 0.01. That is to say, more existent links are missed for denser networks. This can be explained by the fact that with a larger p, more links are connected to the concerned nodes, which means that there are more hidden variables.

In the right panel of Fig. 16, the influence of the network size on the proposed identification method is displayed. Specifically, the size of the generated NW small-world network increases from 100 to 1000 with a step size 100, where $c = 20$, $\delta = 4$, and the probability of adding links $p = 0.01$. Similarly, five time series are adopted to infer the connections among them, assuming that all the others are latent or unconcerned. It can be seen from the panel that both the true positive rate and the true negative rate decline slightly with the increase of the network size. This can also be explained by the fact that with a larger network size, more hidden variables are connected to the concerned five nodes. In a word, both panels in Fig. 16 illustrate that the partial Granger causality works well when the influence of hidden nodes is not so strong.

D. Application to experimental data

As a practical application, our method is tested using yeast cell cycle gene expression data downloaded from the Yeast Cell Cycle project at Stanford University (http://genome-www.stanford.edu/cellcycle/data/rawdata/). These studies profiled expression changes in 6178 genes at about 80 time points under four different conditions. Many genes have missing data points. In Ref. [32], Wang *et al.* proposed a novel method to combine multiple time-course microarray datasets from different conditions for inferring gene regulatory networks. Therein, they applied their method to 140 differentially expressed genes and generated consistent subnetworks with 64 links, 431 links, etc. depending on the scalar parameter used to control the sparsity or consistency of the subnetwork. The 64-link inferred network was presented in their paper. In our test, we selected five genes (YPL158C, HHF2, PRY1, DIP5, and TSL1) for illustration. Clearly, due to unrecorded inputs and the fact that we only used five time series, the genes are influenced both by substantial exogenous inputs and by a large set of latent variables. Since many genes have missing data points, we discarded those time points with missing data and chose 60 data points for each gene. The time series were partitioned into two consecutive parts of identical length. In the left panel of Fig. 17, the red squares represent the PPGCI for the 20 possible causal connections for the five genes. Here, nodes 1, 2, 3, 4, and 5 represent YPL158C, HHF2, PRY1, DIP5, and TSL1, respectively. The inferred topology of the five genes is shown in the right panel, which is the same as that inferred in Ref. [32] except for an additional link from gene DIP5 to gene PRY1. This is reasonable, since the 64-link network therein was inferred using a certain scalar parameter to control the sparsity, while the underlying network can have more links. To see the robustness of our method, two different levels of observation noise are added to the gene expression data, and similar calculations are performed. The causality indexes are displayed in the left panel, with black circles and blue stars representing that for $\delta = 0.1$ and 0.2, respectively. One can see that the method is robust to these disruptions.

IV. CONCLUSIONS

In this paper, a simple and feasible approach called "piecewise partial Granger causality" has been proposed that is designed to recover the interactions among simultaneously obtained nonlinear noisy time series from complex dynamical networks with hidden nodes. Comparison with the previous partial Granger causality has been done using a standard testing network to show the superiority of the proposed approach. The validity of the approach has been demonstrated further on two regular networks containing a few additional links, both with one hidden node. The latent variables' strength and the network coupling strength have been studied as two key factors influencing the effectiveness of this piecewise partial Granger causality. Additionally, NW small-world network models have been tested to further verify the ability of the method and to investigate the impact of hidden nodes. Finally, an application to experimental data further demonstrates the validity and robustness of the method. It has been illustrated that piecewise partial Granger causality is effective in detecting the connections among observed noisy time series in the presence of hidden variables on the conditions that the hidden variables do not have much of an influence on the observable ones, and the coupling strength among the observable variables is strong enough.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions, which substantially improve the paper, and Pei Wang for downloading the gene expression data. This work is supported in part by the Chinese National Natural Foundation (Grant No. 61174028), the Doctoral Fund of the Ministry of Education of China (Grant No. 20080486172), and a research grant from the Australian Research Council.

[1] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, Phys. Rep. **424**, 175 (2006).

[2] D. Yu, M. Righero, and L. Kocarev, Phys. Rev. Lett. **97**, 188701 (2006).

[3] J. Zhou and J. Lu, Physica A **386**, 481 (2007).

[4] X. Wu, Physica A **387**, 997 (2008).

[5] H. Liu, J. Lu, J. Lü, and D. J. Hill, Automatica **45**, 1799 (2009).

[6] J. Zhou, W. Yu, X. Li, M. Small, and J. Lu, IEEE Trans. Neural Networks **20**, 1679 (2009).

[7] J. Zhao, L. Qin, J. Lu, and Z. P. Jiang, Chaos **20**, 023119 (2010).

[8] S. Frenzel and B. Pompe, Phys. Rev. Lett. **99**, 204101 (2007).

[9] B. Pompe and J. Runge, Phys. Rev. E **83**, 051122 (2011).

[10] J. Ren, W. X. Wang, B. Li, and Y. C. Lai, Phys. Rev. Lett. **104**, 058701 (2010).

[11] M. C. Romano, M. Thiel, J. Kurths, and C. Grebogi, Phys. Rev. E **76**, 036211 (2007).

[12] J. Nawrath, M. C. Romano, M. Thiel, I. Z. Kiss, M. Wickramasinghe, J. Timmer, J. Kurths, and B. Schelter, Phys. Rev. Lett. **104**, 038701 (2010).

[13] S. Hempel, A. Koseska, J. Kurths, and Z. Nikoloski, Phys. Rev. Lett. **107**, 054101 (2011).

[14] N. Wiener, *The Theory of Prediction*, in *Modern Mathematics for the Engineer*, edited by E. F. Beckenbach (McGraw-Hill, New York, 1956).

[15] C. W. J. Granger, Econometrica **37**, 424 (1969).

[16] J. Geweke, J. Am. Stat. Assoc. **77**, 304 (1982).

[17] J. Geweke, J. Am. Stat. Assoc. **79**, 907 (1984).

[18] X. Wang, Y. Chen, S. L. Bressler, and M. Ding, Int. J. Neural Syst. **17**, 71 (2007).

[19] Y. Chen, G. Rangarajan, J. Feng, and M. Ding, Phys. Lett. A **324**, 26 (2004).

[20] D. Marinazzo, M. Pellicoro, and S. Stramaglia, Phys. Rev. Lett. **100**, 144103 (2008).

[21] G. Wu, X. Duan, W. Liao, Q. Gao, and H. Chen, Phys. Rev. E **83**, 041921 (2011).

[22] L. Faes, G. Nollo, and A. Porta, Phys. Rev. E **83**, 051112 (2011).

[23] S. Guo, A. K. Seth, K. M. Kendrick, C. Zhou, and J. Feng, J. Neurosci. Meth. **172**, 79 (2008).

[24] X. Wu, C. Zhou, G. Chen, and J. Lu, Chaos **21**, 043129 (2011).

[25] W. Lin and H. F. Ma, Phys. Rev. E **75**, 066212 (2007).

[26] W. Yu, G. Chen, J. Cao, J. Lü, and U. Parlitz, Phys. Rev. E **75**, 067201 (2007).

[27] L. Chen, J. Lu, and C. K. Tse, IEEE Trans. Circ. Syst. II **56**, 310 (2009).

[28] R. FitzHugh, Biophys. J. **1**, 445 (1961).

[29] J. Nagumo, S. Arimoto, and S. Yoshizawa, Proc. IRE **50**, 2061 (1962).

[30] D. J. Watts and S. H. Strogatz, Nature (London) **393**, 440 (1998).

[31] M. E. J. Newman and D. J. Watts, Phys. Lett. A **263**, 341 (1999).

[32] Y. Wang, T. Joshi, X. S. Zhang, D. Xu, and L. Chen, Bioinformatics **22**, 2413 (2006).