SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
APPLIED MATHEMATICS
COMPLEX NETWORKS

Received
26 September 2013

Accepted
13 January 2014

Published
3 February 2014

Correspondence and requests for materials should be addressed to Y.D. (yhdo@knu.ac.kr)

Uncovering hidden nodes in complex networks in the presence of noise

Ri-Qi Su[1], Ying-Cheng Lai[1,2], Xiao Wang[3] & Younghae Do[4]

[1]School of Electrical, Computer, and Energy Engineering, Arizona State University, Tempe, AZ 85287, USA, [2]Department of Physics, Arizona State University, Tempe, AZ 85287, USA, [3]School of Biological and Health Systems Engineering, Arizona State University, Tempe, AZ 85287, USA, [4]Department of Mathematics, Kyungpook National University, Daegu, 702-701, South Korea.

Ascertaining the existence of hidden objects in a complex system, objects that cannot be observed from the external world, not only is curiosity-driven but also has significant practical applications. Generally, uncovering a hidden node in a complex network requires successful identification of its neighboring nodes, but a challenge is to differentiate its effects from those of noise. We develop a completely data-driven, compressive-sensing based method to address this issue by utilizing complex weighted networks with continuous-time oscillatory or discrete-time evolutionary-game dynamics. For any node, compressive sensing enables accurate reconstruction of the dynamical equations and coupling functions, provided that time series from this node and all its neighbors are available. For a neighboring node of the hidden node, this condition cannot be met, resulting in abnormally large prediction errors that, counterintuitively, can be used to infer the existence of the hidden node. Based on the principle of differential signal, we demonstrate that, when strong noise is present, insofar as at least two neighboring nodes of the hidden node are subject to weak background noise only, unequivocal identification of the hidden node can be achieved.

When dealing with an unknown complex system that has a large number of interacting components organized hierarchically, curiosity demands that we ask the following question: are there hidden objects that are not accessible from the external world? The problem of inferring the existence of hidden objects from observations is quite challenging but it has significant applications in many disciplines of science and engineering. Here by "hidden" we mean that no direct observation of or information about the object is available, and so it appears to the outside world as a black box. However, due to the interactions between the hidden object and other observable components in the system, it may be possible to utilize "indirect" information to infer the existence of the hidden object and to locate its position with respect to objects that can be observed. The difficulty to develop effective solutions is compounded by the fact that the indirect information on which any method of detecting hidden objects relies can be subtle and sensitive to changes in the system or in the environment. In particular, in realistic situations noise and random disturbances are present. It is conceivable that the "indirect" information can be mixed up with that due to noise or be severely contaminated. The presence of noise thus poses a serious challenge to detecting hidden nodes, and some effective "noise-mitigation" method must be developed.

To formulate the problem in a concrete way and to gain insights into the development of a general methodology, we note that the basic principle underlying the detection of hidden objects is that their existence typically leads to "anomalies" in the quantities that can be calculated or deduced from observation. Simultaneously, noise, especially local random disturbances applied at the nodal level, can also lead to large variance in these quantities. This is so because, a hidden node is typically connected to a few nodes in the network that are accessible to the external world, and a noise source acting on a particular node in the network may also be regarded as some kind of hidden object. Thus, the key to any detection methodology is to identify and *distinguish* the effects of hidden nodes on measures for detection from those due to *local* noise sources.

In this paper, we focus on complex networks and develop a general method to differentiate hidden nodes from local noise sources. This problem is intimately related to the works on reverse engineering of complex networks, where the goal is to uncover the full topology of the network based on measured time series[1–22]. Our method is based on the recent work[23] on utilizing compressive sensing[24–29] to detect hidden nodes in the absence of noise sources. To explain our method in a concrete setting, we use the network configuration shown schematically in Fig. 1, where there are 20 nodes, the couplings among the nodes are weighted, and the entire network is in a noisy environment, but a number of nodes also receive relatively strong random driving. We assume an oscillator network so that the nodal dynamics are described by nonlinear differential equations, and that time series can be

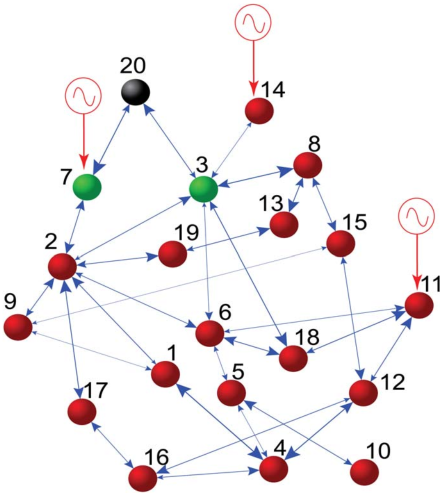

Figure 1 | An example of a complex network with a hidden node. Time series from all nodes except hidden node #20 can be measured, which can be detected when its immediate neighbors, nodes #3 and #7 are unambiguously identified. Nodes #7, #11, and #14 are driven by local noise sources.

measured simultaneously from all nodes in the network except one, labeled as #20, which is a hidden node. The task of ascertaining the presence and locating the position of the hidden node are equivalent to identifying its immediate neighbors, which are nodes #3 and #7 in Fig. 1. Note that, in order to be able to detect the hidden node based on information from its neighboring nodes, the interactions between the hidden node and its neighbors must be directional from the former to the latter or be bidirectional. Otherwise, if the coupling is solely from the neighbors to the hidden node, the dynamics of the neighboring nodes will not be affected by the hidden node and, consequently, time series from the neighboring nodes will contain absolutely no information about the hidden node, which is therefore undetectable. The action of local noise source on a node is naturally directional, i.e., from the source to the node.

Our recent work[23] demonstrated that, when the compressive-sensing paradigm is applied to uncovering the network topology[15], the predicted linkages associated with nodes #3 and #7 are typically anomalously dense, and this piece of information is basically what is needed to identify them as the neighboring nodes of the hidden node. In addition, when different segments of measurement data are used to reconstruct the coupling weights for these two nodes, the reconstructed weights associated with these two nodes exhibit significantly larger variances than those associated with other nodes. However, the predicted linkages associated with the nodes driven by local noise sources can exhibit behaviors similar to those due to the hidden nodes, leading to uncertainty in the detection of the hidden node. To address this critical issue is essential to developing algorithms for real-world applications, which is the aim of this paper. Our main idea is to exploit the principle of *differential signal* to study the behavior of the predicted link weights as a function of the data used in the reconstruction. Due to the advantage of compressive sensing, the required data amount can be quite small and, hence, even if our method requires systematic increase of the data amount, it will still be reasonably small. We shall argue and demonstrate that, when the various ratios of the predicted weights associated with all pairs of links between the possible neighboring nodes and the hidden node are examined, those associated with the hidden nodes and nodes under strong local noise show characteristically distinct behaviors, rendering unambiguous identification of the neighboring nodes of the hidden node. Any such ratio is essentially a kind of differential signal, because it is defined with respect to a pair of edges.

Results

We present our results by using coupled oscillator networks. (Results from evolutionary-game dynamical networks are presented in Supporting Information.) Given such a networked system, we use compressive sensing to uncover all the nodal dynamical equations and coupling functions[15]. This can be done by expanding all the vector fields and functions into series and calculating, from available time series, all the coefficients in the expansion. The expansion base needs to be chosen properly so that the number of non-zero coefficients is small as compared with the total number N_t of unknown coefficients. All N_t coefficients constitute a *coefficient vector* to be estimated. The amount of data used can be conveniently characterized by R_m, the ratio of the number M of data points used in the reconstruction, to N_t. See **Methods**.

Our idea to distinguish the effects of hidden node and local noise sources is based on the following observation. Consider two neighboring nodes of the hidden node, labeled as i and j. Because the hidden node is a common neighbor of nodes i and j, the couplings from the hidden node should be approximately proportional to each other, with the proportional constant determined by the ratio of their link weights with the hidden node. When the dynamical equations of nodes i and j are properly normalized, the terms due to the hidden node tend to cancel each other, leaving the normalization constant as a single unknown parameter that can be estimated subsequently. We name this parameter *cancellation ratio* and denote it by $\mathbf{\Omega}_{ij}$. As the data amount is increased, $\mathbf{\Omega}_{ij}$ tends to its true value. Practically we then expect to observe systematic changes in the estimated value of the ratio as data used in the compressive-sensing algorithm is increased from some small to relatively large amount. If only local noise sources are present, the ratio should show no systematic change with the data amount. Thus the distinct behaviors of $\mathbf{\Omega}_{ij}$ as the amount of data is increased provides a way to distinguish the hidden node from noise and, at the same time, to ascertain the existence of the hidden node. A mathematical formulation of this general principle can be found in **Methods**.

We test our method to differentiate hidden nodes and noise using random networks of nonlinear/chaotic oscillators. To be concrete, we choose the nodal dynamics to be that of the Rössler oscillator, one of the classical models in nonlinear dynamics[30],

$$[\dot{x}_i, \dot{y}_i, \dot{z}_i] = [-y_i - z_i, x_i + 0.2y_i, 0.2 + z_i(x_i - 5.7)],$$

which exhibits a chaotic attractor. The size of the network varies from 20 to 100, and the probability of connection between any two nodes is 0.04. The network link weights are equally distributed in [0.1, 0.5] (arbitrary). Background noise of amplitude ξ is applied (independently) to every oscillator in the network, with amplitude varying from 10^{-4} to 5×10^{-3}. The noise amplitude is thus smaller than the average coupling strength of the network. The tolerance parameter ε in the compressive sensing algorithm can be adjusted in accordance with the noise amplitude (see Supporting Information for details). Time series are generated by using the standard Heun's algorithm[31] to integrate the stochastic differential equations. To approximate the velocity field, we use third-order polynomial expansions in the compressive-sensing formulation. (In Supporting Information, we present more examples using network systems of varying sizes, different weight distributions and topologies, and alternative nodal dynamics.)

Detecting hidden node from time series. As a concrete example, we consider the network in Fig. 1, where only background noise is present and there are no local noise sources. Linear coupling between any pair of connected nodes is from the z-component to the x-component in the Rössler system. From the available time series (nodes #1–19), we solve the coefficient vector using a standard compressive-sensing algorithm [http://users.ece.gatech.edu/justin/l1magic/]. In particular, for node i, the terms associated with couplings from the z-components of other nodes appear in the ith row of the coupling matrix. As shown in Fig. 2(a), when the data amount is $R_m = 0.7$, the network's coupling matrix can be predicted. The predicted links and the associated weights are sparse for all nodes except for nodes #3 and #7, the neighbors of the hidden node. While there are small errors in the predicted weights due to background noise, the predicted couplings for the two neighbors of the hidden node, which correspond to the 3rd and the 7th row in the coupling matrix, appear to be from almost all other nodes in the network and some coupling strength is even negative. Such anomalies associated with the predicted coupling patterns of the neighboring nodes of the hidden node cannot be removed by increasing the data amount. Nonetheless, it is precisely these anomalies which hint at the likelihood that these two "abnormal" nodes are connected with a hidden node.

While abnormally high connectivity predicted for a node is likely indication that it belongs to the neighborhood of the hidden node, in complex networks there are hub nodes with abnormally large degrees, especially for scale-free networks[32]. In order to distinguish a hidden node's neighboring node from some hub node, we use the variance of the predicted coupling constants, which can be calculated from different segments of the available data sets. Due to the intrinsically low-data requirement associated with compressive sensing, the calculation of the variance is feasible because any reasonable time series can be broken into a number of segments, and prediction can be made from each data segment. For nodes not in the neighborhood of the hidden node, we expect the variance to be small as the predicted results hardly change when different segments of the time series are used. However, for the neighboring nodes of the hidden node, due to lack of complete information needed to construct the measurement matrix, the variance values can be much larger. Figure 2(b) shows the variance σ^2 in the predicted coupling strength for all 19 accessible nodes. We observe that the values of the variance for the neighboring nodes of the hidden node, nodes #3 and #7, are all above the upper dashed line and are in fact significantly larger than those associated with all other nodes that all fall below the lower dashed line. This indicates strongly that they are indeed the neighboring nodes of the hidden node. The gap between the two dashed lines can be taken as a quantitative measure of the detectability of the hidden node. The larger the gap, the more reliable it is to distinguish the neighbors of the hidden node from the nodes that not in the neighborhood. The results in Fig. 2 thus indicate that the locations of the hidden node(s) in the network can be reliably inferred even in the presence of weak background noise. The size of the gap, or the hidden-node detectability depends on the system details. In Supporting Information, we present results of a systematic analysis of the detectability measure, where we find that the variance due to the hidden nodes is mainly determined by the strength of their coupling with the accessible nodes in the network. We also find that system size and network topology have little effect on the hidden-node detectability. It is worth emphasizing that the detectability relies also on successful reconstruction of all nodes that are not in the neighborhood of the hidden nodes, which determine the lower dashed line in Fig. 2.

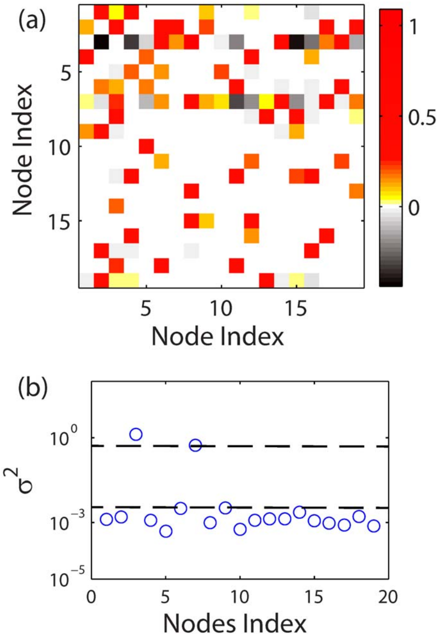

Figure 2 | For the network in Fig. (1), (a) predicted coupling matrix for all nodes except node #20. Time series from nodes #1 to #19 are available, while node #20 is hidden. The predicted weights are indicated by color coding and the amount of data used is $R_m = 0.7$. The abnormally dense patterns in the 3rd and 7th rows suggest that nodes #3 and #7 are the immediate neighbors of the hidden node. (b) Variance σ^2 of the predicted coefficients for all accessible nodes, which is calculated using 20 independent reconstructions based on different segments of the data. The variances associated with nodes #3 and #7 are apparently much larger than those with the other nodes, confirming that these are the neighboring nodes of the hidden node. There is a definite gap between the values of the variance associated with neighboring and non-neighboring nodes of the hidden node, as indicated by the two horizontal dashed lines in (b). When the local noise sources are applied to node #7, #11 and #14, these there nodes have similar dense bars in (a) and large variances in (b) (data are not shown).

To quantify the reliability of the reconstruction results, we investigate how the prediction errors in the link weights of all accessible nodes, except the predicted neighbors of the hidden node, change with the data amount. For an existent link, we use the normalized absolute error E_{nz}, the error in the estimated weight with respect to the true one, normalized by the value of the true link weight. Figure 3 shows the results for $N = 100$. The link weights are uniformly distributed in the interval $[0.1, 0.5]$ and the background noise amplitude is $\xi = 10^{-3}$. The tolerance parameter in the compressive-sensing algorithm is set to be $\varepsilon = 0.5$, which is optimal for this noise amplitude. (In Supporting Information we provide details of determining the optimal tolerance parameter for different values of the background noise amplitude.) We see that for $R_m > 0.4$, E_{nz} decreases to the small value of about 0.01, which is determined by background noise level. As R_m is increased further, the error is bounded by a small value determined by the noise amplitude, indicating that the reconstruction is robust. Although the value of E_{nz} does not decrease further toward zero due to noise, the prediction results are reliable in the sense that the predicted weights and the real values agree with each other, as shown in the inset of Fig. 3, a comparison of the actual and the predicted weights for all existent links. All the predicted results are in the vicinities of the corresponding actual values, as indicated by a heavy concentration of the dots along the diagonal line. The central region in the dot distribution has brighter color than the marginal regions, confirming that vast majority of the predicted results are accurate. In Supporting Information, we further show that robust reconstruction can be achieved regardless of the network size, connection topology and weight distributions, insofar as sufficient data are available.

The error measure E_{nz} to characterize the accuracy of the reconstruction is similar to z-scores, or the standard score in statistics, with the minor difference being that z-scores use the standard derivatives of the distribution to normalize the raw scores, while we use the exact values in our model examples. In realistic applications the exact values are usually not available, so it is necessary to use the z-score measure.

We emphasize that there are two types of "dense" connections: one from reconstruction and another intrinsic to the network. In

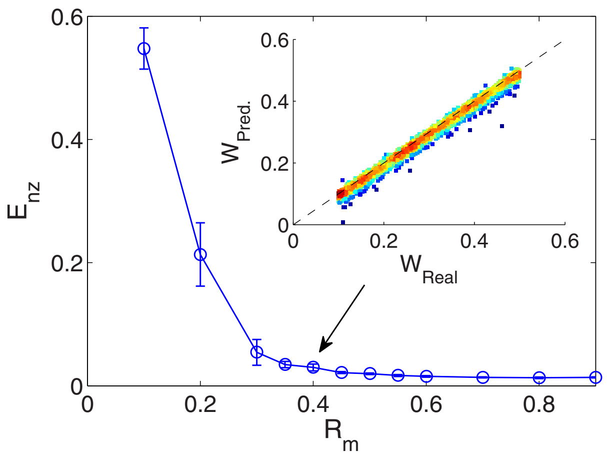

Figure 3 | For random networks of size $N = 100$ with uniform weight distribution in [0.1, 0.5], prediction error E_{nz} associated with nonzero coefficients of the dynamical equations of all nodes except for the neighboring nodes of the hidden node, as a function of normalized data amount R_m. The background noise amplitude is $\xi = 10^{-3}$ for all nodes. All data points are obtained from 10 independent realizations. Inset is a comparison of the predicted and actual weights for all existent links. Each dot represents one such link, and its x-value is the actual weight while the y-value is the corresponding predicted result. The color for each dot is determined by the dot density around it, while bright color signifies high density. The arrow indicates the value of R_m used in the comparison study. The tolerance of the compressive-sensing algorithm is set to be $\varepsilon = 0.5$.

particular, in the two-dimensional representation of the reconstruction results [e.g., Fig. 2(A)], the neighboring nodes of the hidden node typically appear densely linked to many other nodes in the network. These can be a result of lack of incomplete information (i.e., time series) due to the hidden node (in this case, there is indeed a hidden node), or the intrinsic dense connection pattern associated with, for example, a hub node in a scale-free network. Our idea of examining the variances of the reconstructed connections from independent data segments is for distinguishing these two possibilities. As we have demonstrated, extensive computations indicate that a combination of the dense connection and large variance can ascertain the existence of hidden node reliably.

Differentiating hidden node from local noise sources. When strong noise sources are present at certain nodes, the predicted coupling patterns of the neighboring nodes of these nodes will show anomalies. (Here by "strong" we mean that the amplitudes of the random disturbances are order-of-magnitude larger than that of background noise.) We now demonstrate that our proposed method based on the cancellation ratio is effective at distinguishing hidden nodes from local noise sources, insofar as the hidden node has at least two neighboring nodes not subject to such disturbances. To be concrete, we choose a network of $N = 61$ coupled chaotic Rössler oscillators, which has 60 accessible nodes and one hidden node (#61) that is coupled to two neighbors: nodes #14 and #20, as shown schematically in Fig. 4. Assume a strong noise source is present at node #54. We find that the reconstructed weights match their true values to high accuracy. We also find that the reconstructed coefficients including the ratio Ω_{ij} are all constant and invariant with respect to different data segments, a strong signal that the pair of nodes are the neighboring nodes of the same hidden node, thereby confirming its existence.

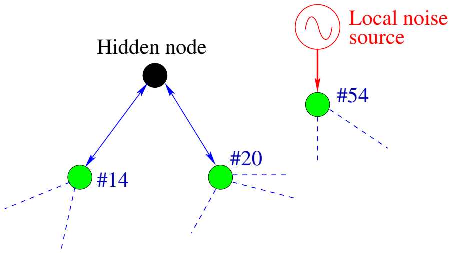

Figure 4 | Schematic illustration of a hidden node and its coupling configuration with two neighbors in a random network of $N = 61$ nodes, where 60 are accessible. A strong noise source is present at node #54.

When there are at least two accessible nodes in the neighborhood of the hidden node which are not subject to strong noisy disturbance, such as nodes #14 and #20, as the data amount R_m is increased towards 100%, the cancellation ratio should also increase and approach unity. This behavior is shown by the open circles in Fig. 5(a). However, when a node is driven by a local noise source, regardless of whether it is in the neighborhood of the hidden node, the cancellation ratio calculated from this node and any other accessible node in the network will show a characteristically different behavior. Consider, for example, nodes #14 and #54. The reconstructed connection patterns of these two nodes both show anomalies, as they appear to be coupled with all other nodes in the network. In contrast to the case where the pair of nodes are influenced by the hidden node only, here the cancellation ratio does not show any appreciable increase as the data amount is increased, as shown by the crosses in Fig. 5(a). In addition, the average variance values of the predicted coefficient vectors of the two nodes exhibit characteristically different behaviors, depending on whether any one node in the pair is driven by strong noise or not. In particular, for the node pair #14 and #20, since neither is under strong noise, the average variance will decrease toward zero as R_m approaches unity, as shown in Fig. 5(b)

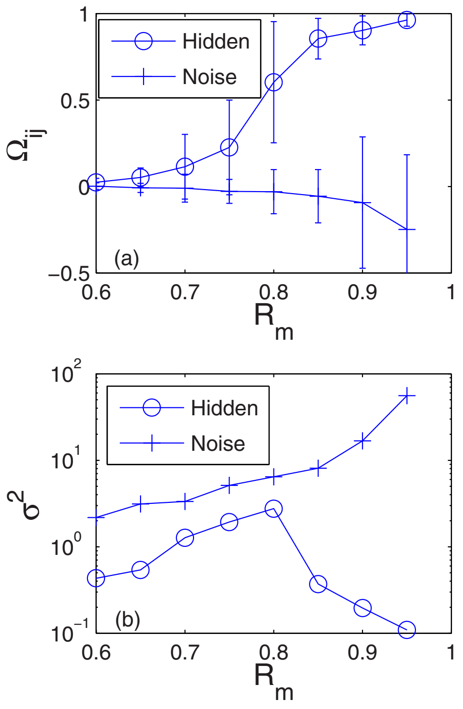

Figure 5 | For the network described in Fig. 4, (a) Predicted values of the cancellation ratio Ω_{ij} obtained from the differential signal of two neighboring nodes of the hidden node (#14 and #20, indicated by circles) and from the differential signal of nodes #14 and #54, where the latter is driven by noise of amplitude $\xi = 10^{-2}$ (crosses). (b) Average variance values of the predicted local coefficient vectors for the two combinations. The background noise amplitude is $\xi = 10^{-5}$. The results are obtained from 20 independent realizations.

(open circles). In contrast, for the node pair #14 and #54, the average variance will increase with R_m, as shown in Fig. 5(b) (crosses). This is because, when one node is under strong random driving, the input to the compressive-sensing algorithm will be noisy and its performance will deteriorate. However, compressive sensing can perform reliably when the input data are "clean," even when they are sparse. Increasing the data amount beyond a threshold is not necessarily helpful, but longer and noisy data sets can degrade significantly the performance. The results in Figs. 5(a,b) thus demonstrate that the cancellation ratio between a pair of nodes, in combination with the average variance of the predicted coefficient vectors associated with the two nodes, can effectively distinguish a hidden node from a local noise source. If there are more than one hidden node or there is a cluster of hidden nodes, the procedure to estimate the cancellation factors is similar but requires additional information about the neighboring nodes of the hidden nodes. Our cancellation-factor based method can be extended to network systems with nodal dynamics not of the continuous-time type, such as evolutionary-game dynamics. See Supporting Information for details.

Discussion

Our program to differentiate a hidden node from local noise sources and then to infer its existence can be summarized into the following steps:

- Collect time series of dynamical variables from accessible nodes;
- Hypothesize suitable expansion bases for nodal dynamics and coupling functions, taking advantage of physical understanding of the underlying networked dynamical system;
- Construct the measurement matrix and derivative vector from time series, and solve the expansion-coefficient vector using compressive sensing;
- Identify all nodes with abnormally dense connections, and calculate the corresponding variances using independent segments of the available time series to eliminate the hub nodes in the network (for those nodes the variances will be much smaller than those of the neighboring nodes of the hidden node or nodes under strong local noise);
- For all the remaining nodes with abnormally dense connections, calculate the cancellation ratio for all possible node pairs and also the average variance of the predicted coefficient vectors using non-overlapping time-series segments for a series of systematically increasing values of the data amount R_m;
- Identify the neighboring nodes of the hidden node as those with cancellation ratios approaching unity and the average variance tending to zero as R_m is increased. For those pairs with cancellation ratio not increasing and/or the average variance not decreasing with R_m, one node in the pair is under the driving of a local noise source.

Detecting hidden nodes in complex networks with *a priori* unknown nodal dynamics, topology, and coupling weights has vast application potential, such as in social and biological networks. Inferring the existence of hidden node in the presence of local random perturbations is an extremely challenging problem. Our efforts represent a step forward in this area of research, where much further work is needed.

Methods

Compressive-sensing based method to uncover network dynamics and topology. We consider the typical setting of a complex network of N coupled oscillators in a noisy environment. The dynamics of each individual node, when it is isolated from other nodes, can be described as $\dot{\mathbf{x}}_i = \mathbf{F}_i(\mathbf{x}_i) + \xi\eta_i$, where $\mathbf{x}_i \in \mathbb{R}^m$ is the vector of state variables, and η_i are an m-dimensional vector whose entries are independent Gaussian random variables of zero mean and unit variance, and ξ denotes the noise amplitude. A weighted network can be described by

$$\dot{\mathbf{x}}_i = \mathbf{F}_i(\mathbf{x}_i) + \sum_{j=1, j\neq i}^{N} \mathbf{W}_{ij} \cdot \left[\mathbf{H}(\mathbf{x}_j) - \mathbf{H}(\mathbf{x}_i)\right] + \xi\eta_i, \tag{1}$$

where $\mathbf{W}_{ij} \in \mathbb{R}^{m\times m}$ is the coupling matrix between node i and node j, and $\mathbf{H}$ is the coupling function. Defining

$$\mathbf{F}'_i(\mathbf{x}_i) \equiv \mathbf{F}_i(\mathbf{x}_i) - \mathbf{H}(\mathbf{x}_i) \cdot \sum_{j=1, j\neq i}^{N} \mathbf{W}_{ij},$$

we have

$$\dot{\mathbf{x}}_i = \mathbf{F}'_i(\mathbf{x}_i) + \sum_{j=1, j\neq i}^{N} \mathbf{W}_{ij} \mathbf{H}(\mathbf{x}_j) + \xi\eta_i, \tag{2}$$

i.e., we have grouped all terms directly associated with node i into $\mathbf{F}'_i(\mathbf{x}_i)$. We can then expand $\mathbf{F}'(\mathbf{x}_i)$ into the following form:

$$\mathbf{F}'_i(\mathbf{x}_i) = \sum_{\gamma} \tilde{\mathbf{a}}_i^{(\gamma)} \cdot \tilde{\mathbf{g}}_i^{(\gamma)}(\mathbf{x}_i), \tag{3}$$

where $\tilde{\mathbf{g}}_i^{(\gamma)}(\mathbf{x}_i)$ are a set of orthogonal and complete base functions chosen such that the coefficients $\tilde{\mathbf{a}}_i^{(\gamma)}$ are sparse. While the coupling function $\mathbf{H}(\mathbf{x}_i)$ can be expanded in a similar manner, for simplicity we assume that they are linear: $\mathbf{H}(\mathbf{x}_i) = \mathbf{x}_i$. We then have

$$\dot{\mathbf{x}}_i = \sum_{\gamma} \tilde{\mathbf{a}}_i^{(\gamma)} \cdot \tilde{\mathbf{g}}_i^{(\gamma)}(\mathbf{x}_i) + \sum_{j=1, j\neq i}^{N} \mathbf{W}_{ij} \cdot \mathbf{x}_j + \xi\eta_i, \tag{4}$$

where all the coefficients $\tilde{\mathbf{a}}_i^{(\gamma)}$ and weights $\mathbf{W}_{ij}$ need to be determined from time series $\mathbf{x}_i$. In particular, the coefficient vector $\tilde{\mathbf{a}}_i^{(\gamma)}$ determines the nodal dynamics and the weighted matrices $\mathbf{W}_{ij}$'s give the full topology and coupling strength of the entire network.

Suppose we have simultaneous measurements of all state variables $\mathbf{x}_i(t)$ and $\mathbf{x}_i(t + \delta t)$ at M different uniform instants of time at interval Δt apart, where $\delta t \ll \Delta t$ so that the derivative vector $\dot{\mathbf{x}}_i$ can be estimated at each time instant. Equation (4) for all M time instants can then be written in a matrix form with the following measurement matrix:

$$\mathbf{G}_i = \begin{pmatrix} \tilde{\mathbf{g}}_i(t_1) & \mathbf{x}_1(t_1) & \cdots & \mathbf{x}_k(t_1) & \cdots & \mathbf{x}_N(t_1) \\ \tilde{\mathbf{g}}_i(t_2) & \mathbf{x}_1(t_2) & \cdots & \mathbf{x}_k(t_2) & \cdots & \mathbf{x}_N(t_2) \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ \tilde{\mathbf{g}}_i(t_M) & \mathbf{x}_1(t_M) & \cdots & \mathbf{x}_k(t_M) & \cdots & \mathbf{x}_N(t_M) \end{pmatrix}, \tag{5}$$

where the index k in $\mathbf{x}_k(t)$ runs from 1 to N, $k \neq i$, and each row of the matrix is determined by the available time series at one instant of time. The derivatives at different time can be written in a vector form as $\mathbf{X}_i = [\dot{\mathbf{x}}_i(t_1), \cdots, \dot{\mathbf{x}}_i(t_M)]^T$, and the coefficients from the functional expansion and the weights associated with all links in the network can be combined concisely into a vector $\mathbf{a}_i$ as

$$\mathbf{a}_i = [\tilde{\mathbf{a}}_i, \quad W_{1i}, \cdots, \quad W_{i-1,i}, \quad W_{i+1,i}, \cdots, \quad W_{N,i}]^T, \tag{6}$$

where $[\cdot]^T$ denotes the transpose. For properly chosen expansion base and a general complex network whose connections are typically sparse, the vector $\mathbf{a}_i$ to be determined is sparse as well. Finally, Eq. (4) can be written as

$$\mathbf{X}_i = \mathbf{G}_i \cdot \mathbf{a}_i + \xi\eta_i. \tag{7}$$

In the absence of noise or if the noise amplitude is negligibly small, Eq. (7) represents a linear equation but the dimension of the unknown coefficient vector $\mathbf{a}_i$ can be much larger than that of $\mathbf{X}_i$, and the measurement matrix will have many more columns than rows. In order to fully recover the network of N nodes with each node having m components, it is necessary to solve $N \times m$ such equations.

Recovering signal from noisy measurement with compressive sensing algorithm. The system of linear equations in Eq. (7 is ill defined. However, since $\mathbf{a}_i$ is sparse, insofar as its number of non-zero coefficients is smaller than the dimension of $\mathbf{X}_i$, the vector $\mathbf{a}_i$ can be uniquely and efficiently determined by the compressive-sensing algorithm[24–29]. In particular, in the equation $\mathbf{X} = \mathbf{G} \cdot \mathbf{a} + \xi$, reliable recovery of the P-dimension sparse vector $\mathbf{a}$ is achievable, according to[25], where $\mathrm{X} \in \mathbb{R}^{Q\times 1}$ and $\mathrm{G} \in \mathbb{R}^{Q\times P}$ but $P \ll Q$. A sufficiently sparse vector $\mathbf{a}$ can be reconstructed by solving the following l_1 regularization problem:

$$\min \|\mathrm{a}\|_{l_1}, \quad \text{subject to} \quad \|\mathrm{G}\cdot\mathrm{a} - \mathrm{X}\|_{l_2} \leq \varepsilon, \tag{8}$$

where the l_1 norm for a vector $\mathbf{x}$ is defined as $\|\mathrm{x}\|_{l_1} = \sum_{i=1}^{n} |x_i|$, its l_2 norm is $\|\mathrm{x}\|_{l_2} = \sum_{i=1}^{n} \left|x_i^2\right|$, ε is the threshold value determined by the noise amplitude. The reconstructed vector $\bar{\mathbf{a}}$ lies within the range: $\|\bar{\mathbf{a}} - \mathrm{a}\| \leq C\cdot\varepsilon$, where C is a constant.

Detection of hidden node. To motivate our consideration, we note that, a meaningful solution of Eq. (7) based on compressive sensing requires that the derivative vector $\mathbf{X}_i$ and the measurement matrix $\mathbf{G}_i$ be entirely known which, in turn, requires time series from all nodes. In this case, we say that information available for reconstruction of the complex networked system is *complete*. In the presence of a hidden node, for its immediate neighbors, the available information will not be complete in the sense that some entries of the vector $\mathbf{X}_i$ and the matrix $\mathbf{G}_i$ become now unknown. Let h denote the hidden node. For any neighboring node of h, the vector $\mathbf{X}_i$ and the matrix $\mathbf{G}_i$ in Eq. (7) now contain unknown entries at the locations specified by the index h. For any other node not in the immediate neighborhood of h, Eq. (7) is unaffected. When compressive-sensing algorithm is used to solve Eq. (7), there will then be large errors in the solution of the coefficient vector $\mathbf{a}_i$ associated the neighboring nodes of h, regardless of the amount of data used. In general, the so-obtained coefficient vector $\mathbf{a}_i$ will not appear sparse. Instead, most of its entries will not be zero, a manifestation of which is that the node would appear to have links with almost every other node in the network. In contrast, for nodes not in the neighborhood of h, the corresponding errors will be small and can be reduced by increasing the data amount, and the corresponding coefficient vector will be sparse. It is this observation which makes identification of the neighboring nodes of the hidden node possible in the noiseless or weak-noise situations[23].

To appreciate the need and the importance to distinguish the effects of hidden node from these of noise, we can separate the term associated with h in Eq. (4) from those with other accessible nodes in the network. Letting l denote a node in the immediate neighborhood of the hidden node h, we have

$$\mathbf{X}_l = \mathbf{G}_l' \cdot \mathbf{a}_l' + (\mathbf{W}_{lh} \cdot \mathrm{x}_h + \xi \eta_l), \tag{9}$$

where $\mathbf{G}_l'$ is the new measurement matrix that can be constructed from all available time series. While background noise may be weak, the term $\mathbf{W}_{lh} \cdot \mathbf{x}_h$ can in general be large in the sense that it is comparable in magnitude with other similar terms in Eq. (4). Thus, when the network is under strong noise, especially for those nodes that are connected to the neighboring nodes of the hidden node, the effects of hidden node on the solution can be entangled with those due to noise. In addition, if the coupling strength from the hidden node is weak, it would be harder to identify the neighboring nodes. For example, hidden node in a network with Gaussian weight distribution will be harder to detect, due to the finite probability of the occurrence of near zero weights. When the coupling strength is comparable or smaller than the background noise amplitude, the corresponding link cannot be detected. See Supporting Information for details.

Method to distinguish hidden nodes from local noise sources - a mathematical formulation. For simplicity, we assume that all coupled oscillators share the same coupling scheme and that each oscillator is coupled to any of its neighbors through one component of the state vector only. Thus, each row in the coupling matrix $\mathbf{W}_{ih}$ associated with a link between node i and h has only one non-zero element. Let p denote the component of the hidden node coupled to the first component of node i, the dynamical equation of which can then be written as

$$[\dot{\mathbf{x}}_i]_1 = \left[\sum_{\gamma} \tilde{\mathbf{a}}_i^{(\gamma)} \cdot \tilde{\mathbf{g}}_i^{(\gamma)}(\mathrm{x}_i)\right]_1 + \left[\sum_{k \neq i,h}^{N} \mathbf{W}_{ij} \cdot \mathrm{x}_j\right]_1 + w_{ih}^{1p} \cdot [\mathrm{x}_h]_p + \xi \eta_i, \tag{10}$$

where $[\mathbf{x}_h]_p$ denotes the time series of the pth component of the hidden node, which is unavailable, and w_{ih}^{1p} is the coupling strength between the hidden node and node i. The dynamical equation of the first component of neighboring node j of the hidden node has a similar form. Letting

$$\Omega_{ij} = w_{ih}^{1p} \Big/ w_{jh}^{1p}, \tag{11}$$

be the cancellation ratio, multiplying Ω_{ij} to the equation of node j, and subtracting from it the equation for node i, we obtain

$$\begin{aligned} [\dot{\mathbf{x}}_i]_1 = {} & \Omega_{ij}\left[\dot{\mathbf{x}}_j\right]_1 + \sum_{\gamma} \tilde{\mathbf{a}}_i^{(\gamma)} \cdot \tilde{\mathbf{g}}_i^{(\gamma)}(\mathrm{x}_i) + \sum_{k \neq i,h} w_{ik}^{1p} [\mathrm{x}_k]_p \\ & - \Omega_{ij} \sum_{\gamma} \tilde{\mathbf{a}}_j^{(\gamma)} \cdot \tilde{\mathbf{g}}_j^{(\gamma)}\left(\mathrm{x}_j\right) - \Omega_{ij} \sum_{k \neq j,h} w_{jk}^{1p} [\mathrm{x}_k]_p \\ & + \left(w_{ih}^{1p} - \Omega_{ij} w_{jh}^{1p}\right) \cdot [\mathrm{x}_h]_p + \xi \eta_i - \Omega_{ij} \xi \eta_j. \end{aligned} \tag{12}$$

We see that terms associate with $[\mathbf{x}_h]_p$ vanish and all deterministic terms on the left-hand side of Eq. (12) are known, which can then be solved by the compressive-sensing method. From the coefficient vector so estimated, we can identify the coupling of nodes i and j to other nodes, except for the coupling between themselves since such terms have been absorbed into the nodal dynamics, and the couplings to their common neighborhood are degenerate in Eq. (12) and cannot be separated from each other. Effectively, we have combined the two nodes together by introducing the cancellation ratio Ω_{ij}.

To give a concrete example, we consider the situation where each oscillator has three independent dynamical variables, named as x, y and z. For the nodal and coupling dynamics we choose polynomial expansions of order up to n. The x component of the nodal dynamics $[\mathbf{F}_i'(\mathrm{x}_i)]_x$ for node i is:

$$[\mathbf{F}_i'(\mathrm{x}_i)]_x = \sum_{l_x=0}^{n} \sum_{l_y=0}^{n} \sum_{l_z=0}^{n} \left[a_{l_x l_y l_z}\right]_x \cdot x_i^{l_x} y_i^{l_y} z_i^{l_z},$$

and the coupling from other node k to the x component can be written as

$$C_{ik}^x = w_{ik}^{xx} \cdot x_k + w_{ik}^{xy} \cdot y_k + w_{ik}^{xz} \cdot z_k,$$

where w_{ik}^{xy} denotes the coupling weight from the y component of node k to the x component of node i, and so on. The nodal dynamical terms in the matrix $\mathbf{G}_i$ are

$$\left[\tilde{\mathrm{g}}_i\right]_x = \left[x_i^0 y_i^0 z_i^0, x_i^1 y_i^0 z_i^0, \cdots, x_i^n y_i^n z_i^n\right],$$

and the corresponding coefficients are $\left[a_{l_x l_y l_z}\right]_x$. The vector of coupling weights is $\left[\mathbf{W}_{ij}\right]_x = \left[w_{ij}^{xx}, w_{ij}^{xy}, w_{ij}^{xz}\right]$. Equation (12) becomes

$$\begin{pmatrix} \dot{x}_i(t_1) \\ \dot{x}_i(t_2) \\ \vdots \\ x_i(t_M) \end{pmatrix} \approx \begin{pmatrix} \dot{x}_j(t_1) & 1 & [\tilde{\mathbf{g}}_i(t_1)]_x & \left[\tilde{\mathbf{g}}_j(t_1)\right]_x & x_1(t_1) & \cdots & z_N(t_1) \\ \dot{x}_j(t_2) & 1 & [\tilde{\mathbf{g}}_i(t_2)]_x & \left[\tilde{\mathbf{g}}_j(t_2)\right]_x & x_1(t_2) & \cdots & z_N(t_2) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ \dot{x}_j(t_M) & 1 & [\tilde{\mathbf{g}}_i(t_M)]_x & \left[\tilde{\mathbf{g}}_j(t_M)\right]_x & x_1(t_M) & \cdots & z_N(t_M) \end{pmatrix} \cdot \begin{pmatrix} \Omega_{ij} \\ c \\ \tilde{\mathbf{a}}_i' \\ -\Omega_{ij} \cdot \tilde{\mathbf{a}}_j' \\ w_{i1}^{xx} - \Omega_{ij} x_{j1}^{xx} \\ \vdots \\ w_{iN}^{xz} - \Omega_{ij} w_{jN}^{xz} \end{pmatrix}, (13) \qquad (13)$$

where c is the sum of constant terms from the dynamical equations of nodes i and j, and $\tilde{\mathbf{a}}_i'$ is the coefficient vector to be estimated which excludes all the constants. Using compressive sensing to solve Eq. (13), we can recover the cancellation ratio Ω_{ij} and the equations of node i. When Ω_{ij} is known we can then recover the dynamics of node j from the coefficient vector $-\Omega_{ij} \cdot \tilde{\mathbf{a}}_j'$.

In Supporting Information we provide an analysis and discussion about the possible extension of our method to systems of characteristically different nodal dynamics and/or with multiple hidden nodes. In particular, we show that the method can be readily adopted to network systems whose nodal dynamics are not described by continuous-time differential equations but by discrete-time processes such as evolutionary-game dynamics. In such a case, the derivatives used for continuous-time systems can be replaced by the agent payoffs. The cancellation factors can then be calculated from data to differentiate the hidden nodes from local noise sources. We also show that, under certain conditions with respect to the coupling patterns between the hidden nodes and their neighboring nodes, the cancellation factors can be estimated even when there are multiple, entangled hidden nodes in the network.

1. Fortunato, S. Community detection in graphs. *Phys. Rep.* **486**, 75–174 (2010).
2. Gardner, T. S., di Bernardo, D., Lorenz, D. D. & Collins, J. J. Inferring Genetic Networks and Identifying Compound Mode of Action via Expression Profiling. *Science* **301**, 102–105 (2003).
3. Gruen, S., Diesmann, M. & Aertsen, A. Unitary Events in Multiple Single Neuron Spiking Activity. I. Detection and Significance. *Neural Comp.* **14**, 43–80 (2002).
4. Gütig, R., Aertsen, A. & Rotter, S. Statistical significance of coincident spikes: count-based versus rate-based statistics. *Neural Comp.* **14**, 121–153 (2002).
5. Pipa, G. & Grün, S. Non-parametric significance estimation of joint-spike events by shuffling and re-sampling. *Neurocomputing* **52–54**, 31–37 (2003).
6. Bongard, J. & Lipson, H. Automated reverse engineering of nonlinear dynamical systems. *PNAS* **104**, 9943–9948 (2007).
7. Timme, M. Revealing network connectivity from response dynamics. *Phys. Rev. Lett.* **98**, 224101 (2007).
8. Napoletani, D. & Sauer, T. D. Reconstructing the topology of sparsely connected dynamical networks. *Phys. Rev. E* **77**, 026103 (2008).
9. Sontag, E. Network reconstruction based on steady-state data. *Essays Biochem.* **45**, 161–176 (2008).
10. Wang, W.-X. *et al.* Scaling of noisy fluctuations in complex networks and applications to network prediction. *Phys. Rev. E* **80**, 016116 (2009).
11. Ren, J., Wang, W.-X., Li, B. & Lai, Y.-C. Noise bridges dynamical correlation and topology in coupled oscillator networks. *Phys. Rev. Lett.* **104**, 058701 (2010).
12. Levnajić, Z. & Pikovsky, A. Network reconstruction from random phase resetting. *Phys. Rev. Lett.* **107**, 034101 (2011).
13. Hempel, S., Koseska, A., Kurths, J. & Nikoloski, Z. Inner composition alignment for inferring directed networks from short time series. *Phys. Rev. Lett.* **107**, 054101 (2011).
14. Shandilya, S. G. & Timme, M. Inferring network topology from complex dynamics. *New J. Phys.* **13**, 013004 (2011).
15. Wang, W.-X. *et al.* Time-series based prediction of complex oscillator networks via compressive sensing. *EPL* **94**, 48006 (2011).
16. Pan, W., Yuan, Y. & Stan, G.-B. Reconstruction of Arbitrary Biochemical Reaction Networks: A Compressive Sensing Approach. 51st IEEE Conference on Decision and Control. Maui, Hawaii, USA (2012, December 10–13).
17. Yuan, Y., Stan, G.-B., Warnick, S. & Goncalves, J. Robust dynamical network reconstruction. *Automatica* **47**, 1230–1235 (2011).
18. Mastromatteo, I., Zarinelli, E. & Marsili, M. Reconstruction of financial networks for robust estimation of systemic risk. *J. Stat. Mech.* **2012**, P03011 (2012).

19. Fagiolo, G., Squartini, T. & Garlaschelli, D. Null models of economic networks: the case of the world trade web. *J. Econ. Interact. Coord.* **8**, 75–107 (2013).
20. Musmeci, Nicol. *et al.* Bootstrapping topological properties and systemic risk of complex networks using the fitness model. *J. Stat. Phys.* **151**, 720–734 (2013).
21. Mastrandrea, R., Squartini, T., Fagiolo, G. & Garlaschelli, D. Enhanced network reconstruction from irreducible local information. arXiv:1307.2104.
22. Caldarelli, G. *et al.* Reconstructing a credit network. *Nature Physics* **9**, 125–126 (2013).
23. Su, R.-Q., Wang, W.-X. & Lai, Y.-C. Detecting hidden nodes in complex networks from time series. *Phys. Rev. E* **106**, 058701 (2012).
24. Candeѐs, E., Romberg, J. & Tao, T. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Theory* **52**, 489–509 (2006).
25. Candeѐs, E., Romberg, J. & Tao, T. Stable signal recovery from incomplete and inaccurate measurements. *Commun. Pure Appl. Math.* **59**, 1207–1223 (2006).
26. Candeѐs, E. Compressive sampling. *in Proceedings of the International Congress of Mathematicians.* Madrid, Spain (2006).
27. Donoho, D. Compressed sensing. *IEEE Trans. Inf. Theory* **52**, 1289–1306 (2006).
28. Baraniuk, R. G. Compressive Sensing. *IEEE Signal Process. Mag.* **24**, 118–121 (2007).
29. Candeѐs, E. & Wakin, M. An introduction to compressive sampling. *IEEE Signal Process. Mag.* **25**, 21–30 (2008).
30. Rössler, O. E. An equation for continuous chaos. *Phys. Lett. A* **57**, 397–398 (1976).
31. Gardiner, C. W. *Handbook of Stochastic Methods* (Springer, Berlin, 1985).
32. Barabási, A.-L. & Albert, R. Emergence of Scaling in Random Networks. *Science* **286**, 509–512 (1999).

Acknowledgments

We thank Dr. W.-X. Wang for discussions. This work was supported by AFOSR under Grant No. FA9550-10-1-0083 and by NSF under Grant. No. CDI-1026710, and by Basic Science Research Program of the Ministry of Education, Science and Technology under Grant No. NRF-2013R1A1A2010067, and by NSF under Grant No DMS-1100309 and by Heart Association under Grant No 11BGIA7440101.

Author contributions

Y.C.L., R.Q.S., X.W. and Y.D. devised the research project. R.Q.S. performed numerical simulations. Y.C.L. and R.Q.S. analyzed the results. Y.C.L., R.Q.S. and Y.D. wrote the paper.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Su, R.-Q., Lai, Y.-C., Wang, X. & Do, Y. Uncovering hidden nodes in complex networks in the presence of noise. *Sci. Rep.* **4**, 3944; DOI:10.1038/srep03944 (2014).