Detecting hidden nodes in complex networks from time series

Ri-Qi Su, Wen-Xu Wang, and Ying-Cheng Lai

Phys. Rev. E **85**, 065201 — Published 29 June 2012

DOI: 10.1103/PhysRevE.85.065201

Detecting hidden nodes in complex networks from time series

Ri-Qi Su,[1] Wen-Xu Wang,[2] and Ying-Cheng Lai[1,3]

[1]*School of Electrical, Computer, and Energy Engineering, Arizona State University, Tempe, AZ 85287, USA*
[2]*Department of Systems Science, School of Management and Center for Complexity Research, Beijing Normal University, Beijing 100875, China*
[3]*Department of Physics, Arizona State University, Tempe, AZ 85287, USA*

We develop a general method to detect hidden nodes in complex networks, using only time series from nodes that are accessible to external observation. Our method is based on compressive sensing and we formulate a general framework encompassing continuous and discrete-time, and evolutionary-game type of dynamical systems as well. For concrete demonstration, we present an example of detecting hidden nodes from an experimental social network. Our paradigm for detecting hidden nodes is expected to find applications in a variety of fields where identifying hidden or black-boxed objects based on limited amount of data is of interest.

PACS numbers: 05.45.-a,89.75.-k

The power of science lies in its ability to infer and predict the existence of objects from which no direct information can be obtained experimentally or observationally. A well known example is to ascertain the existence of black holes of various masses in different parts of the universe from indirect evidence, such as X-ray emissions. In the field of complex networks, the problem of detecting *hidden* nodes can be stated, as follows. Consider a network whose topology is completely unknown but whose nodes consist of two types: one accessible and another inaccessible from the outside world. The accessible nodes can be observed or monitored, and we assume that time series are available from each node in this group. The inaccessible nodes are shielded from the outside and they are essentially "hidden." The question is, can we infer, based solely on the available time series from the accessible nodes, the existence and locations of the hidden nodes? Since no data from the hidden nodes are available, nor can they be observed directly, they act as some sort of "black box" from the outside world. Despite recent works on uncovering network topologies [1–6], to our knowledge, the problem of detecting hidden nodes in complex networks has not been addressed. Solution of the problem, however, has potential applications in different fields of significant current interest. For example, to uncover the topology of a terrorist organization and especially, various ring leaders of the network is a critical task in defense. The leaders may be hidden in the sense that no direct information about them can be obtained, yet they may rely on a number of couriers to operate, which are often subject to surveillance. Similar situations arise in epidemiology, where the original carrier of a virus may be hidden, or in a biology network where one wishes to detect the most influential node from which no direct observation can be made.

In this paper, we present a completely data-driven, compressive-sensing based [7] approach to inferring the existence and locations of hidden nodes in complex networks. The general principle underlying our method can be understood by referring to Fig. 1(a) where, for illustrative purpose, a network of 20 nodes with directed interactions is shown. Suppose nodes No. $1-19$ are accessible to the external world, while node No. 20 (in gray) is hidden and thus inaccessible from the outside. The hidden node has two neighbors: No. 9 and No. 18 (in green), and the remaining 17 nodes are marked as red. Every red node thus has the property that time series from itself and *all* its neighbors are available, but for each green node, although time series from itself is available, the same is not true for all its neighbors due to its link with the hidden node. Generally, the time series can be regarded as being generated by the combination of nodal and coupling dynamics, and one wishes to base on the time series to predict the various dynamical equations so that the dynamical processes on various nodes and the network topology can be uncovered. As we shall demonstrate, for a given node, this can indeed be achieved provided that time series from the node and all its neighbors are available. Referring to Fig. 1(a), this means that the dynamical equations and the links from/to all red nodes can be predicted. However, significant errors would arise in the prediction of the green nodes due to incompleteness of information about their neighbors. By examining the prediction errors of all accessible nodes, the ones that are connected to the hidden node will then show anomalies, providing a way to infer its existence and location (e.g., connected to the two green nodes in Fig. 1(a).

The paradigm of compressive sensing [7] aims to reconstruct a sparse vector $\mathbf{a} \in \mathbb{R}^N$ from linear measurements $\mathbf{M}$ in the form $\mathbf{M} = \mathbf{G} \cdot \mathbf{a}$, where $\mathbf{M} \in \mathbb{R}^K$ and $\mathbf{G}$ is an $K \times N$ matrix. The compressive sensing theory [7] guarantees that, when most components in the unknown vector $\mathbf{a}$ are zero, it can be reconstructed by fewer measurements than the number of components. The unknown vector $\mathbf{a}$ can be solved, for example, by a convex optimization procedure based on L_1 norm. Our recent work [6] demonstrated that the problem of data-based network reconstruction can be casted into the form of $\mathbf{M} = \mathbf{G} \cdot \mathbf{a}$.

We consider networked systems for which the nodal dynamics, described by the vector function $\mathbf{F}_i(\mathbf{x}_i)$, can be separated from the interactions or coupling with other nodes in the network, mathematically described by the coupling function $\mathbf{H}_{ij}(\mathbf{x}_i, \mathbf{x}_j)$. The system can then be written as $\mathbf{M}_i = \mathbf{F}_i(\mathbf{x}_i) + \sum_{j \neq i}^{N} w_{ij}\mathbf{H}_{ij}(\mathbf{x}_i, \mathbf{x}_j)$, where $\mathbf{M}_i$ is the system response, either in discrete or continuous time. For example, for discrete-time mapping system, $\mathbf{M}_i$ are the state variables at the next time step, while in continuous system $\mathbf{M}_i$ are the derivatives of the corresponding variables. To illus-

trate our method to detect hidden nodes in a concrete manner, we assume that the nodal and coupling functions can be written as some series expansion, e.g., power or Fourier series. In particular, we write: $\mathbf{F}_i(\mathbf{x}_i) = \sum_\gamma \tilde{a}_i^{(\gamma)} \tilde{g}_i^{(\gamma)}(\mathbf{x}_i)$ and $\mathbf{H}_{ij}(\mathbf{x}_i, \mathbf{x}_j) = \sum_\beta a_{ij}^{(\beta)} g_{ij}^{(\beta)}(\mathbf{x}_i, \mathbf{x}_j)$, where $\tilde{g}_i^{(\gamma)}$ are the expansion bases associated with $\mathbf{x}_i$ only, and $g_{ij}^{(\beta)}$ are with respect to both $\mathbf{x}_i$ and $\mathbf{x}_j$. Next we combine the bases $\tilde{\mathbf{g}}_i(t)$ and $\mathbf{g}_{ij}(t)$ at time t into a row vector, and the coefficients $\mathbf{a}_i^{(\alpha)}$ and $\mathbf{a}_{ij}^{(\beta)}$ into a constant column vector. The time-series vector of responses $\mathbf{M}_i(t)$ for node i can then be expressed by the product of the matrix $\mathbf{G}_i$ and the *to-be-determined* coefficient vector $\mathbf{a}_i$, with $\mathbf{G}_i$ given by

$$\mathbf{G}_i = \begin{pmatrix} \tilde{\mathbf{g}}_i(t_1) & \mathbf{g}_{i1}(t_1) & \cdots & \mathbf{g}_{ij}(t_1) & \cdots & \mathbf{g}_{iN}(t_1) \\ \tilde{\mathbf{g}}_i(t_2) & \mathbf{g}_{i1}(t_2) & \cdots & \mathbf{g}_{ij}(t_2) & \cdots & \mathbf{g}_{iN}(t_2) \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ \tilde{\mathbf{g}}_i(t_m) & \mathbf{g}_{i1}(t_m) & \cdots & \mathbf{g}_{ij}(t_m) & \cdots & \mathbf{g}_{iN}(t_m) \end{pmatrix}, \quad (1)$$

where $\tilde{\mathbf{g}}_i(t)$ is the set of bases of $\mathbf{F}_i(\mathbf{x}_i)$, and $\mathbf{g}_{ij}(t)$ is the set of expansion bases of $\mathbf{H}_{ij}(\mathbf{x}_i, \mathbf{x}_j)$. Elements in the vector $\mathbf{M}_i(t)$ contain system response $m_i(t)$ at different t. In particular, when the vector $\mathbf{a}_i$ is determined by solving $\mathbf{M} = \mathbf{G} \cdot \mathbf{a}$, the dynamical equations for the set of corresponding variables at all nodes become known. Note that the vector $\mathbf{a}_i$ contains all the coupling weights from other nodes to i as in $\mathbf{g}_{ij}(t)$ and complete information about the nodal dynamical equations as in $\tilde{\mathbf{g}}_i(t)$. Previous works [6] demonstrated that solutions to the compressive sensing problem can be obtained but only when time series from *all* nodes are available, i.e., when there is no hidden object.

To devise a compressive-sensing based methodology for detecting hidden nodes, we consider the case of one hidden node (or one cluster of hidden nodes). Let node i be one of the immediate neighbors of the hidden node. Due to lack of time series from the hidden node, the form $\mathbf{M} = \mathbf{G} \cdot \mathbf{a}$ is violated for node i, despite the available time series from other nodes in the network. That is, due to the missing time series from the hidden node and consequently missing elements in $\mathbf{a}$, it is not possible to obtain the true solution of the dynamical equations of node i. If a node does not neighbor any hidden node, time series from itself and all its direct neighbors are available, rendering valid the form $\mathbf{M} = \mathbf{G} \cdot \mathbf{a}$ for such a node. The practical importance is that the errors in the prediction of the dynamics of the immediate neighbors of the hidden node will be much larger than those associated with nodes that do not have any hidden node in their neighborhoods. The predicted characteristics of all neighboring nodes of the hidden node will then show significant anomalies as compared with those of other nodes. The anomalies can then be used to identify all nearest neighbors of the hidden node, which in turn imply its existence and its position in the network.

While our general idea of detecting hidden nodes can be formulated using different types of dynamical systems, to be concrete we describe how this can be done using evolutionary-game type of dynamics. Such dynamical processes can be used to model generic agent-to-agent interactions in economical, social, or even certain biological networks [8, 9].

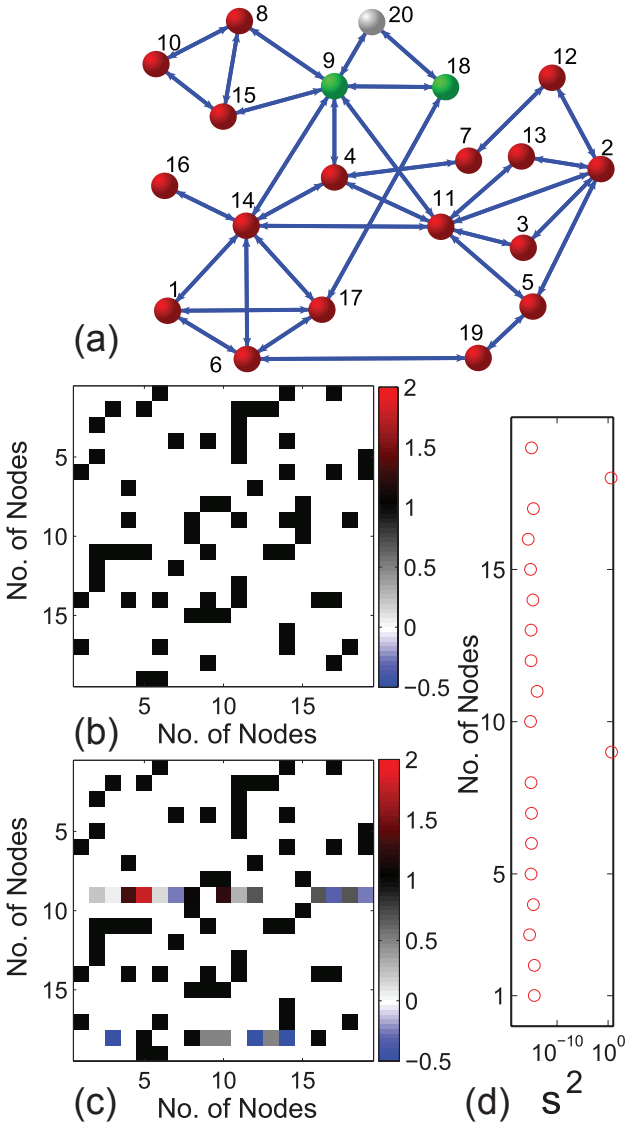

FIG. 1: (Color online.) (a) Illustration of a complex network with a hidden node. (b) Representation of the true adjacency matrix, (c) reconstructed adjacency matrix elements for nodes except the hidden node based on time series from these nodes. (d) Variance σ^2 of the reconstructed coefficient vector $\mathbf{a}$ for all nodes, calculated by using 10 different random segments from the available experimental time series. The variances of the two green nodes (No. 9 and No. 18) are much larger than those of the red nodes, indicating that they are the neighbors of the hidden node.

In an evolutionary-game system, the neighbors of the hidden node can be identified by utilizing the stability criterion with respect to different measurements. More specifically, in an evolutionary-game system, at any time a player can take on one of two strategies: cooperation (C) or defection (D), mathematically represented as $\mathbf{S}(C) = (1, 0)^T$ and $\mathbf{S}(D) = (0, 1)^T$, respectively. The payoffs of the two players in a game are determined by their strategies and the payoff matrix $\mathbf{P}$. For example, for the classical prisoner's dilemma

game (PDG), the matrix elements are $P_{11} = 0$, $P_{12} = 0$, $P_{21} = b$, and $P_{22} = 0$, where $1 < b < 2$ is a parameter characterizing the temptation to defect. At each time step, all agents in the network play the game with their neighbors simultaneously and gain rewards. For agent i, the reward is $m_i = \sum_j a_{ij} \mathbf{S}_i^T \mathbf{P} \mathbf{S}_j$, where $\mathbf{S}_i$ and $\mathbf{S}_j$ denote the strategies of agents i and j taken at the time and a_{ij} is the coupling strength between them. After obtaining its payoff, an agent updates its strategy according to its own and neighbors' payoffs, attempting to maximize its payoff at the next round. We assume that the strategy and payoff data of agents are available except those of the hidden node. In particular, we choose $\mathbf{g}_{ij}(t) = \mathbf{S}_i^T(t) \cdot \mathbf{P} \cdot \mathbf{S}_j(t)$ and ignore $\tilde{\mathbf{g}}_i$, the payoff of node i at different time t can be expressed as $\mathbf{M}_i(t) = \mathbf{G}_i \cdot \mathbf{a}_i$, where $\mathbf{G}_i$ is to be constructed as specified in Eq. (1), and the vector $\mathbf{a}_i$ to be determined contains all interaction strength between nodes i and other accessible nodes in the network. The network structure is uncovered after $\mathbf{a}$'s for all nodes are determined.

As an example, we present results of experimentally detecting a hidden node from a social network hosting evolutionary-game dynamics. In the experiment, 20 participants from Arizona State University played the prisoner's dilemma game (PDG) iteratively, with a pre-specified payoff parameter. The player with the highest normalized payoff (total payoffs normalized by their degrees) summed over time was the winner. The players can gamble with all their nearest neighbors in the pre-existing social network [Fig. 1(a)]. The network was determined by surveying the friendships among those participants, and it exhibits some typical properties of real social network, such as the much larger degree in some hub nodes. During the experiment, the strategies of each player and the gained payoff were recorded in all the 32 rounds, except for the hidden node No. 20. The true adjacency matrix of accessible nodes is represented in Fig. 1(b), and the predicted matrix is shown in Fig. 1(c). We see that the links of the two neighboring nodes (No. 9 and No. 18) of the hidden node No. 20 cannot be predicted. Especially, the two nodes are predicted to have links with almost all nodes in the network, which is highly unlikely for a random network that is typically sparse. While the predicted loss of sparsity of certain nodes is an indication that they might be in the neighborhood of some hidden node, the condition is not sufficient in general, because of the existence of hub nodes with significantly more links than average in a complex network. Other conditions must then be sought in order to identify the neighbors of the hidden nodes. Our idea is to exploit the stability of the predicted solution with respect to different measurements used for compressive sensing. In particular, for the neighboring nodes of the hidden node, due to the lack of information needed to solve the underlying compressive-sensing problem, when different segments of the time series are used, the algorithm will yield different coefficient vectors $\mathbf{a}$. However, for a node not in the immediate neighborhood of the hidden node, the predicted vector $\mathbf{a}$ should be the same for different data segments, for the corresponding coefficients with the hidden node are zero. As shown in Fig. 1(d), the variances in $\mathbf{a}$ of nodes No. 9 and No. 18 from a number of predictions are much larger than

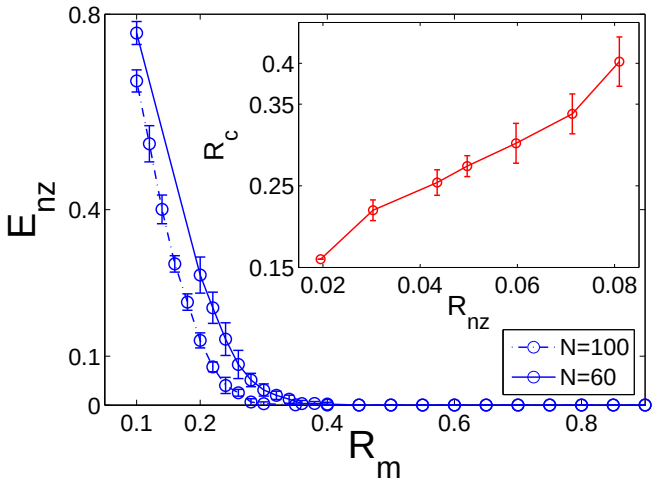

FIG. 2: (Color online.) For directed, weighted random networks of 60 nodes and 100 nodes, prediction error E_{nz} as a function of the ratio R_m. The ratio R_c as a function of the ratio $R_{nz} \equiv N_{nz}/(N_{nz} + N_z)$ is shown in the inset. The average connecting probability of the network is $p = 0.04$, and the link weights are uniformly distributed between 1 and 6. The error bars are calculated from 20 independent network realizations.

those (essentially zero) of other nodes. Violation of sparsity in combination with the instability of the predicted solution then allows us to identify all neighbors of the hidden node, and consequently itself, with high confidence.

To systematically characterize the accuracy and efficiency of our method to detect hidden nodes, we calculate the prediction error of links of all nodes (except the hidden node and its neighbors) in terms of the amount of required data. For an individual node, the prediction error is defined as the ratio of the absolute difference between the true adjacency matrix elements of all links associated with this node and the predicted elements to the nonzero true element values. The average over all nodes, excluding the neighbors of the hidden node, gives the total prediction error E_{nz}. To explore the effect of network size, we study networks of systematically varying size, ranging from 20 to 200 nodes. Figure 2 shows, for networks of 60 nodes and 100 nodes, E_{nz} as functions of the required data, which are the number of measurements N_m normalized by the number of terms $N_{nz} + N_z$ in the unknown vector $\mathbf{a}$. We see that, for the network of 60 nodes, when the measurement ratio exceeds 0.4, E_{nz} is close to zero, demonstrating that 40% data is sufficient to reconstruct the links and detect the location of the hidden node. For the network of 100 nodes, the data requirement is slightly smaller because the unknown coefficient vector is sparser. To further explore the relation between the data requirement and $N_{nz}/(N_{nz} + N_z)$, the sparsity measure of the vector $\mathbf{a}$ to be predicted, we define a threshold of normalized measurement R_c required for full reconstruction of the network dynamical system when the error E_{nz} is less than 10^{-2}. The sparsity measure can actually be adjusted by varying the network size while keeping the average node degree unchanged. As shown in the inset of Fig. 2, we observe that,

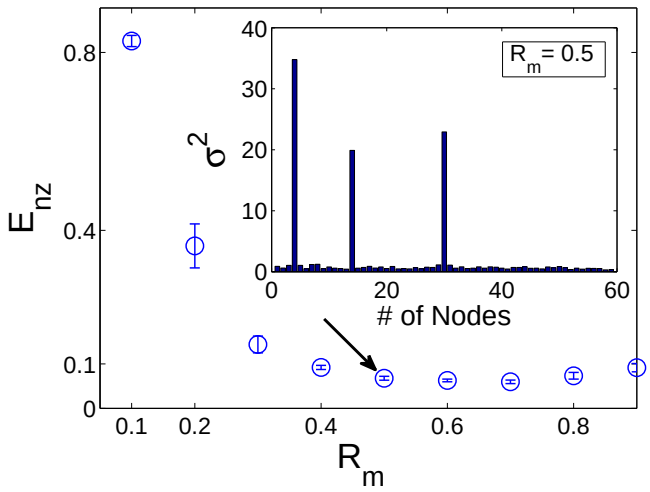

FIG. 3: (Color online.) For a random network of 60 nodes in the presence of noise, prediction error E_{nz} as a function of normalized measurements R_m, after excluding neighbors of the hidden node. Inset shows, for $R_m = 0.5$ as the arrow indicates, the variances of the coefficient vectors for all the nodes. There is only one hidden node in all cases and its neighborhoods are node No. 4, No. 14 and No. 30, which correspond to the tall bars. Uniform noise of amplitude 1% is added to the payoff vector and the measurement matrix.

as $\mathbf{a}$ becomes more sparse, the measurement threshold R_c is reduced accordingly. This also demonstrate the efficiency of our method for different network scales. These results illustrate the power of our compressive-sensing based method to locate hidden nodes with low data requirement.

We now address the effects of noise. As shown in Fig. 3, for a network of 60 nodes, the prediction errors decrease with the amount of the measurement data, with relative error of about 10% in the weights of the existing links. In this case, the links for all nodes except the neighborhoods of the hidden node are still predictable. The variances of the predicted vectors, as shown in the inset of Fig. 3, are larger compared with those in noiseless situation, but the neighborhoods of hidden node still have significantly larger variances than the others, indicating that the hidden node can still be detected reliably when the noise amplitude is weak as compared with the coupling strength of the hidden node. It is also possible to distinguish the effects due to noise and hidden node. The idea is that, when a hidden node is present, its influences on other nodes in the network are distinct, while the effect of noise on different nodes is statistically uniform and independent.

While we have demonstrated the principle of detecting hidden nodes using the setting of evolutionary-game dynamics, our formulation is general and applies to other types of network dynamics. For example, we have applied our method to detecting hidden node in networks with continuous-time oscillatory nodal dynamics by expanding $\tilde{\mathbf{g}}_i$ and $\mathbf{g}_{ij}$ into power series and obtaining a similar matrix $\mathbf{G}$, where the system response is the derivatives of the corresponding variables [6]. The unknown coefficients vectors $\mathbf{a}$ can then be solved, giving rise to full knowledge about the nodal and coupling dynamics. By examining the variances in $\mathbf{a}$, we can confirm and locate precisely the location of the hidden node in the network. We have applied our method to both continuous- and discrete-time oscillatory dynamics. Extensive numerical tests indicate that the method is robust with respect to different complex-network structures such as random, scale-free and clustered topologies, and large variations in the network size as well.

In summary, we have developed a completely data-driven approach to detecting hidden nodes in complex networks, which are inaccessible to external observation or measurement. The basic idea is to locate the immediate neighbors of the hidden node through reconstruction of the dynamical processes on these nodes that generate the time series or data. Because of their direct links with the hidden node, information used for the reconstruction is incomplete, leading to anomalies and instabilities in the prediction of their dynamics, which can then be used to infer that they are in the immediate neighborhood of the hidden node. Our reconstruction process is based on compressive sensing. Detecting hidden or black-boxed objects is an extremely challenging but fascinating task in science, and our work opens an avenue to addressing this problem in complex network science and engineering.

This work was supported by AFOSR under Grant No. FA9550-10-1-0083, and by NSF under Grants No. CDI-1026710 and No. BECS-1023101.

[1] M. Timme, Phys. Rev. Lett. **98**, 224101 (2007); S. G. Shandilya and M. Timme M, New J. Phys.**13**, 013004 (2011).
[2] D. Napoletani and T. D. Sauer, Phys. Rev. E **77**, 026103 (2008).
[3] W.-X. Wang, Q.-F. Chen, L. Huang, Y.-C. Lai, and M. A. F. Harrison, Phys. Rev. E **80**, 016116 (2009); J. Ren, W.-X. Wang, B. Li, and Y.-C. Lai, Phys. Rev. Lett. **104**, 058701 (2010).
[4] Z. Levnajić and A. Pikovsky, Phys. Rev. Lett. **107**, 034101 (2011).
[5] S. Hempel, A. Koseska, J. Kurths, and Z. Nikoloski, Phys. Rev. Lett. **107**, 054101 (2011).
[6] W.-X. Wang, R. Yang, Y.-C. Lai, V. Kovanis, and C. Grebogi, Phys. Rev. Lett. **106**, 154101 (2011); W.-X. Wang, R. Yang, Y.-C. Lai, V. Kovanis, and M. A. F. Harrison, EPL **94**, 48006 (2011); W.-X. Wang, Y.-C. Lai, C. Grebogi, and J.-P. Ye, Phys. Rev. X **1**, 021021 (2011).
[7] E. Candeès, J. Romberg, and T. Tao, IEEE Trans. Inf. Theory **52**, 489 (2006); Commun. Pure Appl. Math. **59**, 1207 (2006); D. Donoho, IEEE Trans. Inf. Theory **52**, 1289 (2006); R. G. Baraniuk, IEEE Signal Process. Mag. **24**, 118 (2007); E. Candeès and M. Wakin, IEEE Signal Process. Mag. **25**, 21 (2008).
[8] M. A. Nowak, *Evolutionary Dynamics: Exploring the Equations of Life* (Harvard Univ Press, Cambridge, MA, 2006).
[9] G. Szabó and G. Fath, Phys. Rep. **446**, 97 (2007).