Universal data-based method for reconstructing complex networks with binary-state dynamics
Jingwen Li, Zhesi Shen, Wen-Xu Wang, Celso Grebogi, and Ying-Cheng Lai
Phys. Rev. E 95, 032303 — Published 2 March 2017
DOI: 10.1103/PhysRevE.95.032303
A universal data based method for reconstructing complex networks with binary-state dynamics

Jingwen Li, Zhesi Shen, Wen-Xu Wang, Celso Grebogi, and Ying-Cheng Lai

1 School of Systems Science, Beijing Normal University, Beijing, 100875, China
2 Business School, University of Shanghai for Science and Technology, Shanghai 200093, China
3 Institute for Complex Systems and Mathematical Biology, King’s College, University of Aberdeen, Aberdeen AB24 3UE, UK
4 School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, Arizona 85287, USA
5 Department of Physics, Arizona State University, Tempe, Arizona 85287, USA

To understand, predict, and control complex networked systems, a prerequisite is to reconstruct the network structure from observable data. Despite recent progress in network reconstruction, binary-state dynamics that are ubiquitous in nature, technology, and society still present an outstanding challenge in this field. Here we offer a framework for reconstructing complex networks with binary-state dynamics by developing a universal data-based linearization approach that is applicable to systems with linear, nonlinear, discontinuous, or stochastic dynamics governed by monotonic functions. The linearization procedure enables us to convert the network reconstruction into a sparse signal reconstruction problem that can be resolved through convex optimization. We demonstrate generally high reconstruction accuracy for a number of complex networks associated with distinct binary-state dynamics from using binary data contaminated by noise and missing data. Our framework is completely data driven, efficient and robust, and does not require any a priori knowledge about the detailed dynamical process on the network. The framework represents a general paradigm for reconstructing, understanding, and exploiting complex networked systems with binary-state dynamics.

I. INTRODUCTION

Complex networked systems consisting of units with binary-state dynamics are common in nature, technology, and society [1]. In such a system, each unit can be in one of the two possible states, e.g., being active or inactive in neuronal and gene regulatory networks [2], cooperation or defection in networks hosting evolutionary game dynamics [3], being susceptible or infected in epidemic spreading on social and technological networks [4], two competing opinions in social communities [5], etc. The interactions among the units are complex and a state change can be triggered either deterministically (e.g., depending on the states of their neighbors) or randomly. Indeed, deterministic and stochastic state changes can account for a variety of emergent phenomena, such as the outbreak of epidemic spreading [6], cooperation among selfish individuals [7], oscillations in biological systems [8], power blackout [9], financial crisis [10], and phase transitions in natural systems [11]. A variety of models have been introduced to gain insights into binary-state dynamics on complex networks [12], such as the voter models for competition of two opinions [13], stochastic propagation models for epidemic spreading [14], models of rumor diffusion and adoption of new technologies [15], cascading failure models [16], Ising spin models for ferromagnetic phase transition [17], and evolutionary games for cooperation and altruism [18]. A general theoretical approach to dealing with networks hosting binary state dynamics was developed recently [19] based on pair approximation and master equations, providing a good understanding of the effect of the network structure on the emergent phenomena.

In this paper, we address the inverse problem of binary-state dynamics on complex networks, i.e., the problem of reconstructing the network structure and binary dynamics from data. Deciphering the network structure from data has always been a fundamental problem in complexity science, as the structure can determine the type of collective dynamics on the network [20]. More generally, for a complex networked system, reductionism is not effective and it is necessary to reconstruct and study the system as a whole [21]. The importance of network reconstruction has been increasingly recognized and effective methodologies have been developed [22–34]. Of particular relevance to our work is spreading dynamics on complex networks, where the available data are binary: a node is either infected or healthy. In such cases, a recent work [33] demonstrated that the propagation network structure can be reconstructed and the sources of spreading can be detected by exploiting compressive sensing [35–40]. However, for binary state network dynamics, a general reconstruction framework was lacking (prior to the present work). The problem of
reconstructing complex networks with binary-state dynamics is extremely challenging, for the following reasons. (i) The switching probability of a node depends on the states of its neighbors according to a variety of functions for different systems, which can be linear, nonlinear, piecewise, or stochastic. If the function that governs the switching probability is unknown, a tremendous difficulty would arise in obtaining a solution of the reconstruction problem. (ii) Structural information is often hidden in the binary states of the nodes in an unknown manner and the dimension of the solution space can be extremely high, rendering impractical (computationally prohibitive) brute-force enumeration of all possible network configurations. (iii) The presence of measurement noise, missing data, and stochastic effects in the switching probability make the reconstruction task even more challenging, calling for the development of effective methods that are robust against internal and external random effects.

To meet the challenges, we develop a general and robust framework for reconstructing complex networks based solely on the binary states of the nodes without any knowledge about the switching functions. Our idea is centered around developing a general method to linearize the switching functions from binary data. The data-based linearization method is applicable to linear, nonlinear, piecewise, or stochastic switching functions. The method allows us to convert the network reconstruction problem into a sparse signal reconstruction problem for local structures associated with each node. Exploiting the natural sparsity of complex networks, we employ the lasso [41], an L1 constrained fitting method for statistics and data mining, to identify the neighbors of each node in the network from sparse binary data contaminated by noise. We establish the underlying mechanism that justifies the linearization procedure by conducting tests using a number of linear, nonlinear and piecewise binary-state dynamics on a large number of model and real complex networks. We find universally high reconstruction accuracy even for small data amount with noise. Because of its high accuracy, efficiency and robustness against noise and missing data, our framework is promising as a general solution to the inverse problem of network reconstruction from binary-state time series, which is key to articulating effective strategies to control complex networks with binary state dynamics using, e.g., the recently developed network controllability frameworks [42–47]. The data-based linearization method is also useful for dealing with general nonlinear systems with a wide range of applications.

### II. BINARY-STATE DYNAMICS

We consider a large number of representative binary state processes on complex networks, which model a plethora of physical, social and biological phenomena [19]. In such a dynamical process, the state of a node can be 0 (inactive) or 1 (active). In general, the process can be characterized by two switching functions, $F(m, k)$ and $R(m, k)$, which determine the probabilities for a node to change its state from 0 to 1 and vice versa, respectively. The variables in these functions, $k$ and $m$, are the degree of the node and the number of active neighbors of the node, respectively. The switching functions can be linear, nonlinear, piecewise, bounded and stochastic for characterizing and generating all kinds of binary-state dynamical processes occurring on complex networks. Despite the difference among the switching functions, the feature that a node’s switching probability depends on its degree and its number of active neighbors is generic. Table I lists the switching functions of different models, and the brief descriptions of each model can be found in Appendix.

| Model        | $F(m, k)$                   | $R(m, k)$                   |
|--------------|-----------------------------|-----------------------------|
| Voter [13]   | $\frac{k-m}{m}$            | $\frac{k-m}{m}$            |
| Kirman [49]  | $c_1 + dm$                  | $c_2 + d(k - m)$            |
| Ising Glauber [17, 50] | $\frac{1}{1 + e^{rac{1}{k-2m}}}$ | $\frac{1}{1 + e^{rac{1}{k-2m}}}$ |
| SIS [14]     | $1 - (1 - \lambda)^m$     | $\mu$                      |
| Game [3]     | $\alpha e^{\frac{1}{m}[(a-c)(k-m)+(b-d)m]}$ | $(1 - s)\left(\frac{k-m}{m}\right)^\alpha$ |
| Language [51] | $s(\frac{m}{m})^\alpha$   | $\alpha e^{\frac{1}{m}[(a-c)(k-m)+(b-d)m]}$ |
| Threshold [52] | $0$ if $m \leq M_k$        | $0$                        |
| Majority vote [53] | $Q$ if $m < k/2$            | $1 - Q$ if $m < k/2$       |
|               | $1/2$ if $m = k/2$         | $1/2$ if $m = k/2$         |
|               | $1 - Q$ if $m > k/2$       | $Q$ if $m > k/2$           |

### III. RECONSTRUCTION METHOD

Our goal is to articulate a general framework to reconstruct the network structure from binary states of nodes without
knowing a priori the specific switching functions. A key step is to develop a universal procedure to obtain the linearization of the switching functions from binary data. We demonstrate that this can be accomplished by taking advantage of certain common features of the binary state dynamics.

A. Data based linearization of switching functions

To proceed, we note that the number of active neighbors at time \( t \) can be expressed as

\[
m_i(t) = \sum_{j=1,j\neq i}^N a_{ij}s_j(t),
\]

(1)

where \( a_{ij} = 1 \) if nodes \( i \) and \( j \) are connected and \( a_{ij} = 0 \) otherwise, and \( s_j(t) \) denotes the state of node \( j \) at time step \( t \). In general, the switching probability \( P_i^{01}(t) \) for node \( i \) to change its state from 0 to 1 at time step \( t \) can be written as

\[
P_i^{01}(t) = F\left(m_i(t), k_i\right) = F\left(\sum_{j=1,j\neq i}^N a_{ij}s_j(t), k_i\right),
\]

(2)

where \( F \) is a monotonous function characterizing different dynamical models, e.g., those listed in Table I. In Eq. (2), all the matrix elements \( a_{ij} \) \((i,j = 1, \ldots, N)\) that are to be inferred from data characterize the network structure. In general this is a difficult problem, because in Eq. (2), only nodal state \( s_j(t) \) is measurable, whereas neither of the quantites \( k_i \) and \( P_i^{01}(t) \) nor the form of \( F \) is known. In fact, not knowing the function \( F \) is the main difficulty in reconstructing the adjacency matrix \( \{a_{ij}\} \). To overcome this difficulty, we propose a merging process to linearize \( F \), i.e.,

\[
F \sim c_i \cdot \sum_{j=1,j\neq i}^N a_{ij}s_j(t) + d_i,
\]

(3)

where \( c_i \) and \( d_i \) are constants associated with node \( i \). Insofar as the linearization is realized, we can solve \( a_{ij} \). The idea of linearization is first proposed and used in Ref. [33], but the mathematical form of \( F \) is assumed to be known in that case. It is worth noting that the linearization approach is highly non-trivial and is fundamentally different from that in the standard canonical nonlinear analysis because, in our case, the mathematical form of \( F \) is not available, which can be a nonlinear, discrete and piecewise function.

B. Procedure of dealing with binary-state data

We present the procedure of dealing with binary-state data. The merging based linearization process enables the probability \( P_i^{01}(t) \) to be estimated according to the law of large numbers, from which the solution of \( a_{ij} \) can be obtained. In particular, as shown in Fig. 1(a), for an arbitrary node \( i \), we first identify all the time steps with \( s_i(t) = 0 \) as information about the switching probability \( P_i^{01}(t) \) is contained only in the flipping behavior from state 0. To accurately estimate the value of \( P_i^{01} \), we need to collect sufficient time strings which has the same state as each other. However, we almost cannot find enough such time strings for aggregation because of the dynamical stochasticity. Thus we relaxed the criterion to finding sufficient similar time strings. In each set of similar time strings, we first pick a base string \( s_{-i}(t) \), and then collect time strings similar to it. Then, the key process comes to selecting base strings that optimize the performance of reconstruction. Figure 1(b) shows our method of selecting the optimal base strings solely based on recorded data. Specifically, we first construct a network whose vertices represent strings composed of \( s_j(t)\)\((j \neq i)\) at different time steps when \( s_i(t) = 0 \), and edges are weighted by the normalized pairwise Hamming distances among the strings. Then, We eliminate edges whose weight is smaller than a threshold, say \( \Delta \). Setting another threshold \( \sigma \), we extract a subnetwork where only the top \( \sigma \) proportion of vertices with largest degree are preserved, while other vertices and their edges are removed. In this way, all remaining strings have a sufficient number of similar strings. Finally, we pick out \( M \) vertices with smallest degrees to ensure that the selected base strings are sufficiently different, where \( M \) is the number of equations in Eq. (16). The process of selecting base strings ensures us both good estimation for \( P_i^{01} \) and dissimilarity among the averaged neighborhood states. For each chosen base string, we use the threshold \( \Delta \) in the normalized Hamming distance between strings to select a set of subordinate strings that belong to each base string, as shown in Fig. 1(c). A subordinate string is a string whose normalized Hamming distance to the base string is less than the selected threshold \( \Delta \). Using the average of \( s_j(t) \) to represent the state of node \( j \) and the average of \( s_i(t+1) \) to estimate the switching probability \( P_i^{01}(t) \) of node \( i \) according to the law of large numbers, we obtain \( P_i^{01}(t) \approx \langle s_i(\hat{t} + 1) \rangle \), where \( \hat{t} \) denotes the time of the base string (see Fig. 1(d)).

The whole process leads to the linearization of \( F \) with the following data-based relationship

\[
\langle s_i(\hat{t} + 1) \rangle \approx c_i \cdot \sum_{j=1,j\neq i}^N a_{ij} \langle s_j(\hat{t}) \rangle + d_i,
\]

(4)

where \( \langle \cdot \rangle \) is the average over all time \( t \) of the subordinate strings within \( \hat{t} \). The constant parameter \( k_i \) is incorporated into the linear coefficient \( c_i \) and the intercept \( d_i \). It is not
necessary to estimate the quantities $c_i$, $a_{ij}$ and $d_i$ in Eq. (4) separately - it is only necessary to infer value of the product $c_i \times a_{ij}$. In particular, if $i$ and $j$ are not connected, we have $c_i \times a_{ij} = 0$, but a nonzero value of $c_i \times a_{ij}$ means that there is a link between the two nodes. As we will show, the value of $d_i$ can be obtained but this quantity plays little role in the reconstruction.

Figure 2 shows some representative examples to validate the linearization procedure. Four types of dynamics, including two with continuous and nonlinear switching functions and two with discontinuous and piecewise functions, are tested. We see that the switching functions $F$ for different parameter values are linearized, enabling the network structure in the linearized system (4) to be reconstructed by distinguishing between zero and nonzero values of the reconstructed product $c_i \times a_{ij}$. As compared to the original function $F$, the range of $m$ in the linearized function typically shrinks considerably as a result of the merging process, as shown in Figs. 2(a) and 2(b). For the discrete piecewise functions in Figs. 2(c) and 2(d), approximately linear functions arise for different parameter values. This is particularly striking, because even given a switching function, it is still difficult to linearize a piecewise function. We have achieve a data-based linearization of nonlinear and piecewise functions without any knowledge a priori.

C. Theoretical validation of data-based linearization

We provide an analysis for the completely data-based linearization that gives rise to the general relationship (Eq. (4)) from general binary-state dynamics characterized by the switching probability (Eq. (2)).

For nodes with only one neighbor, the linear relationship (4) can be rigorously proved. In this scenario, the number of active neighbors is either 0 or 1. Let $P_1$ denote the proportion of strings with single active neighbors in the set of base $t$, and denote the proportion of strings with null active neighbors as $1 - P_1$. Let the switching probability of null active neighbors and single active neighbors be $f(0)$ and $f(1)$. Then we
FIG. 2. Data based linearization procedure for nonlinear and piecewise binary-state dynamics. Linearization of the switching probability function \( F(m, k) \) for (a) Ising model, (b) evolutionary game model, (c) threshold model, and (d) majority model. The grey lines represent Eq. (2) with the function \( F(m, k) \) from the different models, where \( k \) is the node’s degree and \( m \) is the number of active neighbors. Data points are the result of the linearization procedure from time series, which corresponds to Eq. (4). For the linearized function, \( m \) is obtained from 
\[
\sum_{j=1, j \neq i}^{N} a_{ij} \langle s_j(\hat{t}) \rangle
\]
and the value of the function is obtained via 
\[
\langle s_i(\hat{t}+1) \rangle
\]
for the evolutionary game model and the threshold model. We see explicitly linear relationship for both models. With respect to different number of active neighbors in the original bases, two sets of groups are classified.

For nodes with more than one neighbor, the linear relationship can be justified and predicted based on binomial distribution and Taylor linear approximation. For an arbitrary node, say, node \( i \) with \( k \) neighbors, we will substantiate the linear relationship between \( \langle s_i(\hat{t}+1) \rangle \) and 
\[
\sum_{j=1, j \neq i}^{N} a_{ij} \langle s_j(\hat{t}) \rangle
\]
resulting from the data-based linearization, where
\[
\langle s_i(\hat{t}+1) \rangle \approx \langle P_i^{01}(t) \rangle = \sum_{m=0}^{k_i} F(m, k_i) P_i(m),
\]
and
\[
\sum_{j=1, j \neq i}^{N} a_{ij} \langle s_j(\hat{t}) \rangle = \sum_{m=0}^{k_i} m P_i(m).
\]
FIG. 3. Theoretical analysis of the data-based linearization. (a, b) Linearization of switching function for nodes with a single neighbor for the game model (a) and the threshold model (b). The grey solid curves are the original switching functions, data points are the results of data-based linearization (Eq. (4)), and the dashed lines are theoretical predictions from Eq. (7). The color of data points represents two sets of subordinate strings whose base string has no active neighbors \((m = 0)\) or has a single active neighbors \((m = 1)\). For both nonlinear and piecewise switching functions, the theoretical predictions are in exact agreement with data-based linearization, because for \(k_i = 1\) the linearization is rigorous without any approximation. (c, d) The distribution of active neighbors \(m\) in subordinate strings subject to each base string and binomial distributions for reconstructing node \(i\) with \(k_i = 3\) for the game model (c) and \(k_i = 6\) for the majority model (d), respectively. Each color of curves represents a set of subordinate strings whose base string has \(m\) active neighbors. The distribution can be well described by binomial distributions under different success probability in each trial, as exemplified by black curves. There is a good agreement between the distribution of active neighbors in subordinate strings and binomial distributions. (e, f) The original switching function and the linearized function with theoretical prediction based on binomial distribution for the game model (e) and the majority model (f), respectively. The color of data points represents different sets of subordinate strings whose base string has different number of active neighbors \(m\) (the same meaning as in (c), (d)). The grey curves are the original switching function in the binary-state dynamics. The black dashed lines are the theoretical prediction of the linear relationship through Eq. (15) based on binomial distribution and Taylor linear approximation. The theoretical predictions are in good agreement with numerical results.

where \(P_i(m)\) represents the proportion of strings with \(m\) active neighbors among all strings that belong to the set of base \(i\). The key to validating the linear relationship lies in the distribution that \(P_i(m)\) obeys.

Regarding the effect of the merging process as shown in Fig. 1, we hypothesize that \(P_i(m)\) follows binomial distributions with different success probability \(p_i\). We denote the proportion of state 0 in data to be \(p_0\). If the strings are randomly chosen for each set of a base, \(P_i(m)\) exactly obeys binomial distribution with success probability \(p_0\). However, due to the process of selecting strings that are similar to each set of a base, the distribution will be biased toward the number of active neighbors in the base. Despite the original complex influence of the base and string selections based on Hamming distance, their effects can be simply regarded as selecting a group of strings with similar proportion of state 0 since we actually do not know which the node’s neighbors are. This process leads to the success probability that depends on the base string. Figs. 3(c, d) shows the comparison between the actual distribution of \(P_i(m)\) obtained from numerical simulations and the binomial distributions with different success probability in each trial in the game and majority model, where the success probability in each trial approximately range from 0.4 to 0.6 because \(p_0 \approx 0.5\) in the data. We see that \(P_i(m)\) can be well approximated by binomial distributions with different parameter values, which indeed validates our binomial distribution hypothesis.

Based on the binomial distribution hypothesis, we have

\[
P_i(m) = C^m_{k_i} p^m_i (1 - p_i)^{k_i - m}.
\]
Inserting Eq. (10) into Eq. (8) yields
\[
\langle s_i(\hat{t}+1) \rangle \approx \sum_{m=0}^{k_i} F(m, k_i) C_{k_i}^m \sum_{l=0}^{m} (-1)^{m-l} C_m F(\hat{t}, k_i) (1 - p_t)^{k_i - m} p_t^m
\]

\[= \sum_{m=0}^{k_i} C_{k_i}^m \sum_{l=0}^{m} \left[ (-1)^{m-l} C_m F(\hat{t}, k_i) \right] p_t^m. \tag{11} \]

The fact that \(p_t\) fluctuates around \(p_0\) allows us to apply the Taylor series expansion around \(p_0\) to Eq. (11), leading to
\[
\langle s_i(\hat{t}+1) \rangle \approx \sum_{m=0}^{k_i} C_{k_i}^m \sum_{l=0}^{m} \left[ (-1)^{m-l} C_m F(\hat{t}, k_i) \right] p_0^m \]
\[+ \sum_{m=0}^{k_i} C_{k_i}^m \sum_{l=0}^{m} \left[ (-1)^{m-l} C_m F(\hat{t}, k_i) \right] m p_0^{m-1} (p_t - p_0) + O(p_t - p_0). \tag{12} \]

Omitting the high-order term \(O(p_t - p_0)\), we have
\[
\langle s_i(\hat{t}+1) \rangle \approx \sum_{m=0}^{k_i} C_{k_i}^m \sum_{l=0}^{m} \left[ (-1)^{m-l} C_m F(\hat{t}, k_i) \right] (1 - m) p_0^m \]
\[+ \sum_{m=0}^{k_i} C_{k_i}^m \sum_{l=0}^{m} \left[ (-1)^{m-l} C_m F(\hat{t}, k_i) \right] m p_0^{m-1} p_t. \tag{13} \]

On the other hand, substitute Eq. (10) into Eq. (9) yields
\[
\sum_{j=1, j \neq i}^{N} a_{ij} \langle s_j(\hat{t}) \rangle = \sum_{m=0}^{k_i} m C_{k_i}^m p_t^m (1 - p_t)^{k_i - m} \]
\[= k_i p_t. \tag{14} \]

Combining Eq. (13) and Eq. (14), we have
\[
\langle s_i(\hat{t}+1) \rangle \approx \sum_{m=0}^{k_i} C_{k_i}^m \sum_{l=0}^{m} \left[ (-1)^{m-l} C_m F(\hat{t}, k_i) \right] (1 - m) p_0^m \]
\[+ \left\{ \frac{1}{k_i} \sum_{m=0}^{k_i} C_{k_i}^m \sum_{l=0}^{m} \left[ (-1)^{m-l} C_m F(\hat{t}, k_i) \right] m p_0^{m-1} \right\} \]
\[\times \sum_{j=1, j \neq i}^{N} a_{ij} \langle s_j(\hat{t}) \rangle. \tag{15} \]

Note that all variables in the first term on the right hand side of Eq. (15) are only determined by the binary-state dynamics and the node degree of \(i\). Hence, the first term corresponding to \(d_i\) is a constant with respect to node state \(s_i\). In analogy, all variables in the coefficient of the second term are determined by the binary-state dynamics and the node degree of \(i\) as well, indicating the coefficient is a constant corresponding to \(c_i\) in Eq. (4). Taken together, we theoretically justified that Eq. (15) is approximately a linear equation in the form of Eq. (4).

Figures 3(e, f) shows the relationship between \(\langle s_i(\hat{t}+1) \rangle\) and \(\sum_{j=1, j \neq i}^{N} a_{ij} \langle s_j(\hat{t}) \rangle\) (namely \(\langle m \rangle\)) of each set of bases and the linear relationship calculated by using Eq. (15) for the game model and the majority model with nonlinear and piece-wise switching dynamics. We see that the theoretical predictions are in good agreement with the results from the merging process for linearization, which strongly validates the data-based linearization for general binary-state dynamics.

It is noteworthy that the key to the success of the data-based linearization lies in selecting similar strings subject to a base and the average over each set of bases. The selection of similar strings accounts for the binomial distribution of active neighbors in a set, and different bases induces different success probability in each trial. Then the average of the binomial distributions leads to the relatively small range of \(\langle m \rangle\) compared to the original range in the switching function, allowing us to use Taylor linear approximation. Moreover, high-order terms in the Taylor series expansion contribute little to the binomial distribution, which justifies the low-order approximation. Based on the linear relationship, the reconstruction of local structure can be realized by employing the lasso without requiring the linear coefficients and intercept. In other words, the data-based linearization is general valid for arbitrary binary-state dynamics without any knowledge of the switching function.

### D. Reconstruction of local structure based on the lasso

The linear relationship, Eq. (4) allows us to ascertain the neighbors of any node \(i\) from \(M\) different values of the base time, e.g., \(\hat{t}_1, \cdots, \hat{t}_M\), and their subordinate times. In particular, with respect to \(\hat{t}_1, \cdots, \hat{t}_M\), Eq. (4) can be expressed in the matrix form \(Y_i = \Phi_i \times X_i\) as Eq. (16), where the vector \(X_i\) is to be solved for obtaining the neighbors of \(i\), and the vector \(Y_i\) and the matrix \(\Phi_i\) can be constructed entirely from binary time series without requiring any other information.
The natural sparsity of complex networks ensures that, on average, the number of neighbors for a node is much smaller than the network size \( N \), implying that \( X_i \) is typically sparse with most of its elements being zero and the number of nonzero elements is in fact the node degree \( k_i \) with \( k_i \ll N \). We can then exploit the sparsity to reconstruct \( X_i \) by employing the lasso [41], a convex optimization method for sparse signal reconstruction. The lasso incorporating an L1-norm and an error control term is efficient and robust, enabling a reliable reconstruction of the local network structure as represented by \( X_i \) from a small amount of data. In particular, the problem is to optimize 

\[
\min_{X_i} \left\{ \frac{1}{2M} \| \Phi_i X_i - Y_i \|^2_2 + \lambda \| X_i \|_1 \right\}
\]

where \( \| X_i \|_1 = \sum_{j=1,j \neq i}^{N} |x_{ij}| \) is the \( L_1 \) norm of \( X_i \) assuring the sparsity of the solution, and the least squares term \( \| \Phi_i X_i - Y_i \|^2_2 \) guarantees the robustness of the solution against noise in data. In Eq. (17), \( \lambda \) is a nonnegative regularization parameter that affects the reconstruction performance in terms of the sparsity of the network, which can be determined by a cross-validation method [62]. An advantage of using the lasso is that \( M \), i.e., the number of bases needed, can be much less than the length of \( X_i \). For each base of each node, the strings included can be collected and calculated from only one set of data sample in the time series, ensuring the sparse data requirement.

After the vector \( X_i \) has been reconstructed, the direct neighbors of node \( i \) are simply those associated with nonzero elements in \( X_i \). In the same manner, we can uncover the neighborhoods of all other nodes, so that the full structure of the network can be obtained by matching the neighbors of all nodes.

IV. RECONSTRUCTION PERFORMANCE

A. Measurement indices

To quantify the performance of our reconstruction method, we introduce two standard measurement indices, the Area Under the Receiver Operating Characteristic curve (A\text{UROC}) and the Area Under the Precision-Recall curve (A\text{UPR}). True positive rate (TPR, \( R_{TP} \)), false positive rate (FPR, \( R_{FP} \)), Precision(\( \Theta_{\text{precision}} \)) and Recall(\( \Theta_{\text{recall}} \)) that are used to calculate \( A_{\text{UROC}} \) and \( A_{\text{UPR}} \) are defined as follows:

\[
R_{TP}(l) = \frac{N_{TP}(l)}{N_{P}},
\]

where \( l \) is the cutoff in the edge list, \( N_{TP}(l) \) is the number of true positives in the top \( l \) predictions in the edge list, and \( N_{P} \) is the number of positives in the gold standard.

\[
R_{FP}(l) = \frac{N_{FP}(l)}{N_{N}},
\]

where \( N_{FP}(l) \) is the number of false positive in the top \( l \) predictions in the edge list, and \( N_{N} \) is the number of negatives in the gold standard.

\[
\Theta_{\text{precision}}(l) = \frac{N_{TP}(l)}{N_{TP}(l) + N_{FP}(l)} = \frac{N_{TP}(l)}{l},
\]

\[
\Theta_{\text{recall}}(l) = \frac{N_{TP}(l)}{N_{P}},
\]

where \( \Theta_{\text{recall}}(l) \), which is called sensitivity, is equivalent to \( R_{TP}(l) \). By varying \( l \) from 0 to \( N \), two sequences of points \( (R_{TP}(l), R_{FP}(l)) \) and \( (\Theta_{\text{recall}}(l), \Theta_{\text{precision}}(l)) \) are measured respectively, and the receiver operating characteristic curve and the precision-recall curve are obtained, as shown in Fig. 3(d) and (f). The area under the two curves, denoted as \( A_{\text{UROC}} \) and \( A_{\text{UPR}} \) respectively, represent the reconstruction performance: \( A_{\text{UROC}}(A_{\text{UPR}}) \) ranges from \( A_{\text{UROC}} = 0.5(A_{\text{UPR}} = N_{P}/2N) \) for random guessing to \( A_{\text{UROC}} = 1(A_{\text{UPR}} = 1) \) for perfect reconstructibility.
Because the links of each node are actually identified separately, the AUROC and AUPR are calculated for each node, and we use the mean index values over all the nodes to characterize the reconstruction performance for the whole network.

B. Reconstruction performance affected by network structure and amount of data

We test our method by implementing different dynamical processes on Erdős-Rényi random (ER) [54](circle), scale-free (SF) [55](square), small-world (SW) [56](diamond), and empirical networks. For network reconstruction, knowledge about the switching dynamics and network details is not necessary – only the states of the nodes at different time steps need to be recorded. See Sec. 1 in Supplementary Materials [48] for computational details.

Figure 4 illustrates the reconstruction performance, where Fig. 4(a) shows the element values $x_{ij}$ in the reconstructed neighboring vector $X_i$ of all nodes for SW and SF networks with the voter model. $n_t$ is defined as the number of used base strings normalized by network size $N$. We note that the values of $x_{ij}$ corresponding to actual links are markedly and distinctly greater than those of null connections. Setting a cut-off value in the gap between the two groups of points in Fig. 4(a), we can separate the actual links from the null connections, enabling a reconstruction of the whole SW network. For the SF network, it is difficult to fully reconstruct the neighbors of the hub nodes, for the following two reasons: (i) in general the linearization procedure works better for small node degree, as shown in Fig. 2; (ii) the lasso based reconstruction requires smaller data amount and offers better accuracy for sparser vector $X_i$ associated with small degree nodes. However, for an SF network, a vast majority of the nodes in an SF network are not hubs, which can be precisely reconstructed. The reconstructed SW and SF networks are shown in Figs. 4(b) and 4(c), respectively.

To assess how the number of base strings $t$ affects the reconstruction accuracy, we define $n_t$ to be the number of $t$ divided by the network size $N$ to quantify the relative amount of the base strings. As shown in Figs. 4(d-g), the receiver operating characteristic (ROC) and the precision-recall (PR) curves show better performance as $n_t$ is increased for both SW and SF networks, implying that high accuracy can be achieved for reasonably large values of $n_t$. Fig. 5 shows the AUROC and AUPR measures as a function of $n_t$ for different dynamical models on ER, SW and SF networks. Due to the advantage of the lasso for sparse vectors, nearly perfect reconstruction is achieved after $n_t$ exceeds a relatively small critical value, e.g., 0.4.

It is also important to assess how the length of the binary time series affects the reconstruction accuracy and efficiency. We have calculated the AUROC and AUPR measures as a function of the relative time-series length $n_t$ (defined as the total length of time series divided by $N$) for various dynamical processes on ER, SF and SW networks. Figure 6 shows the reconstruction performance for Voter, Ising and Majority models in combination with different types of networks. We find that AUROC and AUPR rapidly increases as $n_t$ increases. After $n_t$ exceeds a relatively small value, nearly full reconstruction can be achieved, which provides additional evidence for the high efficiency of our reconstruction method (see Supplementary Material [48] Sec. 2 for full results of performance for all models versus $n_t$). In general, high reconstruction accuracy can be achieved for relatively short time series. We systematically test our method on a variety of model and real networks in combination with eight binary-state dynamics (Table II) and find high values of AUROC and AUPR for all cases.

We explore the effects of network properties such as the average degree $\langle k \rangle$ and the size $N$ on reconstruction performance. As shown in Fig. 7. The reconstruction accuracy decreases as $\langle k \rangle$ increases. The main reason for this result is that the low-order approximation in the data-based linearization is better for smaller node degree. Moreover, with the increase of $\langle k \rangle$, the vector $X_i$ to be reconstructed will become denser. Note that it usually requires larger amounts of data to reconstruct a denser signal by using the lasso according to the compressive sensing theory. Thus, in general a network with larger $\langle k \rangle$ will be more difficult to be reconstructed. Fig. 8 shows the minimum relative length of time series $n_t^{\min}$ to acquire at least 0.95 AUROC and AUPR simultaneously as a function of network size $N$. We see that $n_t^{\min}$ decreases as $N$ increases, which is because of network sparsity as well. In general, for the same average node degree $\langle k \rangle$, a network with larger size will be sparser, leading to a sparser vector $X_i$. According to the compressive sensing theory, less data are required for reconstructing a sparser $X_i$, accounting for the decrease of $n_t^{\min}$ with the increase of $N$. These results indicate that our reconstruction method is scalable and of practical importance for dealing with large real networked systems.

C. Robustness of reconstruction against noise and missing data

In real applications, time series are often contaminated by noise and the data from certain nodes may be lost or inaccessible. To address these practical issues, we test the robustness of
by randomly flipping a fraction \( n_f \) of binary states and assume a fraction \( n_m \) of nodes are inaccessible. The results are shown in Table III, where voter, game, and majority models are used as examples of linear, nonlinear and piecewise dynamics, respectively. Strikingly, we obtain high values of \( A_{UROC} \) and \( A_{UPR} \) even in presence of 10% measurement noise or 30% inaccessible nodes, providing strong evidence for the robustness of our framework against noise and missing data. More detailed characterization associated with the results in Table III, i.e., \( A_{UROC} \) and \( A_{UPR} \) as functions of \( n_f \) and \( n_m \), are provided in Supplementary [48] Sec. 3.

TABLE II. \( A_{UROC} \) and \( A_{UPR} \) measures for various dynamics on a variety of model and empirical networks. The parameter values in the dynamical models are listed in Supplementary Table S1. The size and mean degree of ER(circle), SF(square) and SW(diamond) networks are \( N = 500 \) and \( \langle k \rangle = 6 \), and the length of time series used is \( 6 \times 10^4 \). The length of time series used for empirical networks is \( 1.5 \times 10^4 \).

\[
\begin{array}{cccccccc}
\text{AUCROC/AUPR} & \text{Voter} & \text{Kiran} & \text{Ising} & \text{SIS} & \text{Game} & \text{Language} & \text{Threshold} & \text{Majority} \\
\hline
\text{ER} & 1.000/0.983 & 0.999/0.954 & 1.000/0.982 & 0.997/0.960 & 0.999/0.981 & 0.995/0.934 & 1.000/0.988 & 1.000/0.986 \\
\text{SF} & 0.992/0.959 & 0.985/0.920 & 0.998/0.976 & 0.984/0.924 & 0.988/0.951 & 0.986/0.925 & 0.986/0.985 & 0.999/0.980 \\
\text{SW} & 1.000/0.988 & 1.000/0.982 & 1.000/0.988 & 1.000/0.988 & 1.000/0.988 & 1.000/0.986 & 0.994/0.979 & 1.000/0.987 \\
\text{Dolphins} & 1.000/0.916 & 0.997/0.908 & 0.999/0.911 & 0.978/0.867 & 0.993/0.900 & 0.985/0.870 & 0.991/0.890 & 1.000/0.913 \\
\text{Football} & 0.999/0.884 & 1.000/0.898 & 0.999/0.899 & 0.999/0.884 & 0.996/0.882 & 0.992/0.859 & 0.918/0.637 & 0.999/0.896 \\
\text{Karate} & 0.997/0.856 & 0.969/0.838 & 0.981/0.836 & 0.954/0.823 & 0.984/0.839 & 0.960/0.803 & 0.971/0.810 & 0.966/0.847 \\
\text{Leader} & 1.000/0.838 & 0.991/0.912 & 0.991/0.823 & 0.968/0.789 & 0.990/0.818 & 0.966/0.780 & 0.970/0.760 & 0.998/0.832 \\
\text{Polbooks} & 0.999/0.912 & 0.991/0.829 & 0.998/0.908 & 0.932/0.779 & 0.986/0.888 & 0.978/0.857 & 0.971/0.858 & 0.999/0.913 \\
\text{Prison} & 1.000/0.936 & 0.999/0.896 & 1.000/0.935 & 0.992/0.915 & 0.981/0.909 & 0.991/0.909 & 0.999/0.931 & 1.000/0.935 \\
\text{Santa Fe} & 0.998/0.967 & 0.990/0.933 & 1.000/0.969 & 0.982/0.937 & 0.997/0.965 & 0.996/0.959 & 0.994/0.961 & 1.000/0.970 \\
\end{array}
\]
TABLE III. Robustness of reconstruction against noise and missing data. $\text{AUROC}$ and $\text{APR}$ measures for voter, game, and majority models on ER, SF and SW networks for measurement noise $n_f = 10\%$ and the fraction of inaccessible nodes $n_m = 30\%$. The network size is $N = 500$ and the mean degree is $\langle k \rangle = 6$. The length of the time series used is $6 \times 10^4$. Details of the parameter values in the dynamical models are listed in Supplementary Table S1.

|                | $n_f = 10\%$ | $n_m = 30\%$ |
|----------------|---------------|---------------|
| $N_{\text{ER}}$ | $0.9950.938$  | $1.0000.985$  |
| $N_{\text{SF}}$ | $0.9830.903$  | $0.9950.968$  |
| $N_{\text{SW}}$ | $1.0000.984$  | $1.0000.988$  |

FIG. 5. Reconstruction performance with respect to the number of base strings. (a,b,c) $\text{AUROC}$ and (d,e,f) $\text{APR}$ as functions of the normalized number of base strings $n_f$ for the voter, game and majority model on ER (circle), SF (square) and SW (diamond) networks. The network size $N = 100$ and $\langle k \rangle = 6$. The length of time series is $1.5 \times 10^4$. Other parameter values of binary-state dynamics are shown in Supplementary Table S1.

FIG. 6. Reconstruction performance with respect to the length of time series. (a-c) $\text{AUROC}$ and (d-f) $\text{APR}$ as functions of the relative length of time series $n_t$ for the voter, ising and majority model on ER(circle), SF(square) and SW(diamond) networks. The network size $N = 500$ and $\langle k \rangle = 6$. Other parameter values of binary-state dynamics are shown in Supplementary Table S1.

FIG. 7. Reconstruction performance affected by average node degree. (a,b,c) $\text{AUROC}$ and (d,e,f) $\text{APR}$ as functions of the average node degree $\langle k \rangle$ for the voter, game and majority model on ER(circle), SF(square) and SW(diamond) networks. The network size $N = 500$ and relative length of time series $n_t = 100$. Other parameter values of binary-state dynamics are shown in Supplementary Table S1.

FIG. 8. Reconstruction performance affected by network size. The minimum relative length $n_{\text{rel}}$ to acquire at least 0.95 $\text{AUROC}$ and $\text{APR}$ simultaneously as a function of network size $N$ for the voter(red circle), Ising(green square) and majority model(blue diamond) on (a) ER, (b) SF and (c) SW networks. The mean degree of networks is 6. Other parameter values of binary-state dynamics are shown in Supplementary Table S1.
V. DISCUSSION

Reconstructing the topological structure and dynamics of complex systems from data is a central issue in both network science and engineering community [22, 24, 25, 27, 28, 32, 58, 59]. A framework [29, 60, 61] of network reconstruction is based on compressive sensing [35–40], a sparse signal recovery method developed in applied mathematics and engineering signal processing. A recent work [33] also demonstrated that compressive sensing can be exploited for network reconstruction in situations where the available time series are polarized (binary), e.g., virus spreading and information diffusion in social and computer networks. While the structure of the virus propagation network and the spreading sources can be obtained, the method is unable to predict the network dynamical systems that generate the binary data.

The contribution of this paper is a general framework to solve the challenging problem of reconstructing complex networks hosting binary-state dynamics, based only on time series without any knowledge of the network structure and the switching functions that generate the binary data. The key to our success is the formulation of a universal data-based linearization method, which is powerful for reconstructing the neighborhood of nodes for any type of nodal dynamics: linear, nonlinear, discontinuous, or stochastic. The natural sparsity of real complex networks allows us to address the local reconstruction as a sparse signal reconstruction problem that can be solved by employing the lasso, a convex optimization method, from small amounts of binary data. The optimization is robust against measurement noise and missing data. Once the neighborhoods of all nodes have been reconstructed, the whole network can be mapped out by assembling all the local structures and making adjustments to ensure consistency. We have validated our framework using a variety of binary-state dynamical models on a number of model and real complex networks. High reconstruction accuracy has been obtained for all cases, even for relatively small amounts of binary data contaminated by noise and when partial data are lost. These results suggest the practical applicability of our framework. In practical applications, instead of evaluating $A_{UROC}$ and $A_{UPR}$, we often need to distinguish the true links from the non-links based on the reconstructed values of $X$. To accomplish this goal, we can generate a histogram from all the elements of $X$ and find a appropriate cutting threshold between the two peaks representing links and non-links.

While our framework potentially offers a general, completely data driven approach to reconstructing binary dynamical processes on complex networks, there are still challenges. For example, our framework can deal with various types of switching functions underlying the binary-state dynamics, but in its present form the framework is not applicable to non-monotonic functions or non-Markovian type of dynamics. Especially, when the switching functions are not monotonic, the data-based linearization would fail due to the violation of the one-to-one correspondence between the switching probability and the number of active neighbors. For non-Markovian dynamics, the merging procedure inherent in our method would fail. To predict the interaction strength among nodes presents another challenge, especially where noise is present and there is missing data. The results reported in this paper suggest strongly that our present framework can serve as a starting point to meet the challenges, eventually leading to a complete and universally applicable solution to the inverse problem of binary network structure and dynamics.

[1] Barrat, A., Barthelemy, M. & Vespignani, A. Dynamical Processes on Complex Networks (Cambridge University Press, 2008).
[2] Kumar, A., Rotter, S. & Aerts, A. Spiking activity propagation in neuronal networks: reconciling different perspectives on neural coding. Nature Rev. Neuro. 11, 615–627 (2010).
[3] Szabó, G. & Fath, G. Evolutionary games on graphs. Phys. Rep. 446, 97–216 (2007).
[4] Pastor-Satorras, R., Castellano, C., Van Mieghem, P. & Vespignani, A. Epidemic processes in complex networks. Rev. Mod. Phys. 87, 925–979 (2015).
[5] Shao, J., Havlin, S. & Stanley, H. E. Dynamic opinion model and invasion percolation. Phys. Rev. Lett. 103, 018701 (2009).
[6] Granell, C., Gómez, S. & Arenas, A. Dynamical interplay between awareness and epidemic spreading in multiplex networks. Phys. Rev. Lett. 111, 128701 (2013).
[7] Santos, F. C. & Pacheco, J. M. Scale-free networks provide a unifying framework for the emergence of cooperation. Phys. Rev. Lett. 95, 098104 (2005).
[8] Koseska, A., Volkov, E. & Kurths, J. Oscillation quenching mechanisms: Amplitude vs. oscillation death. Phys. Rep. 531, 173–199 (2013).
[9] Buldyrev, S. V., Parshani, R., Paul, G., Stanley, H. E. & Havlin, S. Catastrophic cascade of failures in interdependent networks. Nature 464, 1025–1028 (2010).
[10] Galbiati, M., Delpini, D. & Battiston, S. The power to control. Nature Phys. 9, 126–128 (2013).
[11] Balcan, D. & Vespignani, A. Phase transitions in contagion processes mediated by recurrent mobility patterns. Nature Phys. 7, 581–586 (2011).
neighbors at the rate parameter to the Glauber dynamics \[50\], as shown in Table I, where the ferromagnetic-interaction parameter. In our study, we chose the transition rates according to the states of the neighbors and the probability of misalignment is described by the payoff matrix \[
\begin{pmatrix}
  s_1 & a & b \\
  s_2 & c & d
\end{pmatrix}
\]
where \(a, b, c\) and \(d\) are parameters. Different games can be generated by adjusting \(a, b, c\) and \(b\). The payoff of a player is the sum of profit from playing game with all its neighbors. A player switches the strategy with a probability that depends on the payoff it may gain in the next round under the current circumstance by switching its strategy, as illustrated in Table I, where the parameter \(\alpha\) qualifies the willingness for an individual to change its strategy according to those of its neighbors, and \(\beta\) is associated with the effect of the expected payoff.

For the language model \[51\], the two states denote two different language choices of a person. Transition from the primary language to the secondary occurs with the probability that is proportional to the fraction of speakers in the neighbors with the power \(\alpha\), multiplied by the parameter \(s\) (or \(1 - s\)) according to the respective language.

The threshold model \[52\] is a deterministic model, where for each node a certain threshold \(M_k\) is set which can be, for example, a function of the node’s degree. At each time step, a node becomes active if the number \(m\) of its active neighbors exceeds the threshold \(M_k\), and no recovery transformation is permitted.

**ACKNOWLEDGEMENTS**

W.-X.W. was supported by NSFC under Grant No. 11105011, CNNSF under Grant No. 61074116 and the Fundamental Research Funds for the Central Universities. Y.-C.L. was supported by ARO under Grant W911NF-14-1-0504.

**AUTHOR CONTRIBUTIONS**

W.-X.W. designed research; J.L. and Z.S. performed research; all analyzed data; J.L., W.-X.W. and Y.-C.L. wrote the paper; all edited the paper.
ADDITIONAL INFORMATION

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