Reconstructing complex networks with binary-state dynamics

Jingwen Li,[1] Wen-Xu Wang,[1,2,*] Ying-Cheng Lai,[3,4] and Celso Grebogi[5]

[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]*School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, Arizona 85287, USA*
[4]*Department of Physics, Arizona State University, Tempe, Arizona 85287, USA*
[5]*Institute for Complex Systems and Mathematical Biology, Kings College, University of Aberdeen, Aberdeen AB24 3UE, UK*

Abstract

The prerequisite for our understanding of many complex networked systems lies in the reconstruction of network structure from measurable data. Although binary-state dynamics occurring in a broad class of complex networked systems in nature and society and has been intensively investigated, a general framework for reconstructing complex networks from binary states, the inverse problem, is lacking. Here we offer a general solution to the reconstruction problem by developing a data-based linearization approach for binary-state dynamics with linear, nonlinear, discrete and stochastic switching functions. The linearization allows us to convert the network reconstruction problem into a sparse signal reconstruction problem that can be resolved efficiently and credibly by convex optimization based on compressed sensing. The completely data-based linearization method and the sparse signal reconstruction constitutes a general framework for reconstructing complex networks without any knowledge of the binary-state dynamics occurring on them in an extremely efficient and robust manner. Our framework has been validated by several different kinds of binary-state dynamics in combination with a large number of artificial and real complex networks. A universal high reconstruction accuracy is achieved in spite of the measurement noise and missing data of partial nodes. Our approach opens a new route to the inverse problem in complex networked systems with binary-state dynamics and improves our ability to fully understand and control their emergent dynamics in a comprehensive way.

*Electronic address: wenxuwang@bnu.edu.cn

I. INTRODUCTION

Complex networked systems consisting of dynamical units with binary states are common in nature and society [1]. Each unit can be in one of two possible states, e.g., active or inactive in neuronal and gene regulatory networks [2, 3], cooperate or defect in networks associated with evolutionary games [4], susceptible or infected in epidemic spreading on social and technological networks [5, 6], two competing opinions in communities with social networks [7], etc. The interactions among units often exhibit complex structure and units switch their states in a stochastic manner that depends on the states of their neighbors, which jointly account for a variety of emergent phenomena, such as the outbreak of epidemic spreading [8], cooperation among selfish individuals [9], oscillation in many biological systems [10], big blackout and financial crisis [11, 12], and phase transitions in many scenarios [13].

A variety of models have been introduced to explore binary-state dynamics occurring on complex networks [14, 15]. Representative models include voter models for competition of two opinions [16], stochastic propagation models for epidemic spreading [5], rumor propagation and adoption of new technologies [17], cascading failure models for crisis events [11], Ising spin models for paramagnetic phase transition [18], and evolutionary games for cooperation and altruism [4]. At present, general approaches based on pair approximations and approximate master equations have been provided to theoretically investigated the binary-state dynamics and deepen our understanding of the effect of network structure on the emergent phenomena [19].

Our goal here is to address the inverse problem of binary-state dynamics on complex networks, i.e., network reconstruction based solely on binary states. This is a fundamental problem for exploring binary-state dynamics on complex networks, because networks play a deterministic role in many collective dynamics [20]. Much evidence has demonstrated that reductionism is no longer available for complex networked systems, raising the need for exploring a complex networked system as a whole rather than reducing it into independent components [21]. Network reconstruction is necessary for studying many systems in that a direct measure of interaction structure is often not applicable and alternatively, one has to rely on measurable data of binary states to infer network topology [22]. Although the importance of network reconstruction has been increasingly recognized and some effective approaches and tools have been developed [22–34], a general reconstruction framework for complex networks with binary-state dynamics is lacking. The task is extremely challenging because of the following reasons. (i) The switching probability of a node depends on the states of neighbors according to a variety of functions for different systems, including linear, nonlinear, piecewise and stochastic functions. In particular, if the function or the form that governs the switching probability is unknown, it will be very difficult to solve the reconstruction problem. (ii) Structural information is hidden in the binary states of nodes in an unknown manner and the solution space is extremely high, rendering brute-force enumeration of all possible network configuration impossible. (iii) The presence of measurement noise, missing data and the stochastic effect in the switching probability raises the need for developing a robust method against internal and external noisy effects.

To overcome these obstacles, we develop a generally available and robust framework for reconstructing complex networks based solely on binary states of nodes in the absence of the knowledge of switching function. The key to our success lies in the development a general data-based method

for linearizing switching functions from binary observation. The data-based linearization method is generally applicable to nonlinear and piecewise stochastic switching functions. The task of reconstructing the whole network is then decomposed into infer local structures centered at each nodes. By exploiting the natural sparsity of complex networks, we employ the Lasso [35] to identify neighbors of each nodes from sparse binary data contaminated by noise. We articulate the underlying mechanism that enables the linearization by applying our method to several typical linear, nonlinear and piecewise binary-state dynamics occurring on many model and real complex networks. We found a universal high reconstruction accuracy from using relative small amounts of measurement contaminated by noise. With respect to its extremely high accuracy, efficiency and robustness against noise and missing information, our approach goes much beyond conventional methods in information theory and statistic physics, such as transfer entropy and maximum likelihood estimation that are useful for inferring network structure to some extent. Our approach offers a promising prospect of generally solving the inverse problem of network reconstruction from binary-state time series. The solution for the inverse problem is of paramount importance in understanding the dynamical behaviors of a large number of complex networked systems in nature and society [36]. Consequently, effective control strategies may be devised to guide the dynamics towards desired states by combining recently developed theory for controlling complex networked systems [37–40].

II. RESULTS

Binary-state dynamics. We consider some representative binary-state models on complex networks for modeling many physical, social and biological dynamics [19]. The dynamics of these models are characterized by switching functions $F(m,k)$ and $R(m,k)$, where k is the number of neighbors and m is the number of active neighbors. The switching functions determine the probability of a node to flip from 0 to 1 and from 1 to 0, respectively. The switching functions could be in linear, nonlinear, piecewise, bounded and stochastic for characterizing many dynamical processes occurring on complex networks, constituting a broad classes of binary-state dynamics. Despite the difference among the switching functions, a common feature is that a node's switching probability depends on the number of its active neighbors m and the number of its neighbors (degree) k. The switching functions of different models are shown in Table I. The detailed description of the models is presented in Methods.

Data-based linearization of switching functions by a merging process. Our goal is to develop a general framework to reconstruct network structure from binary states of nodes without knowing the specific dynamical function. The key lies in the establishment of a universal data-based linearization of switching functions from the common feature of the binary-state dynamics. Specifically, the number of active neighbors at time t can be expressed as

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

where $a_{ij} = 1$ if node i and j are connected and $a_{ij} = 0$ otherwise, and $s_j(t)$ captures the state of node j in time step t and can be obtained directly from the nodal states. We can generally

TABLE I: **Switching functions of binary-state dynamic models on complex networks.** $F(m,k)$ is the probability that a node turns from 0 to 1 while $R(m,k)$ is the probability of a node flipping from 1 to 0. k is the node's degree. m is the number of the node's neighbors with state 1. The models and the other parameters are introduced in Methods. Values of the parameters used in simulations are listed in Supplementary Table I.

Model	$F(m,k)$	$R(m,k)$
Voter	$\frac{m}{k}$	$\frac{k-m}{k}$
Kirman	$c_1 + dm$	$c_2 + d(k-m)$
Ising Glauber	$\frac{1}{1+e^{\beta(k-2m)/k}}$	$\frac{e^{\beta(k-2m)/k}}{1+e^{\beta(k-2m)/k}}$
SIS	$1-(1-\lambda)^m$	μ
Game	$\frac{1}{\alpha+e^{\beta a(k-m)/k}}$	$\frac{1}{\alpha+e^{\beta bm/k}}$
Language	$s(\frac{m}{k})^\alpha$	$(1-s)(\frac{k-m}{k})^\alpha$
Threshold	$\begin{cases} 0 & \text{if } m \leqslant M_k \\ 1 & \text{if } m > M_k \end{cases}$	0
Majority vote	$\begin{cases} Q & \text{if } m < k/2 \\ 1/2 & \text{if } m = k/2 \\ 1-Q & \text{if } m > k/2 \end{cases}$	$\begin{cases} 1-Q & \text{if } m < k/2 \\ 1/2 & \text{if } m = k/2 \\ Q & \text{if } m > k/2 \end{cases}$

formulate the switching probability $P_i^{01}(t)$ of node i from 0 to 1 at time step t to be

$$P_i^{01}(t) = F\left(m_i(t), k_i\right) = F\left(\sum_{j=1, j\neq i}^{N} a_{ij} s_j(t), k_i\right), \tag{2}$$

where F is a general monotonous function and can characterize different dynamical models listed in Table I and beyond.

Note that in Eq. (2), a_{ij} captures the network structure and is to be inferred. However, it is an extremely challenging problem, because that in Eq. (2), only node state $s_j(t)$ is measurable, whereas the constant k_i, $P_i^{01}(t)$ and the form of F are all unknown. Here, the unknown of the function F leads to the main difficulty in the recovery of a_{ij}. Thus, we propose a data-based merging process to linearize F, i.e.,

$$F \sim c_i \cdot \sum_{j=1, j\neq i}^{N} a_{ij} s_j(t) + d_i, \tag{3}$$

where c_i and d_i are constants of node i. Insofar as the linearization is realized, it is possible to solve a_{ij}. It is worth of nothing that the linearization approach is highly nontrivial and is fundamentally different from that in canonical nonlinear science, because the mathematical formula of

F is unavailable and could be nonlinear, discrete and piecewise function. The completely data-based linearization is our main contribution to the network reconstruction problem. Accompanied by linearization through a merging process, the probability $P_i^{01}(t)$ is estimated as well according to the law of large numbers, enabling the solution of a_{ij} exclusively from binary time series.

In particular, as shown in Fig. 1(a), we first pick out all the time steps with $s_i(t) = 0$ because the switching probability $P_i^{01}(t)$ is only reflected in the flipping behavior starting from state 0. Then we choose proper base strings from these time steps to represent different states of the system (see Methods and Fig. 1b for detailed procedure). For each chosen base string, we set a threshold Δ of the normalized Hamming distance between strings to select a set of subordinate strings that belong to each base string (how to choose the threshold is detailed in Supplementary Information Section 2). Then we use 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. This yields $P_i^{01}(t) \approx \langle s_i(\hat{t}+1)\rangle$. The whole process (as schematically illustrated in Fig. 1) finally leads to the linearization of F and a linear relationship

$$\langle s_i(\hat{t}+1)\rangle \approx c_i \cdot \sum_{j=1, j\neq i}^{N} a_{ij}\langle s_j(\hat{t})\rangle + d_i, \tag{4}$$

where $\hat{t}$ denotes the time of base string and $\langle\cdot\rangle$ denotes the average over all time t of subordinate string subject to $\hat{t}$ of base string. The effect of constant k_i is incorporated into the linear coefficient c_i and intercept d_i. In the linear formula (4), it is not necessary to derive c_i, a_{ij} and d_i separately, but infer $c_i \times a_{ij}$ as a whole (if i and i are not connected, $c_i \times a_{ij} = 0$; otherwise, nonzero value of $c_i \times a_{ij}$ stands for a link). As we will show, d_i can be reproduced but does not useful for our reconstruction.

Figure 2 exhibits some representative examples to validate the linearization effect. Four dynamics, including two continuous and nonlinear switching functions, and two discontinuous and piecewise functions, are presented. We see that the switching functions F with different dynamic parameter values are indeed linearized, which allows us to reconstruct network structure in the linearized system (4) through distinguishing between zero and nonzero values of the reconstructed $c_i \times a_{ij}$. Compared to the original F, the range of m in the linearized function considerably shrinks induced by the merging process, which partially accounts for the general feasibility of the data-based linearization for the continuous nonlinear F, as shown in Fig. 2(a)(b). For the discrete piecewise function in Fig. 2(c),(d), a theoretical explanation of data-based linearization is provided in Supplementary Sec. I.

Reconstruction of local structure based on the Lasso. Equation (4) allows us to infer the neighbors of node i from M different base time, e.g., $\hat{t}_1, \cdots, \hat{t}_M$ and their subordinate times. Specifically, with respect to $\hat{t}_1, \cdots, \hat{t}_M$, Eq. (4) can be expressed in the matrix form $\mathbf{Y}_i = \Phi_i \times \mathbf{X}_i$ (see Methods for a detailed matrix form), where vector $\mathbf{Y}_i$ and matrix Φ_i can be constructed completely from binary time series without relying on any other information and vector $\mathbf{X}_i$ is to be solved for revealing neighbors of i. In particular, 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 $\mathbf{X}_i$ is sparse with a large fraction of null elements and the number of nonzero elements is actually the node degree k_i with $k_i \ll N$. We thus exploit the sparsity of $\mathbf{X}_i$ to reconstruct $\mathbf{X}_i$

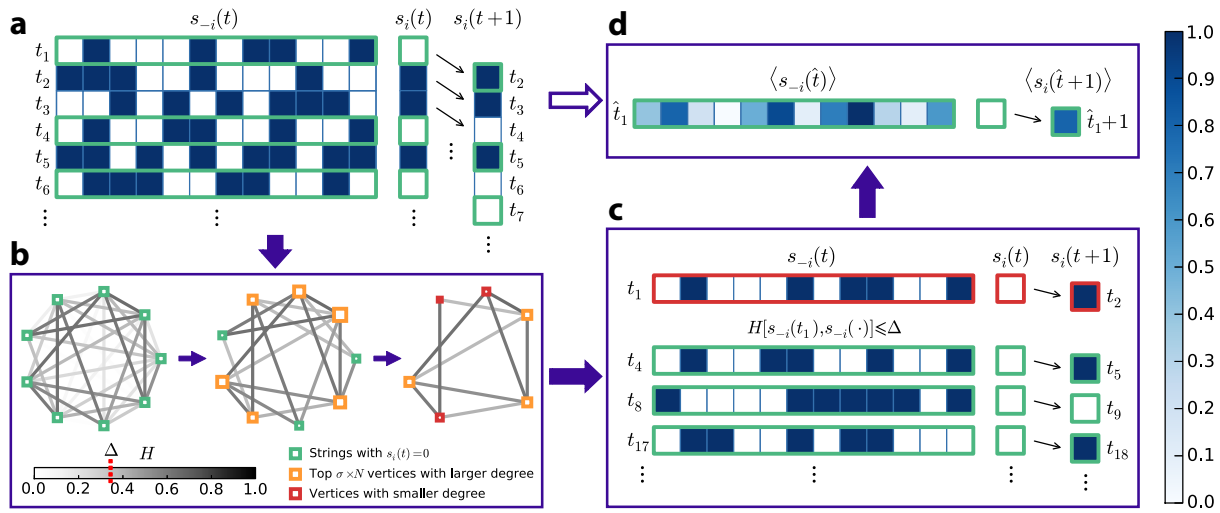

FIG. 1: **Schematic illustration of data-based linearization from a merging process.** (**a**) The original Binary-state time series. Dark blue square denotes 1 state and white square denotes 0 state. $s_{-i}(t)$ consists of $s_j(t)$ for all $j \neq i$. Only strings with $s_i(t) = 0$ as highlighted by green frames contain useful information for reconstruction. Thus, we pick out time step with $s_i(t) = 0$ and relevant $s_i(t+1)$. (**b**) Method of choosing bases. We first construct a network where vertices denote strings of $s_{-i}(t)$ with $s_i(t) = 0$ (green squares) and edges are weighted by normalized Hamming distance H between strings. We then eliminate edges whose weight is smaller than a threshold Δ. By setting another threshold σ, we select out the top $\sigma \times N$ vertices with larger degree (yellow squares), and remove the other vertices and their edges. Finally, we pick out the vertices with smaller degree (red squares) according to the number of base strings $n_{\hat{t}}$ needed for reconstruction. (**c**) Selection of subordinate strings subject to a based. We take t_1 as a base $\hat{t}_1$. We calculate H between $s_{-i}(t_1)$ and other strings $s_{-i}(t)$, and sort out time steps satisfying $H[s_{-i}(t_1), s_{-i}(\cdot)] < \Delta$ in this set. (**d**) Establishing average node states. We calculate the average value $\langle s_{-i}(\hat{t}) \rangle$ to represent the state of the base set, and the average value $\langle s_i(\hat{t}+1) \rangle$ to linearize the switching probability $P_i^{01}(t)$, see Eq. (4) and Eq. (5). The average values are in blue. In a similar fashion, we obtain a series of $\hat{t}_M$ and the associated average values for reconstructing network structure by employing the lasso to solve $\mathbf{Y}_i = \Phi_i \times \mathbf{X}_i$ (see Methods for details).

by employing the Lasso [35], a convex optimization method for sparse signal reconstruction. The Lasso by incorporating an L1-norm and an error control term enables a reliable reconstruction of $\mathbf{X}_i$ from small amounts of data, giving rise to efficient and robust reconstruction of local structures. (more details of the Lasso are presented in Methods). In a similar fashion, we can reconstruct the local structure of all nodes from the same set of data measurement, accounting for the sparse data requirement. The whole network can be recovered by simply assembling all local structures of nodes.

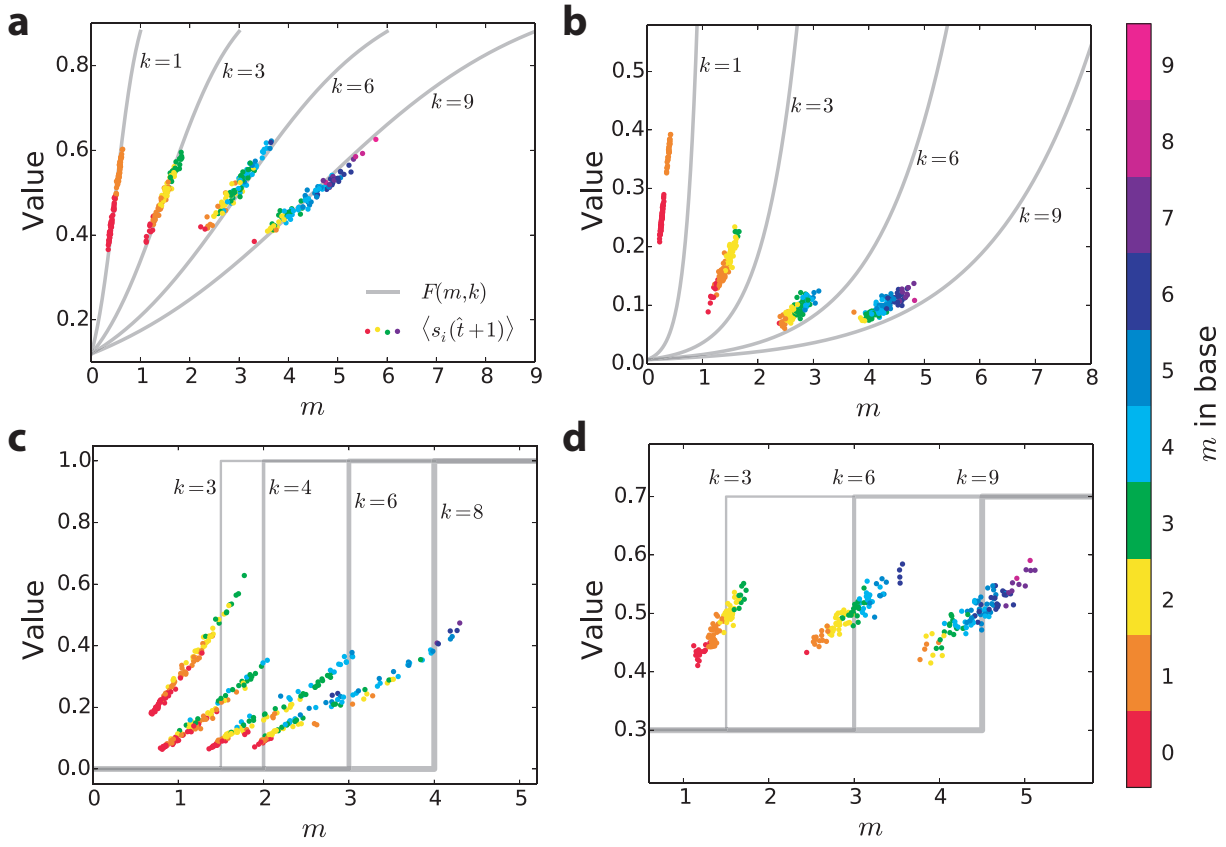

FIG. 2: **Data-based linearization for nonlinear and piecewise binary-state dynamics.** Linearization of the switching probability function $F_{k,m}$ for (a) Ising model, (b) evolutionary Game, (c) Threshold model and (d) Majority model. The grey lines represent Eq. (2) with $F_{k,m}$ for the models, where k is the node's degree and m is the number of active neighbors. Data points are the results of data-based linearization from time series and corresponding to linear 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 function value is obtained via $\langle s_i(\hat{t}+1)\rangle$. Each color of data points represents a set of subordinate strings whose base string has m active neighbors. The colors of data points demonstrate that bases with different m are necessary to produce a linear function with sufficient range of m for reconstruction, which justifies the base selection based on normalized Hamming distance in Fig. 1. For both nonlinear and piecewise switching function, linear function in the form of Eq. 4 is generated by the data-based linearization method, which is the key to the successful network reconstruction. The data points are obtained from an ER random network with $N = 100$ and $\langle k\rangle = 6$. More details of the data-based linearization can be seen in Supplementary Information Section 1.

III. NUMERICAL VALIDATION

We explore various dynamics on ErdösCRényi random (ER) [41], scale-free(SF) [42], small-world(SW) [43] and several empirical networks. For implementing network reconstruction, only states of nodes in different time steps are recorded and used, without any other knowledge of switching dynamics and network structure. To qualify the performance of reconstruction, we em-

ploy two standard indices: the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPR) [44] (see Supplementary Section II for the definitions of AUROC and AUPR). 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 nodes to measure the performance of reconstructing the whole network.

Figure 3 illustrates the reconstruction performance. Specifically, Fig. 3a shows the element values x_{ij} in the reconstructed neighboring vector $\mathbf{x}_i$ of all nodes for SW and SF network with Voter model. x_{ij} corresponding to links are generally greater than those of null connections. Given a cut-off in the the gap between two groups of points in Fig. 3a, links and null connections can be separated, leading to the reconstruction of the whole SW network and most nodes in SF networks. In SF networks, the neighbors of hubs are more difficult to be fully reconstructed, which is because of two facts: (i) in general the linearization is better for smaller node degrees, as exhibited in Fig. 2; (ii) the reconstruction based on the Lasso requires smaller amounts of data and offers better accuracy for sparser $\mathbf{X}_i$ associated with smaller degree nodes. Hub nodes because of the violation of the two requirements are hard to be fully reconstructed. However, a vast majority of nodes other than hubs can be still precisely reconstructed, giving rise to high accuracy of the whole network. The reconstruction results for SW networks and SF networks are shown in Fig. 3(b) and (c), respectively.

We explore how the number of base strings $\hat{t}$ affects the reconstruction accuracy. We define $n_{\hat{t}}$ as the number of $\hat{t}$ divided by the network size N to quantify relative amounts of base strings. As shown in Fig. 3d-g, receiver operating characteristic (ROC) curve and precision-recall (PR) curve show better performance as $n_{\hat{t}}$ increases for both SW and SF network, implying that high accuracy can be achieved from sufficient amounts of $n_{\hat{t}}$. Figure 3h,i shows the AUROC and AUPR as functions of $n_{\hat{t}}$ for SW and SF network respectively. We see that due to the advantage of the Lasso for dealing with sparse vectors, nearly perfect reconstruction is achieved after $n_{\hat{t}}$ exceeds a relatively small value, e.g., 0.4. Reconstruction results for the other dynamic models are exhibited in Supplementary Fig. 2. The length of time series is also significant for evaluating reconstruction efficiency. We investigate the AUROC and AUPR as functions of normalized length of time series for various dynamics on ER, SF and SW network (see Supplementary Fig. 3). We find that high reconstruction accuracy can be achieved from relatively small amounts of time series and the normalized length of time series needed to ensure 0.95 AUROC and AUPR decreases as N increases. These results indicate the high efficiency of our method and it is scalable for dealing with large networks.

We systematically apply our method to a variety of model and real networks in combination with the eight binary-state dynamics (see Table II), finding extremely high AUROC and AUPR for all combinations. We also investigated how representative network properties influence reconstruction performance, such as N and the average node degree $\langle k \rangle$ (see Supplementary Fig. 4,5). In practice, time series are usually contaminated by noise, and the data of some nodes may be lost or inaccessible, which call for the robustness to against the obstacle. We test the robustness of our method in more realistic situation. Specifically, we impose noise on the time series by randomly flipping a fraction of binary states in time series n_f (errors in time series), and assume the existence of a fraction of missing nodes n_m to mimic inaccessible nodes, as shown in Table III. We take Voter, Game, and Majority model as representative examples of linear, nonlinear and piecewise

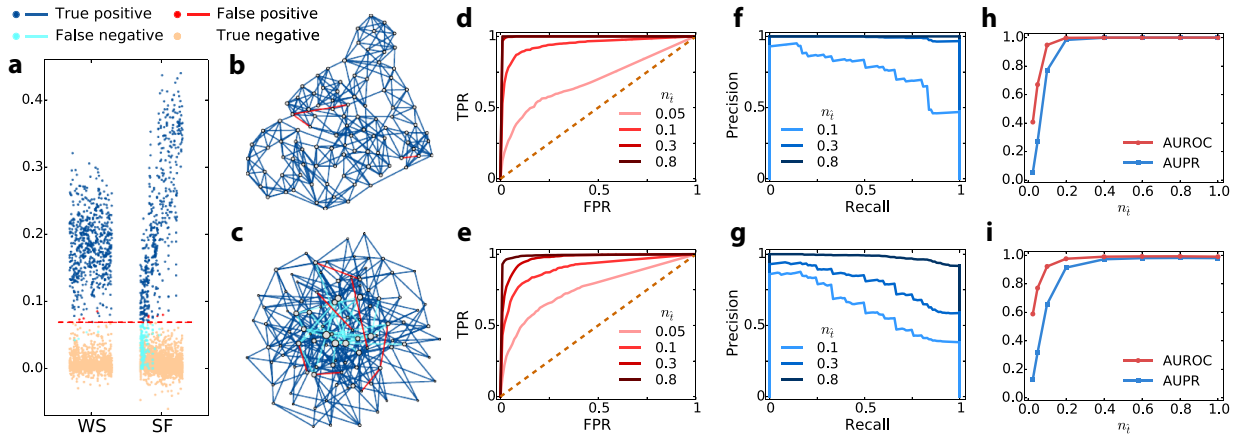

FIG. 3: **Reconstruction performance.** (a) Reconstructed values in the neighboring vector $\mathbf{x}_i$ of all nodes on SW and SF network with Voter model, where $N = 100$, $\langle k \rangle = 6$, $n_t = 15000$ and $n_{\hat{t}} = 80$. The red dashed line represents the threshold for determining whether a reconstructed value is considered to be linked or not (a value larger then the threshold will be deemed a link). The correctly reconstructed links (true positive), falsely reconstructed links (false positive) and missing links (false negative) are in dark blue, red and light blue points, respectively, while true negative links are in yellow. (b, c) Visualization of the reconstructed the SW and SF network, respectively. The color of reconstructed links are the same as that of the data points in (a). We see that missing links (false negative) in the SF network are more than that in the ER network. (d, e) ROC curve of reconstructed values for SW and SF network using different normalized amount of bases $n_{\hat{t}}$. (f, g) PR curve of reconstructed values for SW and SF network using different amount of $n_{\hat{t}}$. (h, i) AUROC and AUPR as functions of the normalized number of bases $n_{\hat{t}}$ for SW and SF network.

TABLE II: AUROC and AUPR for various dynamics in combination with model and empirical networks. Details of parameter values in dynamics are shown in Supplementary Table 1. The network size and mean degree of ER, SF and SW network are $N = 500$ and $\langle k \rangle = 6$ and n_t of time series used is 6×10^4. Information of empirical networks is shown in Supplementary Table 2 and n_t of time series used is 1.5×10^4.

AUROC/AUPR	Voter	Kirman	Ising	SIS	Game	Language	Threshold	Majority
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
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
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
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
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
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	0.996/0.847
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
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
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
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

dynamics. Strikingly, we find that high AUROC and AUPR remains even in the presence of 10% measurement noise or 30% inaccessible nodes, providing strong evidence for the robustness of our framework against measurement noise and inherent limits in accessing all nodes or missing data. More detailed results associated with Table III, i.e., AUROC and AUPR as functions of n_f and n_m, are displayed in Supplementary Fig. 6,7.

TABLE III: **Robustness against noise and missing data.** AUC and AUPR for Voter, Game, and Majority model on ER, SF and SW networks for measurement noise $n_f = 10\%$ and the fraction of inaccessible nodes $n_m = 30\%$, respectively. The network size $N = 500$ and mean degree $\langle k \rangle = 6$. The length of time series used is 6×10^4. Details of parameter values in dynamics are shown in Supplementary Table 1.

	$n_f = 10\%$			$n_m = 30\%$		
AUROC/AUPR	Voter	Game	Majority	Voter	Game	Majority
ER	0.995/0.938	0.955/0.707	0.991/0.864	1.000/0.985	0.999/0.983	1.000/0.988
SF	0.983/0.903	0.954/0.800	0.990/0.894	0.995/0.968	0.991/0.957	0.995/0.984
SW	1.000/0.984	0.976/0.741	0.994/0.874	1.000/0.988	1.000/0.988	1.000/0.988

IV. DISCUSSION

We have developed a general framework for addressing the challenging problem of reconstructing complex networks with binary-state dynamics, only from binary time series without any knowledge of switching function and structural information. Our main contribution lies in the development of a universal data-based linearization approach, which offers a general solution to the reconstruction of neighborhood of nodes for linear, nonlinear and discrete stochastic nodal dynamics. The task of reconstructing the whole network can thus be decomposed into the reconstruction of local structure centered at each node. The entire network can be recovered by simply assembling all local structures. The natural sparsity of real complex networks allows us to deal with the local reconstruction as a sparse signal reconstruction problem that can be addressed by employing the Lasso, a convex optimization method, from using a quite small amount of binary data. The optimization is also robust against measurement noise and missing data because of our limited accessibility to all nodes. The data-based linearization approach and the optimization based on the Lasso thus constitutes a general and purely data-based framework for reconstructing complex networks exclusively from binary time series, which is lacking prior to our current work. Our framework has been validated by using a variety of binary-state dynamic models in combination with a number of model and real complex networks. A generally high reconstruction accuracy has been achieved for all the studied cases, from using relatively small amounts of binary data contaminated by noise and the loss of partial data. These results suggest that potential applications of our framework in a wide range can be expected and addressing the inverse problem eventually will remarkably deepen our understanding of many complex networked systems with binary-state dynamics in nature and society.

Although our framework provides promising prospective of solving the inverse problem, some challenging problems remain. For example, although our framework is generally available for different types of switching function in binary-state dynamics, it may fail for non-monotonous function or non-Markovian dynamics. This is due to the facts that for the former cases, the data-based linearization is invalid because of the violation of one-to-one correspondence between the switching probability and active neighbors; for the latter, the merging process is inapplicable. Moreover, our framework is incapable of inferring interaction strength between nodes, especially in the presence of noise and missing observation. Despite these open questions, our framework provide significant insight into the inverse problem of complex networked systems with binary-

state dynamics and may motivate further effort in the pursuit of eventually solving the inverse problem completely.

V. METHODS

Models of binary-state dynamics. The voter model [16] assumes that a node randomly chooses one of its neighbors' states in each time step. If the total number of a node's neighbor is k and m among them are active, then the probability it becomes active is m/k while the probability of becoming inactive is $(k-m)/k$. In the majority-vote model[45], a node tends to align with the major state of its neighbors, but with a probability Q of misalignment.

In the Kirman's ant colony model [46], nodes transfer from state 0 to 1 with the probability $F_{k,m} = c_1 + dm$ when there are m active neighbors, and change back from 1 to 0 with the rate $R_{k,m} = c_2 + d(k-m)$ correspondingly. The parameters c_1 and c_2 quantify the individual action that is independent to the states of neighbors, while the parameter d represents the the action of copying from neighbors.

Ising model [18] is a classical model delineating magnetic spins, where each node is either in spin-up or spin-down state. The switching is adopted with a certain form of probability, driving the system to minimization of the Hamiltonian. Here we choose the transition rates in Glauber dynamics [47] as shown in Table I. The parameter β stands for a combination of temperature and ferromagnetic-interaction parameter.

The SIS model [5] describes a disease-spreading dynamics with infection and recovery. Each susceptible individual contracts disease from each of its infected neighbors at a rate λ. Thus, a susceptible node with m infected neighbors has the probability $(1-\lambda)^m$ of remaining susceptible at each time step, leading to the infection rate $1-(1-\lambda)^m$. Meanwhile, the recovery rate of a infected node is μ in every moment of time.

The game model [4] comes from the game theory. When embedding on networks, each node is occupied by a player, and the two states stand for different strategies. Each player plays with each of his/her neighbors using one chosen strategy in each time step. According to the game theory, the profit of a rational player i when playing with a neighbor j can always be characterized by a payoff matrix $\begin{array}{c} \\ s_1 \\ s_2 \end{array}\begin{array}{c} \begin{array}{cc} s_1 & s_2 \end{array} \\ \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix} \end{array}$ where a and b are game parameters. Different games can be produced by adjusting a and b. The payoff of a player is the sum of profits with all his/her neighbors. A player switches the strategy with a probability depending on the payoff it may gain in the next round under the current circumstance, as shown in Table I. Parameter α qualifies the willing of an individual changing his/her mind according to the strategies of their neighbors, and β is associated with the influence of expected payoffs.

For the language model [48], the two states denote different language choices of a person. Transitions from the primary language to another occur proportionally to the fraction of speakers in the neighbors with the power α, multiplied by the parameter s or $1-s$ according to the respective language.

The threshold model [49] is a deterministic model. A certain threshold M_k, which may be a function of the node's degree, is set to each node. In each time step, a node turns to active if

the number of its active neighbor m exceeds the threshold M_k, and no recovery transformation is permitted.

Procedure of choosing bases. Theoretically, a base string should not be chosen arbitrarily. On the one hand, if a base string is too special to find its subordinate strings, the estimation of the switching probability via the average will deviate from the true value. On the other hand, if the bases resemble each other closely, little differences in the switching probabilities will lead to difficulty in reconstruction, because of the small range in the linearized function. To choose the most proper bases among all available strings, we propose a method to select base strings in the network composed of base strings. For an arbitrary node i, we first construct a network where 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 normalized Hamming distance between strings. We then eliminate edges whose weight is smaller than the threshold Δ. The remaining edges indicate sufficient similarity between vertices. By setting another threshold σ, we extract a subnetwork where only the top $\sigma \times N$ vertices with larger degree are preserved, while other vertices and their edges are removed. In this way, all remaining strings have relatively sufficient amount of subordinate strings similar to them. Finally, we pick out the vertices with smaller degree according to the data requirement, so that the selected base strings will sufficiently different. Figure 1b shows the process of choosing base, and see Supplementary Information Sec. I for detailed parameter values and discussion.

The Lasso for reconstructing $\mathbf{x}_i$ from $\mathbf{y}_i = \Phi_i \times \mathbf{x}_i$. Using our method of pretreating data, $\mathbf{y}_i$ and Φ_i can be collected and calculated solely from the time series. Thus the problem of recovering the node i's links has been converted into reconstructing a vector $\mathbf{x}_i$ from a linear measurement $\mathbf{y}_i = \Phi_i \times \mathbf{x}_i$:

$$\begin{bmatrix} \langle s_i(\hat{t}_1+1)\rangle \\ \langle s_i(\hat{t}_2+1)\rangle \\ \vdots \\ \langle s_i(\hat{t}_M+1)\rangle \end{bmatrix} = \begin{bmatrix} 1 & \langle s_1(\hat{t}_1)\rangle & \cdots & \langle s_{i-1}(\hat{t}_1)\rangle & \langle s_{i+1}(\hat{t}_1)\rangle & \cdots & \langle s_N(\hat{t}_1)\rangle \\ 1 & \langle s_1(\hat{t}_2)\rangle & \cdots & \langle s_{i-1}(\hat{t}_2)\rangle & \langle s_{i+1}(\hat{t}_2)\rangle & \cdots & \langle s_N(\hat{t}_2)\rangle \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \langle s_1(\hat{t}_M)\rangle & \cdots & \langle s_{i-1}(\hat{t}_M)\rangle & \langle s_{i+1}(\hat{t}_M)\rangle & \cdots & \langle s_N(\hat{t}_M)\rangle \end{bmatrix} \begin{bmatrix} d_i \\ c_i \cdot a_{i1} \\ \vdots \\ c_i \cdot a_{i,i-1} \\ c_i \cdot a_{i,i+1} \\ \vdots \\ c_i \cdot a_{iN} \end{bmatrix}. \quad (5)$$

Note that $\mathbf{x}_i$ is usually sparse since the number of the neighbors of node i is much less than the network scale N in most systems. The sparsity of $\mathbf{x}_i$ satisfies the prerequisite of the Lasso [35], a convex optimization method, which fittingly solves our reconstruction problem. The problem the Lasso addresses is to optimize

$$\min_{\mathbf{x}_i}\Big\{\frac{1}{2M}\|\Phi_i\mathbf{x}_i - \mathbf{y}_i\|_2^2 + \lambda\|\mathbf{x}_i\|_1\Big\}, \quad (6)$$

where $\|\mathbf{x}_i\|_1 = \sum_{j=1,j\neq i}^{N} |x_{ij}|$ is the L_1 norm of $\mathbf{x}_i$ assuring the sparsity of solution, the least square term $\|\Phi_i\mathbf{x}_i - \mathbf{y}_i\|_2^2$ provides robustness of the solution against noise in data. λ is a nonnegative regularization parameter which affects performance of reconstruction according to the sparsity of networks, and can be determined by cross-validation method [50](see Supplementary Information

Session 2). A striking advantage of using the Lasso is that M, i.e., the number of bases needed is much less than the length of $\mathbf{x}_i$. And for each base of each node, the strings included can be collected and calculated from only one set of data sampling in time series, ensuring relatively sparse data requirement. After vector $\mathbf{x}_i$ is reconstructed, the direct neighbors of node i correspond to the nonzero elements in it. In the same manner, we uncover the neighbors of all other nodes, yielding the full structure of the network by simply matching the neighbors of all nodes.

[1] C. Castellano, S. Fortunato, and V. Loreto, Reviews of modern physics **81**, 591 (2009).

[2] A. Kumar, S. Rotter, and A. Aertsen, Nature reviews neuroscience **11**, 615 (2010).

[3] J. Hasty, D. McMillen, F. Isaacs, and J. J. Collins, Nature Reviews Genetics **2**, 268 (2001).

[4] G. Szabó and G. Fath, Physics Reports **446**, 97 (2007).

[5] R. Pastor-Satorras and A. Vespignani, Physical review letters **86**, 3200 (2001).

[6] J. Balthrop, S. Forrest, M. E. Newman, and M. M. Williamson, arXiv preprint cs/0407048 (2004).

[7] K. Sznajd-Weron and J. Sznajd, International Journal of Modern Physics C **11**, 1157 (2000).

[8] M. E. Newman, Physical review E **66**, 016128 (2002).

[9] H. Ebel and S. Bornholdt, Physical Review E **66**, 056118 (2002).

[10] J. C. Dunlap, Cell **96**, 271 (1999).

[11] I. Dobson, B. A. Carreras, V. E. Lynch, and D. E. Newman, Chaos: An Interdisciplinary Journal of Nonlinear Science **17**, 026103 (2007).

[12] M. P. Niemira and T. L. Saaty, International Journal of Forecasting **20**, 573 (2004).

[13] H. E. Stanley, Introduction to Phase Transitions and Critical Phenomena, by H Eugene Stanley, pp. 336. Foreword by H Eugene Stanley. Oxford University Press, Jul 1987. ISBN-10: 0195053168. ISBN-13: 9780195053166 **1** (1987).

[14] A. Barrat, M. Barthelemy, and A. Vespignani, *Dynamical processes on complex networks* (Cambridge University Press, 2008).

[15] M. Newman, *Networks: an introduction* (Oxford University Press, 2010).

[16] V. Sood and S. Redner, Physical review letters **94**, 178701 (2005).

[17] D. H. Zanette, Physical review E **65**, 041908 (2002).

[18] P. L. Krapivsky, S. Redner, and E. Ben-Naim, *A kinetic view of statistical physics* (Cambridge University Press, 2010).

[19] J. P. Gleeson, Physical Review X **3**, 021004 (2013).

[20] S. H. Strogatz, Nature **410**, 268 (2001).

[21] A.-L. Barabási, Nature Physics **8**, 14 (2011).

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

[23] T. S. Gardner, D. Di Bernardo, D. Lorenz, and J. J. Collins, Science **301**, 102 (2003).

[24] N. Friedman, Science **303**, 799 (2004).

[25] M. Timme, Phys. Rev. Lett. **98**, 224101 (2007).

[26] A. Clauset, C. Moore, and M. E. Newman, Nature **453**, 98 (2008).

[27] S. Guo, J. Wu, M. Ding, and J. Feng, PLoS Comput. Biol. **4**, e1000087 (2008).

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

[29] W.-X. Wang, Y.-C. Lai, C. Grebogi, and J. Ye, Phys. Rev. X **1**, 021021 (2011).
[30] B. Barzel and A.-L. Barabási, Nat. Biotechnol. **31**, 720 (2013).
[31] S. Feizi, D. Marbach, M. Médard, and M. Kellis, Nat. Biotechnol. **31**, 726 (2013).
[32] G. Caldarelli, A. Chessa, F. Pammolli, A. Gabrielli, and M. Puliga, Nat. Phys. **9**, 125 (2013).
[33] Z. Shen, W.-X. Wang, Y. Fan, Z. Di, and Y.-C. Lai, Nat. Commun. **5** (2014).
[34] X. Han, Z. Shen, W.-X. Wang, and Z. Di, Phys. Rev. Lett. **114**, 028701 (2015).
[35] T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin, The Mathematical Intelligencer **27**, 83 (2005).
[36] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, Phys. Rep. **424**, 175 (2006).
[37] Y.-Y. Liu, J.-J. Slotine, and A.-L. Barabási, Nature **473**, 167 (2011).
[38] T. Nepusz and T. Vicsek, Nature Physics **8**, 568 (2012).
[39] G. Yan, J. Ren, Y.-C. Lai, C.-H. Lai, and B. Li, Physical review letters **108**, 218703 (2012).
[40] Z. Yuan, C. Zhao, Z. Di, W.-X. Wang, and Y.-C. Lai, Nature communications **4** (2013).
[41] P. Erdös and A. Rényi, Publ. Math. Inst. Hungar. Acad. Sci **5**, 17 (1960).
[42] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).
[43] D. J. Watts and S. H. Strogatz, Nature **393**, 440 (1998).
[44] J. Davis and M. Goadrich, in *Proceedings of the 23rd international conference on Machine learning* (ACM, 2006), pp. 233–240.
[45] M. J. de Oliveira, Journal of Statistical Physics **66**, 273 (1992).
[46] A. Kirman, The Quarterly Journal of Economics pp. 137–156 (1993).
[47] R. J. Glauber, Journal of mathematical physics **4**, 294 (1963).
[48] D. M. Abrams and S. H. Strogatz, Nature **424**, 900 (2003).
[49] M. Granovetter, American journal of sociology pp. 1420–1443 (1978).
[50] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., The Journal of Machine Learning Research **12**, 2825 (2011).

Acknowledgements
We thank Zhesi Shen for valuable discussion and help.

Author contributions

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