Reconstructing complex networks without time series

Chuang Ma, Hai-Feng Zhang, and Ying-Cheng Lai

Phys. Rev. E **96**, 022320 — Published 25 August 2017

DOI: 10.1103/PhysRevE.96.022320

Reconstructing complex networks without time series

Chuang Ma,[1] Hai-Feng Zhang,[1,2,3,*] and Ying-Cheng Lai[4]

[1]*School of Mathematical Science, Anhui University, Hefei 230601, China*
[2]*Center of Information Support &Assurance Technology, Anhui University, Hefei 230601, China*
[3]*Department of Communication Engineering, North University of China, Taiyuan, Shan'xi 030051, China*
[4]*School of Electrical, Computer and Energy Engineering, Department of Physics, Arizona State University, Tempe, Arizona 85287, USA*

(Dated: August 7, 2017)

In the real world there are situations where the network dynamics are transient (e.g., various spreading processes) and the final nodal states represent the available data. Can the network topology be reconstructed based on data that are not time series? Assuming that an ensemble of the final nodal states resulting from statistically independent initial triggers (signals) of the spreading dynamics is available, we develop a maximum likelihood estimation based framework to accurately infer the interaction topology. For dynamical processes that result in a binary final state, the framework enables network reconstruction based solely on the final nodal states. Additional information, such as the first arrival time of each signal at each node, can improve the reconstruction accuracy. For processes with a uniform final state, the first arrival times can be exploited to reconstruct the network. We derive a mathematical theory for our framework and validate its performance and robustness using various combinations of spreading dynamics and real-world network topologies.

PACS numbers:

I. INTRODUCTION

Extensive research in the past two decades revealed that network structure can play a fundamental and critical role in the collective dynamics on the network. However, in realistic situations, the network structure and nodal dynamics are often unknown but only limited measured time series are available. To determine the full topology and structure of a complex network from data has thus evolved into an important area of research with significant applications [1]. For example, data-based network reconstruction is pertinent to biomedical sciences such as gene regulatory networks [2], systems biology [3], and psychology [4]. In the past there were advances in this field [5–43], where various network reconstruction methods were developed such as those based on the Pearson correlation [44, 45], phase synchronization [46, 47], delayed feedback control [10, 25], noise induced fluctuations [17, 19, 31], and compressive sensing [26–28, 39–43]. A common feature among many existing methods is that measured *time series* of a finite duration from the system are needed for the reconstruction task. That is, it is necessary to have full or partial information about the dynamical trajectory of the networked system [24, 41, 48] to enable reconstruction.

There are real-world situations where the nodal dynamics are transient with a short lifetime and only the final nodal states of the network are available. For example, after a rapid outbreak of an infectious disease, the individuals who have been infected can be observed, and this may be the only available information. A similar situation arises with information diffusion on networks (e.g., online rumor or opinion propagation and spreading), where the final subpopulation that received the information is known. Likewise, for a given web page a record of the individuals who have visited the page is often available, but the detailed browsing history of these individuals is usually not known. To reconstruct the network topology without time series, where the only available information is the final nodal states, poses an extreme challenge in the field of reverse engineering of complex networks.

In this paper, we develop a general framework to infer the network topology based solely on information about the final nodal states in the absence of any time series. Since many diffusion processes can be conceptualized as propagation of "signals" [49] in networks (e.g., virus, rumors, opinions, data packets, or passengers), the final state of any node is often binary: either it has received the signal or not. We exploit maximum likelihood estimation (MLE) [50] to ascertain the existence of actual links among the nodes using only the final binary states. We develop a mathematical theory with proved theorems to establish the framework and demonstrate its performance using a large number of model and empirical networks. We address the issue of robustness by numerically assessing and mathematically analyzing the effects of random signal disturbances. A finding is that, for dynamical processes with a binary final state, when "extra" information is available, such as the first arrival time of the signal at each node, the reconstruction accuracy can be markedly improved. Even for processes with a uniform final state, the first arrival times can be exploited to uncover the network topology.

II. MODEL

A. Reconstruction framework

We consider a network of N nodes and M propagating signals. For each node, a binary dynamical variable can be defined with two states: either the node has received a signal or not. In this work, to verify the universality of our recon-

*Electronic address: haifengzhang1978@gmail.com

struction framework, the propagations of signals on networks can be simulated by Susceptible-Infected-Recovery (SIR) epidemic model, rumor spreading model, or the mixture of them. The detailed descriptions of them are presented in Sec. III A. The data or information needed for reconstruction can then be represented by a matrix S, where every element S_{ij} assumes one of the two values: one or zero, with the first index i specifying the signal and the second index j labeling the node. Specifically, we have $S_{ij} = 1$ if signal i has transmitted to node j, otherwise $S_{ij} = 0$. For convenience, we call S the *global information matrix*. Our goal is to reconstruct the full topology of the network according to matrix S, which contains information about final nodal states only. Figure 1 presents a schematic illustration of this task for the concrete case of $N = 8$ and $M = 16$.

For convenience, we use Fig. 1 to explain how the global information matrix S can be used to infer the network structure. From each row, we can identify the nodes that have received the signal and the infected nodes provide limited structural information about the whole network. Using many rows of the matrix S, we can infer the whole network structure. Take the first row of S as an example. It can be seen that nodes 3, 5, 7 and 8 have received the first signal, which contains the three existing links (3, 8), (5, 8) and (7, 8), as shown in Fig. 1(d). Using information about the first row only is not sufficient to infer the three links, because the row also contains the possible links (3, 5), (3, 7) and (5, 7). Exploiting other rows can eliminate the redundant links. For instance, from the matrix S in Fig. 1(a), we see that the states of nodes 3 and 5 in many rows are different, indicating lack of a link between them.

Step 1 - construction of global and local information matrices. To reconstruct the whole network, we begin with reconstructing the local connection topology for each node. For each node j, we define a *local information matrix* S^j characterizing the signals arriving at node j, which can be extracted from the global information matrix S. In particular, from S, we remove the jth column and keep only the rows corresponding to signals actually received by node j. A non-zero element of the matrix S_j thus has the following meaning. Say $S^j_{ik} = 1$. This means that signal i, which has been received by node j, also arrives at and is received by node k. Otherwise, $S^j_{ik} = 0$. To better describe the relation between global and local information matrices, we present a schematic illustration in Fig. 1, where Fig. 1(a) presents an example of the global information matrix S, and Fig. 1(b) shows the local information matrix S^6 of node 6 extracted from S. We see that S^6 consists of 9 rows in S (marked by the red dotted boxes: rows 3, 5, 6, 8, 9, 11, 12, 14 and 16, which are re-indexed from 1 to 9 in S^6). For S^j, column-wise we keep those corresponding to nodes whose final states are not opposite to that of node j for each signal. That is, for nodes whose states are exactly opposite to that of node j for each signal, the corresponding columns in S^j are removed. An example is the column corresponding to node 5 in S^6, where for each and every signal, if it is received by node 6 then it will never arrive at node 5, and vice versa, as shown in Fig. 1(b). As a result, the column corresponding to node 5 is removed. The final local information matrix S^6 for node 6 is shown in Fig. 1(c).

Step 2 - maximum likelihood estimation based inference of local topology.

For each node j, we need to infer its neighbors (the nodes that are connected with node j) based on the local information matrix S^j. To do so, we view node j as the recipient of the signals starting from other nodes. Because signals can only diffuse through the links in the networks, all signals starting from other nodes to node j should be received by some neighbors first and are then passed onto node j. Take the sixth row in S^6 as an example [see Fig. 1(c), highlighted by the dashed red box], the signal is received by nodes 1, 2, 3 and 8 [the green nodes in Fig. 1(d)]. However, as shown in Fig. 1(d), signals from nodes 1, 2 and 8 cannot directly reach node 6 (red node): they must reach node 3 first before arriving at node 6. To quantify this process, we define P^k_j as the probability that node k can *directly* pass signals to node j, namely, the signals received by node j were passed on by node k. As a result, there is a possible connection between the two nodes if $P^k_j > 0$ and no such connection exists if $P^k_j = 0$. For each signal starting from another node to node j, at least one of the neighbors has received the signal. We thus have $\sum_k P^j_k \equiv 1$ for each row of S^j. The probability for node j to receive signal i, $P(S_{ij} = 1)$, is determined by whether a node k also received the signal and their directly connection probability P^k_j, namely,

$$P(S_{ij} = 1) = \sum_{k \neq j} P^j_k \times S^j_{ik} = S^j(i,:) \cdot P^j, \qquad (1)$$

where $S^j(i,:)$ denotes the ith row of S^j and $P^j = [P^j_1, P^j_2, \cdots, P^j_{N_j}]^T$ with N_j being the number of columns in S^j.

If the direct connection probability is correctly predicted, the probability of $P(S_{ij} = 1)$ will be large for those signals received by node j. Our goal is thus to find the direct connection probability $P^j = [P^j_1, P^j_2, \cdots, P^j_{N_j}]^T$ to maximize the product of the probability $P(S_{ij} = 1)$ that signal i is received by node j. Letting $f\left(P^j\right) = \prod_i \left(S^j(i,:) \cdot P^j + \epsilon\right)$, we obtain the direct connection probability $P^j = [P^j_1, P^j_2, \cdots, P^j_{N_j}]^T$ using the principle of MLE:

$$\max f\left(P^j\right), \;\; s.t \qquad \begin{cases} P^j_k \geq 0; \\ \sum\limits_k P^j_k = 1, \end{cases} \qquad (2)$$

where ϵ is an error tolerance parameter, which in numerical simulations can be fixed at some arbitrarily small value (e.g., $\epsilon = 0.01$). We also test other small values of ϵ (e.g., 10^{-4}, 10^{-3}, and 10^{-1}) and find consistent reconstruction accuracy.

Since each factor in Eq. (2) is less than unity: $\left(S^j(i,:) \cdot P^j + \epsilon\right) \leq 1$, their product can be arbitrarily small, leading to an equally small value for $\max f\left(P^j\right)$ and, consequently, to a large estimation error. To overcome this difficulty, we use the logarithmic form of $\max f\left(P^j\right)$ to write

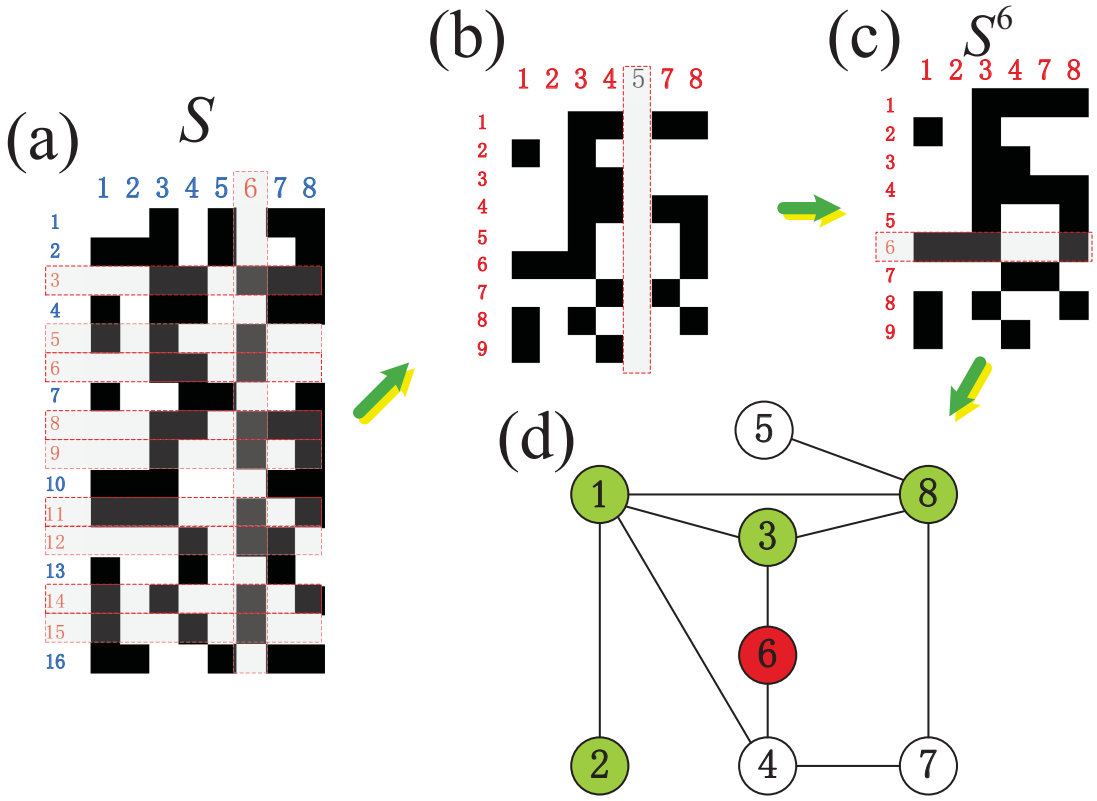

FIG. 1: Schematic illustration of construction of global information matrix S and local information matrix S^j extracted from S. (a) Global information matrix S, where $S_{ij} = 1$ if signal i is received by node j (marked by black); otherwise, $S_{ij} = 0$ (blank). (b) Matrix S^6 for node 6 extracted from S, which contains the rows of S with $S_{i6} = 1$, i.e., the red boxes in (a) with element indices renumbered. Note that S^6 does not contain the column of node 6 in S. Since the state of node 5 is opposite to that of node 6 for each signal, the column corresponding to node 5 is also removed. (c) The final local information matrix S^6. (d) The original network with $N = 8$. The eleventh signal (the sixth row of S^6 in (c), highlighted by dashed red box) is also received by nodes 1, 2, 3 and 8 (green nodes). However, only node 3 can directly diffuse the signal to node 6 (red node). The signals from nodes 1, 2 and 8 must reach node 3 first before arriving at node 6. Other signals must pass through node 3 or node 4 to reach node 6.

formula (2) as

$$\max \sum_i \ln \left(S^j \left(i, : \right) P^j + \epsilon \right),$$
$$s.t \quad \begin{cases} P_k^j \geq 0; \\ \sum_k P_k^j = 1. \end{cases} \tag{3}$$

Equation (3) is a standard optimization problem. Its solution P^j can be used to ascertain the existence of links. In particular, there is an edge connecting nodes k and j if P_k^j can be distinguished from zero statistically.

Figure 2 presents a schematic illustration of our MLE based reconstruction method. Specifically, Fig. 2(a) shows the connecting structure of a model network of size $N = 19$, on which an epidemic spreading process has occurred. To be concrete, we assume that the dynamics is described by SIR epidemic model with the following parameter values: transmission rate $\beta = 0.2$ and recovery rate $\mu = 1.0$. Say the task is to identify the neighbors of red node 11. We generate 500 independent realizations of the dynamics, each starting from a random initial seed of the epidemic source, i.e., $M = 500$. The global information matrix S thus has 500 rows and 19 columns. Figure 2(b) shows the matrix S^{11} extracted from S. Solving the underlying MLE optimization problem (3), we obtain the values of P_k^{11}, where those values that can be distinguished from zero are shown Fig. 2(c). For the values in Fig. 2(c), the corresponding part of S^{11} is shown in the upper right corner of Fig. 2(b). As shown in Fig. 2(d), the MLE based method yields correctly the neighbors of node 11. Repeating this for every node, we can obtain the full topology of the whole network.

B. Dimension reduction

The dimensions of the matrices S and S^j increase with N and M, leading to a rapid increase in the number of unknown parameters in (3) and consequently to inaccurate solutions of the MLE optimization problem. Thus, for large networks and/or large number of signals, it is desired to reduce the dimensions of S and S^j. Intuitively, if a signal i is received by too many nodes, it will have little contribu-

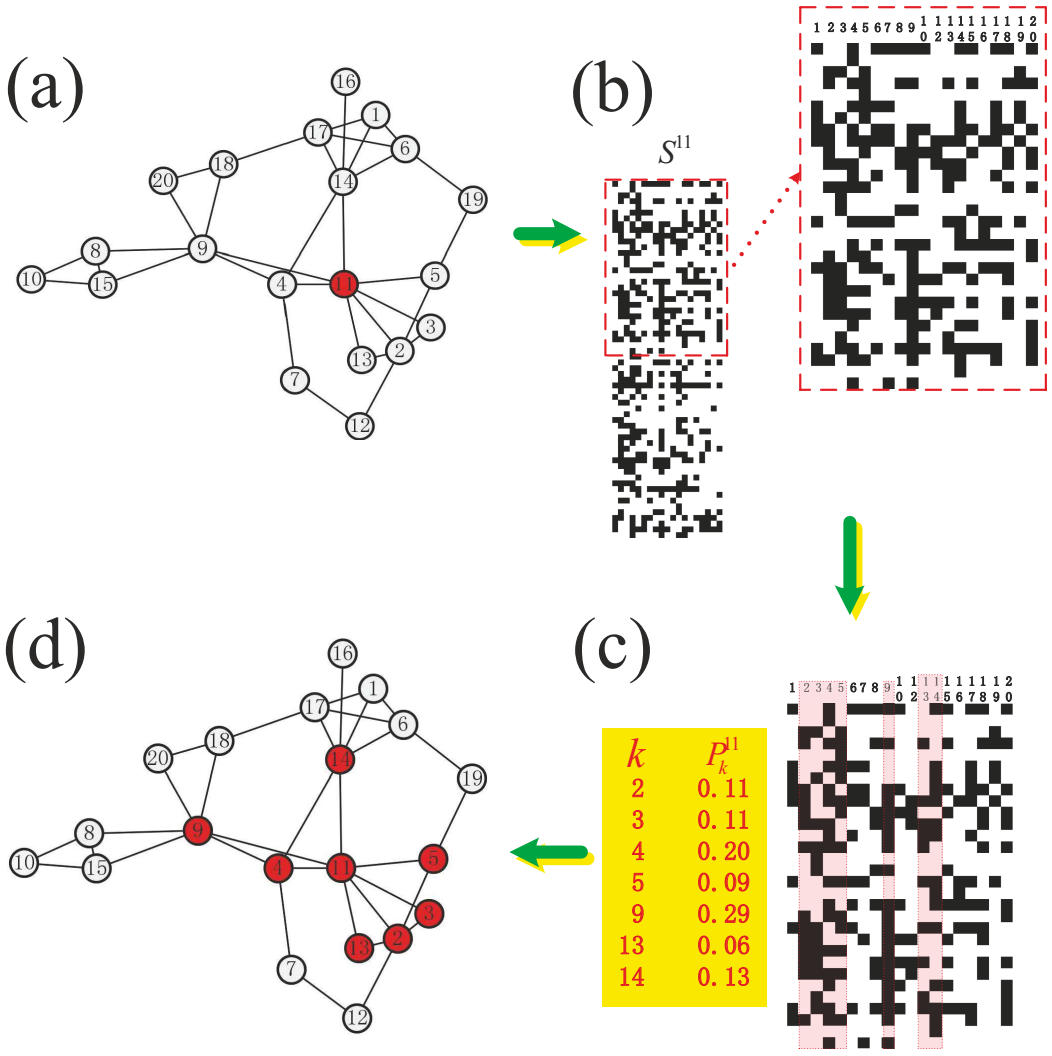

FIG. 2: Illustration of MLE based method for network reconstruction without any dynamical trajectory or time series. (a) A network of 19 nodes, and the task is to identify the neighbors of node 11. (b) Matrix S^{11} of node 11 extracted from the global information matrix S, where S is established using 500 signals generated from 500 independent realizations of SIR epidemic process. (c) The non-zero directional connection probability values associated with node 11 to other nodes obtained from the solution of the optimization problem (3). The subset of the local information matrix S^{11} which corresponds to these directional links is shown in the upper right inset of (b). (d) Inferred neighbors of node 11 from the directional connection probability values. The results are averaged over 10 independent runs.

tion to the reconstruction solution as it leads to indistinguishable nodal states. For example, from formula (2), we see that, when a signal is received by all nodes, it has no effect because $S^j(i,:)\cdot P^j+\epsilon = 1+\epsilon$ is a fixed value. In general, the signals received by a few nodes play a determining role in reconstruction. To identify such signals, we set a threshold parameter δ, where signal i is deemed useless and is removed from the matrix S^j if the following inequality holds: $\sum_k S^j_{ik} > \delta$. If the value of δ is too small, some signals will be discarded. However, if δ is too large, the computational time will increase. Through tests we set $\delta = 20$.

To reduce the dimension of S^j in the horizontal direction (i.e., to reduce the number of columns), we prove a theorem stipulating that the kth column is redundant and can be removed from S^j if all the signals received by k are also received by node l. In this case, we say that the kth node is nested in the lth node, and the number of columns of S^j can be drastically reduced. We state the theorem here and provide a detailed proof in Appendix A.

- *Theorem 1* If column b is nested in column a, i.e., $S^j_{ia} \geq$

S^j_{ib} $(i = 1, 2, \cdots, M^j)$ and $\sum_i S^j_{ia} > \sum_i S^j_{ib}$, then one has $P^j_b = 0$ from Eq. (2).

For the special case where there are *only three* columns in S^j, the following theorem provides a criterion to determine whether information from a particular node is pertinent for the reconstruction task (a detailed proof is given in Appendix B).

- *Theorem 2* Suppose the matrix S^j has three columns only and they meet the following conditions: $S^j_{i1} + S^j_{i2} \geq S^j_{i3}$ $\left(i = 1, 2, \cdots, M^j\right)$, $\sum\limits_i S^j_{i1} > \sum\limits_i S^j_{i3}$, and $\sum\limits_i S^j_{i2} > \sum\limits_i S^j_{i3}$. For $\epsilon \to 0$, we have $P^j_3 = 0$ if either one of two conditions is satisfied:
(1) $n_1 \geq n_5$ and $n_2 > n_4$, or
(2) $n_5/(n_1 + n_3) + n_4/(n_2 + n_3) \leq 1$.

While Theorem 2 holds rigorously for the case where the matrix S^j has three columns, a heuristic argument and numerical evidence in Appendix C and Fig. 12 indicate that, in more general cases, the theorem can still be applied to reducing the number of columns of S^j so as to guarantee accurate reconstruction. To validate that the two Theorems provide an effective guideline to reduce the number of unknown parameters (n) in (2), we compare the values of n for nodes versus the degree in an Email network, as shown in Fig. 3. We see that the value of n can decrease by a factor of two when S^j is reduced according to Theorem 1 (labeled as $S1$). The value of n can be further reduced when Theorem 2 is applied (labeled as $S2$). In this example, the number of unknown quantities can be reduced by a factor of seven. An empirical rule is that, if the network size N is below 1000, we apply Theorem 1. Otherwise, Theorem 2 can be applied, even it is not rigorous if the number of columns is greater than 3.

III. PRELIMINARIES

A. Spreading process

The stochastic propagation of signals on a network is treated, as follows. For the classical SIR model on a network [51], each node can be in one of the three states: Susceptible (S), Infected (I) or Recovered (R). An infected node can infect each susceptible neighbor with transmission rate β, and the original infected node enters the recovery state with rate μ. For the rumor spreading process on network, each node can also be in one of three states: ignorant (have not heard the rumor and are susceptible to be informed), spreader (who knows and spreads the rumor), or stifler (who knows the rumor but has no interest in spreading it). At each time step, a spreader can infect its each ignorant neighbor with transmission rate β or can become a stifler with rate μ once it contacts a spreader or a stifler neighbor. The key difference between SIR and rumor models is that, for the former, infected nodes are recovered nodes by themselves while for the latter, a spreader becomes a stifler only when it makes contact with at least a neighbor that is a spreader or a stifler. Because of this difference, the structure of the network has distinct effects on the stochastic process of signal propagation. For example, for the SIR dynamics, hub nodes can enhance spreading but for rumor dynamics hub nodes often act as firewalls that inhibit spreading [52, 53]. For the mixed model, each signal is diffused on networks according to the SIR or the rumor spreading process with equal probability.

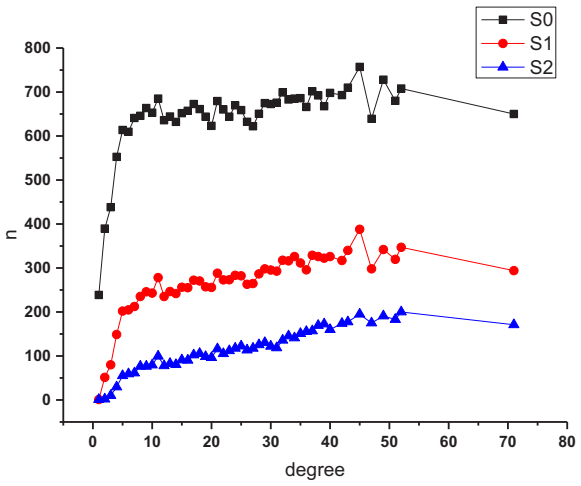

FIG. 3: Dimension reduction based on Theorems 1 and 2. For the Email network, the number of unknown parameter n as a function of degree without and with dimension reduction. S0: without reduction; S1: reduction based on Theorem 1; S2: reduction based on Theorem 2. Ten independent runs are used for statistical average.

B. Data sets

We use four empirical networks to test our reconstruction framework and to analyze the accuracy, which are (1) Zachary's karate club network, a network of friendship among the members of a university karate club [54], (2) Dolphins network, a network of frequent associations between dolphins living near Doubtful Sound, New Zealand [55], (3) Football network, a network of the schedule of games between American college football teams in a single season [56], and (4) Email network, the e-mail network of University at Rovira i Virgili, URV [57]. Detailed information about these empirical networks are presented in Table I.

TABLE I: **Basic topological features of four empirical networks**. N and E are the total numbers of nodes and links, respectively. The average degree of the network is $\langle k \rangle$, C and r are the clustering coefficient and the assortative coefficient, respectively, and H is the degree heterogeneity defined as $H = \langle k^2 \rangle / \langle k \rangle^2$.

Network	N	E	$\langle k \rangle$	C	r	H
Karate	34	78	4.5882	0.5879	-0.4756	1.6933
Dolphins	62	159	5.129	0.2901	-0.0718	1.3255
Football	115	613	10.6609	0.4032	0.1624	1.0069
Email	1133	5451	9.6222	0.2540	0.0782	1.9421

C. Performance indicators

The true positive rate TPR and the false positive rate FPR are used to characterize the accuracy of our reconstruction framework, which are defines as TPR $=$ TP/(TP $+$ FN) and FPR $=$ FP/(FP $+$ TN), respectively, where TP, FN, FP and TN represent true positive, false negative, false positive and true negative, respectively. A larger value of TPR and a smaller value of FPR indicate high accuracy [58, 59]. The F1-score is a summary measure [60] that combines the precision and efficiency of the algorithm, which is defined as F1 $= (2 \times P \times R)/(P + R)$, where $P =$ TP/(TP $+$ FP) and $R =$ TP/(TP $+$ FN).

IV. MAIN RESULTS

A. Performance of reconstruction based on final nodal states

To demonstrate the general applicability of our MLE based reconstruction framework, we study the SIR epidemic model, rumor model, and a mixed spreading process. In numerical simulations, we choose the transmission rate β for the SIR and rumor models randomly from the intervals $[0, 0.3]$ and $[0.4, 0.6]$, respectively. The recovery rate is set to be $\mu = 1.0$. For the mixed model, each signal is diffused on networks according to SIR or the rumor spreading process with equal probability.

To characterize the accuracy of the global information matrix S reconstructed from the final nodal states of the network, the values of TPR and FPR as the functions of the number M of signals are investigated. The results are summarized in Figs. 4-6 for three types of model networks: Erdös-Rényi (ER) random [61], Watts and Strogatz small-world (SW) [62], and Barabási-Ablert scale-free (BA) networks [63], and for SIR, rumor, and mixed spreading processes, respectively. We see that in all cases, TPR increases with M but the opposite trend occurs for FPR. For sufficiently large values of M, we have TPR$= 1$ and FPR$= 0$, indicating that the whole network can be fully reconstructed with zero error. A comparison of Figs. 4(a), 5(a) and 6(a) reveals that the network structure has an effect on the reconstruction. In particular, for a fixed (large) value of M, the ER network [Fig. 4(a)] gives the highest accuracy, while the lowest accuracy occurs for the BA network [Fig. 6(a)], implying that network homogeneity is beneficial to reconstruction. The results in Figs. 4(b), 5(b) and 6(b) suggest that the reconstruction performance tends to decrease with the average degree of networks, indicating that more signals are needed for networks with denser connections [41].

We further test our method using four empirical networks. Figures 7(a-d) show TPR and FPR versus M for Karate, Dolphins, Football, and Email networks, respectively. In each case, SIR, rumor, and mixed spreading dynamics are tested. We see that, regardless of the type of the spreading dynamics, TPR increases and FPR decreases with M. For the Karate, Dolphins, and Football networks, as M is increased, the values of TPR and FPR converge to unity and zero, respectively, as shown in Figs. 7(a-c). However, for the Email network, the value of TPR can reach about 90%, as shown in Fig. 7(d).

The relatively poor reconstruction performance with the Email network can be explained, as follows. In general, we find that the reconstruction accuracy is reduced when the network is more dense and/or more heterogeneous. As shown in Table I, the average degree of the Email network is about 10. In addition, the degree distribution of the network is quite heterogeneous in that there are hub nodes. The Email network is thus relatively more dense and heterogeneous than the other three empirical networks, leading to a relatively poor reconstruction performance. However, the result in Fig. 8(d) indicates that the reconstruction performance can be greatly improved once the first arrival times are used.

B. Performance enhancement with temporal information

The results discussed so far have demonstrated reconstruction of network based solely on the final states. Intuitively, the availability of limited temporal information can facilitate reconstruction and improve accuracy. To demonstrate this, we consider the concrete case where, in addition to the signal received by each node, its first arrival time is also available. A new data matrix ST can then be constructed, where ST_{ij} specifies the first arrival time of signal i received by node j, otherwise, $ST_{ij} = 0$ if signal i never reaches node j. The task is to extract the local information matrix S^j associated with node j from ST. If node j receives signal i at time ST_{ij}, there exists a node that received signal i during the time interval $[ST_{ij} - t_0, ST_{ij}]$ and then passed the signal on to node j, where t_0 is a transmission cycle of the disease. For example, in the SIR model, t_0 is the time interval for transition from an infected state to a recovered state. If the value of t_0 is not known *a priori*, we set $t_0 = +\infty$. In general, the transmission time is short if node j transmits signal i to a neighboring node k. We can then obtain the matrix S^j through a parameter τ:

$$S^j_{ik} = \begin{cases} 1, & ST_{ij} - t_0 \le ST_{ik} \le ST_{ij} + \tau, ST_{ij} \cdot ST_{ik} \neq 0; \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

(We set $\tau = 1$. In fact, the reconstruction performance is slightly reduced when the value of τ is increased.) For either SIR or rumor model with the recovery rate $\mu = 1$, we have $t_0 = 1$. Equation (4) becomes

$$S^j_{ik} = \begin{cases} 1, & |ST_{ij} - ST_{ik}| \le 1,\ ST_{ij} \cdot ST_{ik} \neq 0; \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Once the matrix S^j is constructed, the neighbors of node j can be inferred. Figure 8 shows that better reconstruction performance can be achieved when incorporating the first arrival time, and the number M of signals required can be significantly reduced. Take the Email network as an example [Fig. 8(d)]. Without the time information, the value of TPR is about 0.9 even with $M = 100000$. However, when first arrival time information is used, the value of TPR can exceed 0.99 for $M = 36000$.

If only partial time information is available, reconstruction performance can still be improved. In this case, for signals

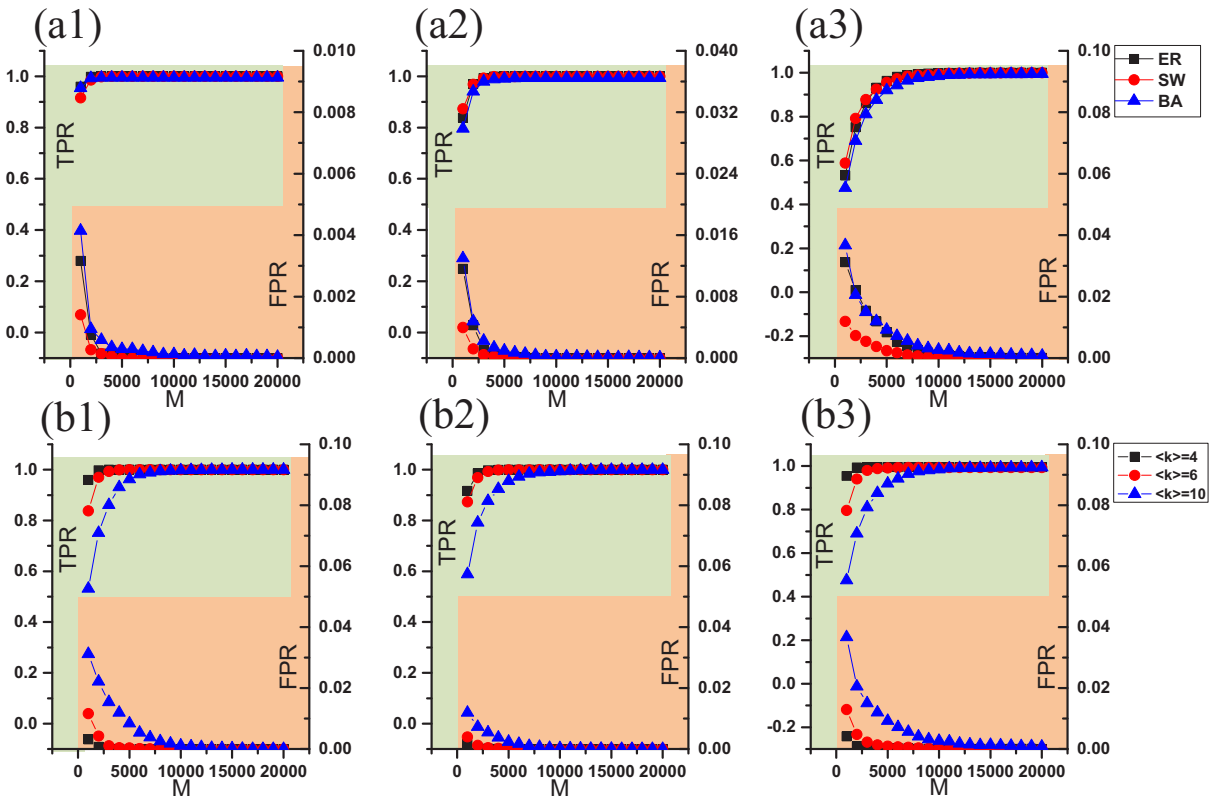

FIG. 4: Effects of the number of signals on reconstruction performance for SIR dynamics. For three types of networks: ER, SW and BA, TPR (left ordinate) and FPR (right ordinate) versus M. The top panels (a1-a3) are for ER networks with $\langle k\rangle = 4$, 6 and 10, respectively, and the bottom panels (b1-b3) are for ER, SW, and SF networks, respectively. The network size is $N = 100$ for all cases.

with which the first arrival times are known, the elements of S^j can be determined through Eq. (5). For other signals for which no such information is available, the matrix elements can be extracted from the global information matrix S. In general, more elements of S^j as determined from Eq. (5) leads to higher reconstruction precision.

C. Uniform final states

For either the SIR epidemic model or the rumor spreading model, the final nodal state for each node has two distinct states (binary states), making it possible to reconstruct the network based on the final states. Nevertheless, for dynamical processes with a single or uniform final state, the framework based solely on the final states is ineffective. An example is the susceptible-infected (SI) epidemic process, for which all nodes are finally infected. However, if the first arrival time of each signal is known, network can still be reconstructed through the local information matrix S^j according to Eq. (4). For the SI model, a unique feature is $t_0 = +\infty$, since no infected nodes can be recovered. We have

$$S_{ik}^j = \begin{cases} 1, & 0 \le ST_{ik} \le ST_{ij} + \tau,\ ST_{ij} \cdot ST_{ik} \neq 0; \\ 0, & \text{otherwise}, \end{cases} \tag{6}$$

where $\tau = 1$. To give a concrete example, we choose the transmission rate β randomly from the interval $[0.3, 0.7]$ and record the first arrival time to get the information matrix ST. The reconstruction results are shown in Fig. 9. A general behavior similar to the cases of SIR or rumor dynamics (Figs. 4-6) is that the reconstruction accuracy can be improved by increasing the number of signal M. However, as shown in Figs. 9(a1-a3), network heterogeneity has little effect on the reconstruction accuracy, which is different from the case of binary final states. This is due to the basic fact that for the SI dynamics, all nodes are finally infected, regardless of the network structure. As the number of edges is increased, the accuracy tends to decrease unless more signals are collected, as shown in Figs. 9(b1-b3).

D. Effects of random signal disturbance

We investigate how random signal disturbances affect reconstruction. Let SN be the number of "1" in the information matrix S. We consider two types of disturbance: (1) replacing $SN \cdot \rho$ of "0" elements by "1" so that the number of false signals ($0 \to 1$) is larger than that of true signals and (2) replacing $SN \cdot \rho$ of "1" elements by "0" so that some signals are not collected (denoted as $1 \to 0$). Disturbance of the first type can cause a pair of originally unconnected nodes to be

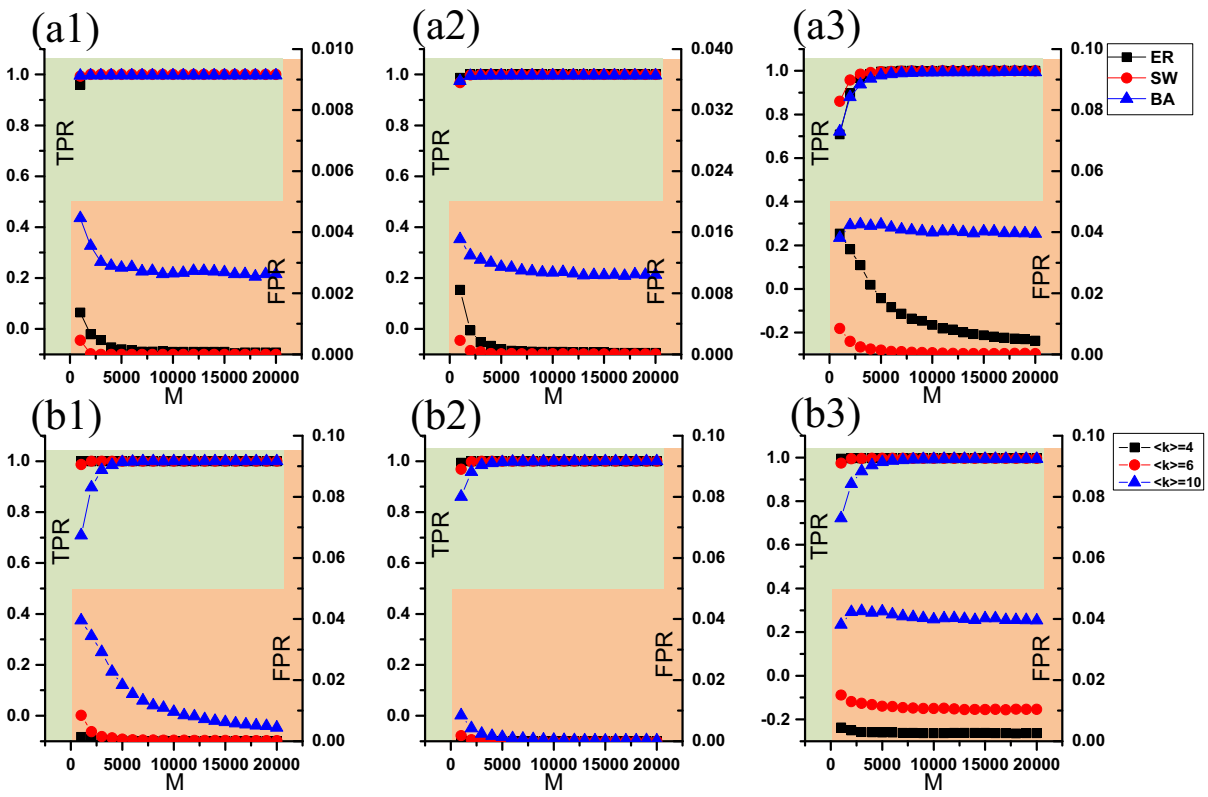

FIG. 5: Effects of the number of signals on reconstruction performance for rumor spreading dynamics. Same as Fig. 4 except that the dynamics is of the rumor spreading type.

regarded as connected. Numerically, we assume that there is a connection between nodes j and k when P_k^j is larger than a small threshold, e.g., $1/N$. To characterize the effects of signal disturbances on reconstruction accuracy, we use F1-score index. Setting $\rho = 10\%$, we calculate the F1-score for the two types of disturbance for ER, SW and BA networks. As shown in Fig. 10, the disturbance type "$0 \to 1$" ($A_{0\to1}$) has a destructive effect on the accuracy as the values of $F1$ are close to zero. This is due to the fact that false signals can lead to wrong prediction of non-existent links. However, the "$1 \to 0$" ($A_{1\to0}$) type of disturbance has a small effect on the reconstruction as the $F1$ assumes high values. In this case, even there are missing signals, the links can be predicted for sufficiently large values of M. An implication is that certain signals can be discarded without severely affecting the reconstruction performance.

To mitigate the effect of random signal disturbance, we break the information matrix into several blocks, assuming that a large number of signals are available, where each block can be used to reconstruct a pseudo-network and there is an edge linking two nodes if they are deemed connected in several pseudo-networks. For instance, we can use 5000 signals to construct each pseudo-network and regard two nodes as connected if the probability of being linked in different pseudo-networks is larger than 80%. There is great flexibility in choosing the blocks of signals and two sets of signals can be overlapped. For example, for $M = 8000$, the first 5000 signals and the last 5000 signals can be chosen as two blocks. As shown in Fig. 10, we see that this approach can lead to a significant increase in the value of $F1$ even in the presence of a large number of false signals (denoted as $B_{0\to1}$) or missing signals (labeled as $B_{1\to0}$). In particular, $F1$ can approach unity even for the case of "$0 \to 1$." Figure 11 demonstrates that, for fixed $M = 25000$, the value of $F1$ decreases with ρ, but the method of signal blocks is still effective at mitigating the effects of random signal disturbances on reconstruction.

V. DISCUSSION

Inferring complex network topologies from data is a problem of current interest. In previous works time series were required for this reverse engineering problem [5–43]. There are real world applications, e.g., spreading dynamics on complex networks, in which time series are not available, raising the question of whether the network structure can be identified in such circumstances. This paper provides an affirmative answer. In particular, we develop a general framework to infer the network topology using information about the final states of the network only. Collecting an ensemble of binary final states originated from independent triggers (or signals) of the spreading dynamics and exploiting the principle of maximum likelihood estimation, we obtain rigorous mathematical results establishing the feasibility of accurate reconstruction

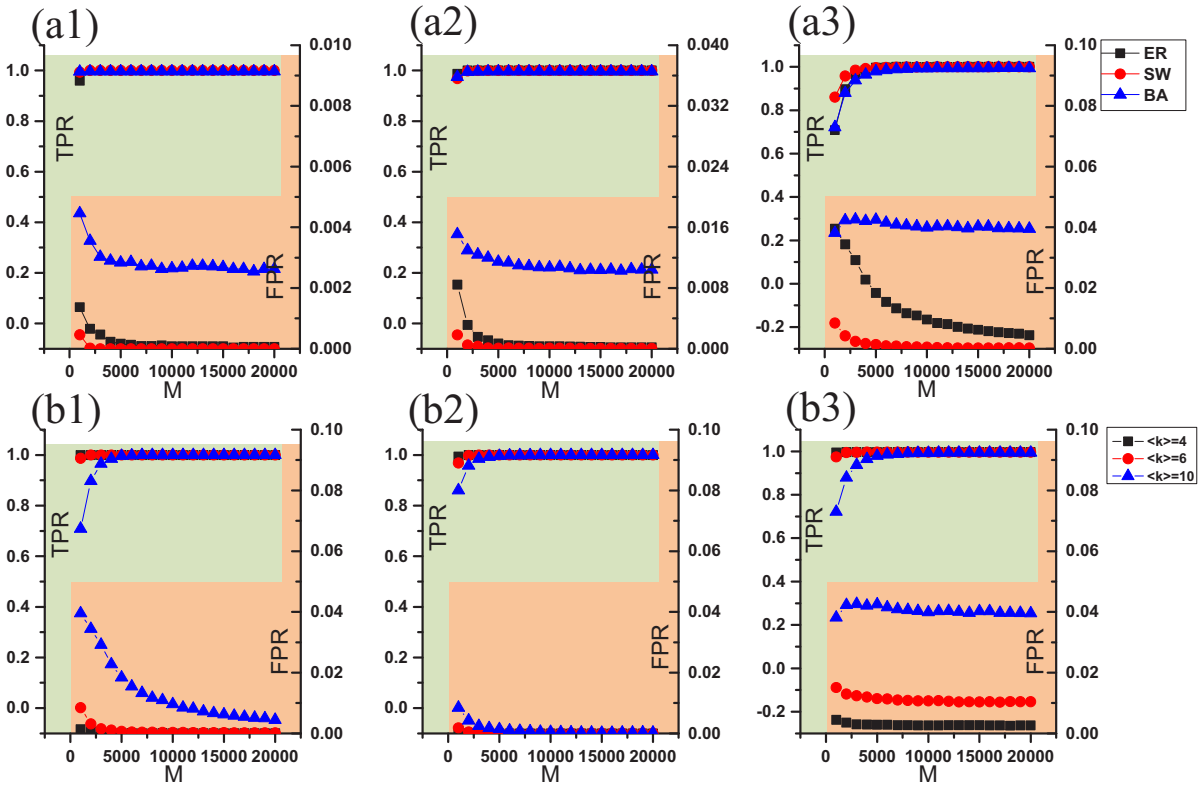

FIG. 6: Effects of the number of signals on reconstruction performance for mixed SIR and rumor spreading dynamics. Same as Fig. 4 except that the dynamics is of the mixed type.

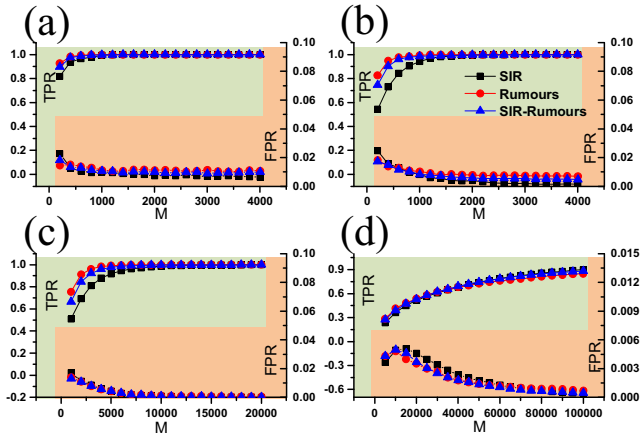

FIG. 7: Performance of our method for empirical networks. (a-d) For four empirical networks (Karate, Dolphins, Football, and Email, respectively), TPR and FPR versus M. In each case, results for SIR, rumor, and mixed spreading dynamics are displayed.

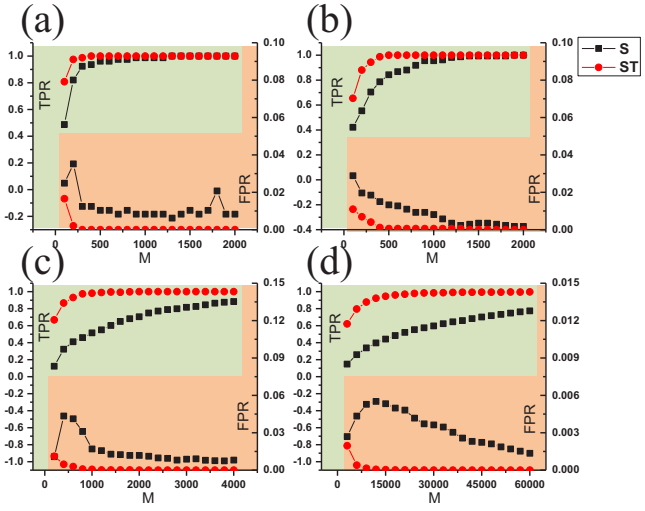

FIG. 8: Network reconstruction with first arrival time information. For four empirical networks, TPR (left ordinate) and FPR (right ordinate) versus M when first arrival time information is incorporated into the reconstruction framework. Comparing with the case where no such information is available, we see that TPR is increased and FPR is decreased, and fewer number of signals are required for high-accuracy reconstruction. The network dynamics is assumed to be of the SIR type. (a) Karate network, (b) Dolphins network, (c) Football network, and (d) Email network.

of the network. We demonstrate the working of our method and quantify its accuracy using a large number of model and empirical networks. For spreading processes with a uniform final state, e.g., SI spreading dynamics, the underlying network can still be reconstructed using certain temporal information about the dynamical process, such as the first arrival

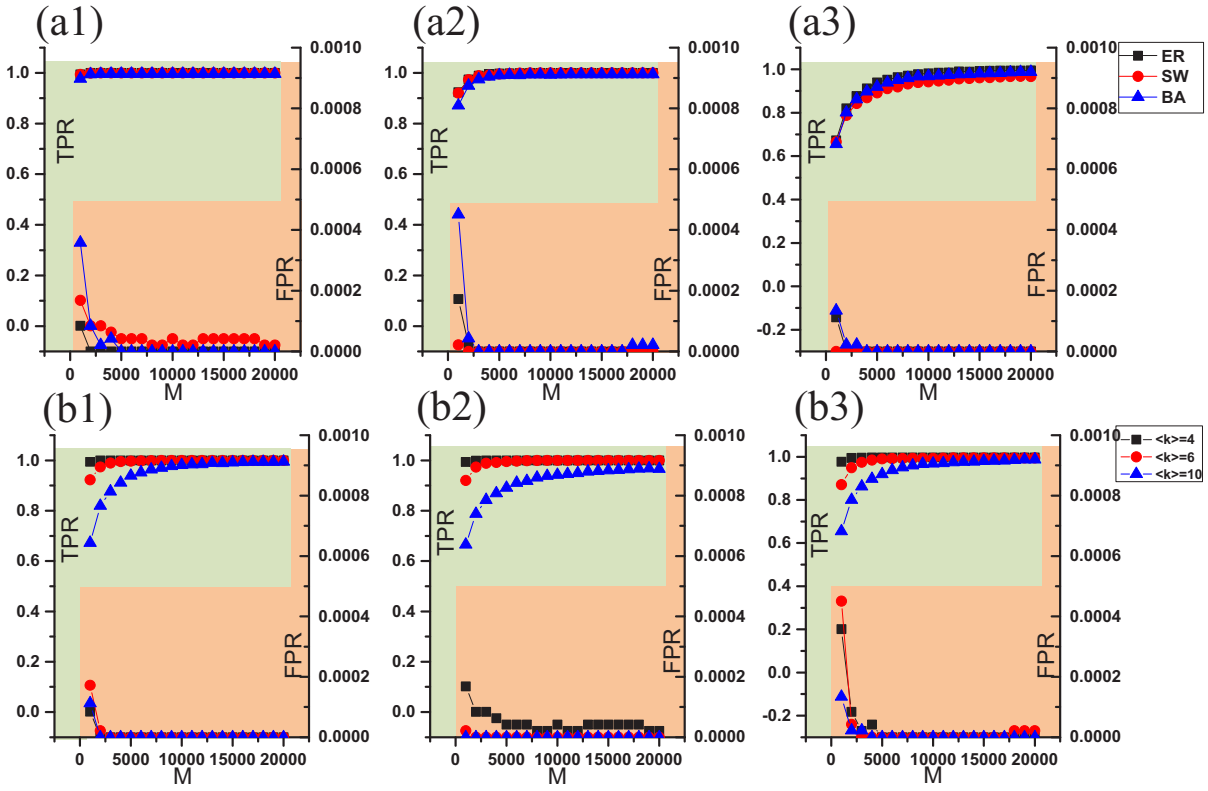

FIG. 9: Network reconstruction with SI epidemic dynamics. For SI dynamics, the information matrix ST is obtained by recording the first arrival time of each signal at each node. Shown are TPR (left vertical axis) and FPR (right vertical axis) versus the number M of signals for three distinct types of model complex networks: ER, SW, and BA. The top panels (a1-a3) are for $\langle k\rangle = 4$, 6 and 10, respectively, while the bottom panels (b1-b3) correspond to ER, SW, and BA networks of fixed average degree, respectively.

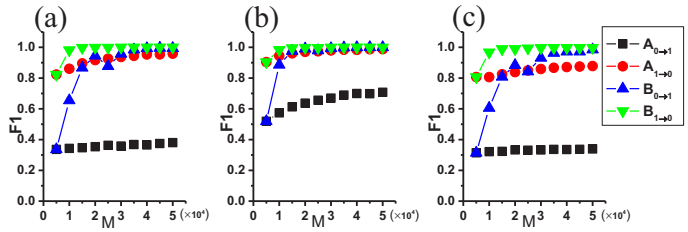

FIG. 10: **Mitigation of random signal disturbances using sub-blocks of signals**. For fixed $\rho = 10\%$, F1 score versus the number of signals M: (a-c) ER, SW, and BA networks, respectively, where the symbol $A_{0\to1}$ ($A_{1\to0}$) denotes that "0" is wrongly regarded as "1" (vice versa) in the unblocked information matrix S, and $B_{0\to1}$ ($B_{1\to0}$) is for the case of blocked matrix. Each pseudo-matrix is constructed with 5000 signals. The network size is $N = 100$ and the average degree is $\langle k\rangle = 6$. The dynamical process is of the SIR type.

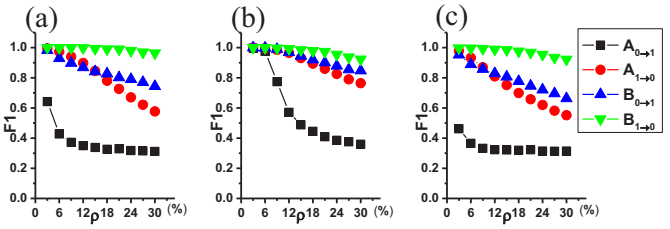

FIG. 11: **Mitigation of random signal disturbances using sub-blocks of signals**: F1 score versus ρ for $M = 25000$ for (a-c) ER, SW, and BA networks, respectively. Legends and parameters are the same as for Fig. 10.

time of a signal at each node. At the present, our method cannot be applied to non-Markovian dynamics due to the time memory and social reinforcement effects [64–66].

The basic philosophy underlying our framework is the principle of ergodicity in statistical and nonlinear physics: for a complex system in the thermodynamical limit the time and ensemble averages are equivalent. That is, when time series are not available, it is possible to use ensemble information about the asymptotic state of the system for reconstruction. Our work is thus demonstration of this universal principle in the specific context of reverse engineering of complex networks. One issue associated with our reconstruction framework is that the number of independent signals required for accurate reconstruction is large. Acquiring additional information, such as the first arrival time of a signal at each node, can reduce the number of signals markedly without compromising accuracy. For large networks, suitable dimension reduction procedures can be used to ensure reconstruction accu-

racy. To articulate methods to significantly reduce the ensemble size is a problem worth further studies.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant Nos. 61473001 and 11331009. YCL would like to acknowledge support from the Vannevar Bush Faculty Fellowship program sponsored by the Basic Research Office of the Assistant Secretary of Defense for Research and Engineering and funded by the Office of Naval Research through Grant No. N00014-16-1-2828.

Appendix A: Proof of Theorem 1

Theorem 1: If $S_{ia}^j \geq S_{ib}^j \left(i = 1, 2, \cdots, M^j\right)$ and $\sum\limits_i S_{ia}^j > \sum\limits_i S_{ib}^j$, then one has $P_b^j = 0$ from Eq. (2).

Proof: Without loss of generality, we set $a = 1$ and $b = 2$. Assume that $P^j = \left[P_a^j, P_b^j, P_3^j, \cdots, P_{N^j}^j\right]^T$ is the optimal solution of Eq. (2), where $P_b^j = \beta > 0$ and $P_a^j = \alpha$. If we set $\tilde{P}_a^j = \alpha + \beta$ and $\tilde{P}_b^j = 0$, we have that $\tilde{P}^j = \left[\tilde{P}_a^j, \tilde{P}_b^j, P_3^j, \cdots, P_{N^j}^j\right]^T$ is also a solution of Eq. (2). We then have

$$\begin{aligned} f\left(\tilde{P}^j\right) &= \prod_i \left(\tilde{P}_a^j S_{ia}^j + \tilde{P}_b^j S_{ib}^j + \sum_{k=3}^{N^j} P_k^j S_{ik}^j + \epsilon\right) \\ &= \prod_i \left(\alpha S_{ia}^j + \beta S_{ia}^j + \sum_{k=3}^{N^j} P_k^j S_{ik}^j + \epsilon\right) \\ &> \prod_i \left(\alpha S_{ia}^j + \beta S_{ib}^j + \sum_{k=3}^{N^j} P_k^j S_{ik}^j + \epsilon\right) \\ &= \prod_i \left(P_a^j S_{ia}^j + P_b^j S_{ib}^j + \sum_{k=3}^{N^j} P_k^j S_{ik}^j + \epsilon\right) \\ &= f\left(P^j\right). \end{aligned}$$

The inequality indicates that P^j is not the optimal solution of Eq. (2), contradicting the original hypothesis. As a result, we must have $P_b^j = 0$.

Appendix B: Proof of Theorem 2

Theorem 2: Suppose the matrix S^j has three columns only which satisfy the following conditions: (1) $S_{i1}^j + S_{i2}^j \geq S_{i3}^j \left(i = 1, 2, \cdots, M^j\right)$, (2) $\sum\limits_i S_{i1}^j > \sum\limits_i S_{i3}^j$, and (3) $\sum\limits_i S_{i2}^j > \sum\limits_i S_{i3}^j$. For $\epsilon \to 0$, the formula (2) can be written as:

$$\begin{aligned} &\max f(P_1^j, P_2^j, P_3^j) \\ =\ & (P_1^j + \epsilon)^{n_1} (P_2^j + \epsilon)^{n_2} (P_1^j + P_2^j + \epsilon)^{n_3} \\ & (P_1^j + P_3^j + \epsilon)^{n_4} (P_2^j + P_3^j + \epsilon)^{n_5} \\ & (P_1^j + P_2^j + P_3^j + \epsilon)^{n_6}, \end{aligned} \tag{B1}$$

where $P_1^j + P_2^j + P_3^j = 1$ and $P_1^j \geq 0, P_2^j \geq 0, P_3^j \geq 0$. We then have $P_3^j = 0$ if one of two conditions holds: (1) $n_1 \geq n_5$ and $n_2 \geq n_4$, or (2) $n_5/(n_1 + n_3) + n_4/(n_2 + n_3) \leq 1$.

Proof: Firstly, we prove $P_1^j \neq 0$ and $P_2^j \neq 0$ if $P^j = \left[P_1^j, P_2^j, P_3^j\right]^T$ is the optimal solution of Eq. (2). To prove $P_1^j \neq 0$, we note that, if $n_1 \neq 0$, we have $P_1^j \neq 0$. Otherwise, the following equation indicates that $f\left(0, P_2^j, P_3^j\right)$ is not the maximum value:

$$f(0, P_2^j, P_3^j) = (\epsilon)^{n_1} (P_2^j + \epsilon)^{n_2} (P_2^j + P_3^j + \epsilon)^{n_3} (P_2^j + P_3^j + \epsilon)^{n_4} (P_2^j + P_3^j + \epsilon)^{n_5} (P_2^j + P_3^j + \epsilon)^{n_6} = 0. \tag{B2}$$

If $n_1 = 0$, then Eq. (B1) can be rewritten as (for $\epsilon \to 0$)

$$\max\ f\left(P_1^j, P_2^j, P_3^j\right) = \left(P_2^j + \epsilon\right)^{n_2} \left(P_1^j + P_2^j + \epsilon\right)^{n_3} \left(P_1^j + P_3^j + \epsilon\right)^{n_4} \left(P_2^j + P_3^j + \epsilon\right)^{n_5}, \tag{B3}$$

where $P_1^j + P_2^j + P_3^j = 1$, $P_2^j \geq 0$, and $P_3^j \geq 0$. Assume $P^j = [0, P_2, P_3]^T$ is the optimal solution of Eq. (B3), i.e.,

$$\max\ f\left(P_1^j, P_2^j, P_3^j\right) = f\left(0, P_2, P_3\right) = (P_2 + \epsilon)^{n_2 + n_3} (P_3 + \epsilon)^{n_4}, \tag{B4}$$

where $P_2 + P_3 = 1$, $P_2 > 0$ and $P_3 > 0$ [the maximum value of Eq. (B4) is zero for $P_2 = 0$ or $P_3 = 0$]. The inequality $\sum_i S_{i1}^j > \sum_i S_{i3}^j$ leads to $n_1 + n_3 + n_4 + n_6 > n_4 + n_5 + n_6$. As a result, we have $n_3 > n_5$ owing to $n_1 = 0$.

Moreover, $P_1^j = P_3$, $P_2^j = P_2$ and $P_3^j = 0$ imply the conditions: $P_1^j + P_2^j + P_3^j = 1$, $P_2^j \geq 0$ and $P_3^j \geq 0$. In this case, we have $P^j = [P_3, P_2, 0]^T$ and

$$f(P_3, P_2, 0) = (P_2 + \epsilon)^{n_2+n_5}(P_3 + \epsilon)^{n_4} > (P_2 + \epsilon)^{n_2+n_3}(P_3 + \epsilon)^{n_4} = \max f\left(P_1^j, P_2^j, P_3^j\right), \tag{B5}$$

which contradicts the original hypothesis, so we must have $P_1^j \neq 0$. The conclusion $P_2^j \neq 0$ can be proved in a similar manner.

Next, we show that $P_3^j = 0$. Note that, for $n_4 = 0$ or $n_5 = 0$, Theorem 2 is a direct consequence of Theorem 1. For $n_4 \neq 0$ and $n_5 \neq 0$, Eq. (B1) can be written as

$$g(x, y) = (x + \epsilon)^{n_1}(y + \epsilon)^{n_2}(x + y + \epsilon)^{n_3}(1 - y + \epsilon)^{n_4}(1 - x + \epsilon)^{n_5}, \tag{B6}$$

where $x + y \leq 1$, $x \geq 0$, and $y \geq 0$. We thus have $P_1^j = x$, $P_2^j = y$ and $P_3^j = 1 - x - y$. Within the bounded region $x + y \leq 1$, $x \geq 0$ and $y \geq 0$, $g(x, y)$ is a continuous function so it must have a maximum value. The possible maximum value is obtained when the variables x and y are on the borders ($x = 0$ or $y = 0$ or $x+y = 1$) or are at a stationary point. Note that that the maximum value of Eq. (B6) cannot be obtained for $P_1^j = x = 0$ or $P_2^j = y = 0$.

We now prove that the maximum value of $g(x, y)$ is achieved for $x + y = 1$. [In fact, $g(x, y)$ does not possess any stationary point]. Letting

$$\begin{cases} g_x{}' = 0; \\ g_y{}' = 0, \end{cases}$$

we have

$$\begin{cases} n_1(x + y + \epsilon)(1 - x + \epsilon) + n_3(x + \epsilon)(1 - x + \epsilon) - n_5(x + \epsilon)(x + y + \epsilon) = 0; \\ n_2(x + y + \epsilon)(1 - y + \epsilon) + n_3(y + \epsilon)(1 - y + \epsilon) - n_4(y + \epsilon)(x + y + \epsilon) = 0. \end{cases} \tag{B7}$$

Note that the inequality $\sum_i S_{i1}^j > \sum_i S_{i3}^j$ gives rise to $n_1 + n_3 > n_5 > 0$. However, the inequality $\sum_i S_{i2}^j > \sum_i S_{i3}^j$ leads to $n_2 + n_3 > n_4 > 0$. From Eq. (B7), we have

$$\begin{cases} (1 - x + \epsilon) = \frac{n_5(x+\epsilon)(x+y+\epsilon)}{n_1(x+y+\epsilon)+n_3(x+\epsilon)}, \\ (1 - y + \epsilon) = \frac{n_4(y+\epsilon)(x+y+\epsilon)}{n_2(x+y+\epsilon)+n_3(y+\epsilon)}. \end{cases} \tag{B8}$$

For $n_1 > n_5 > 0$ and $n_2 > n_4 > 0$, we obtain the following inequality from Eq. (B8):

$$\begin{cases} (1 - x + \epsilon) \leq \frac{n_5(x+\epsilon)}{n_1} \leq (x + \epsilon); \\ (1 - y + \epsilon) \leq \frac{n_4(y+\epsilon)}{n_2} \leq (y + \epsilon). \end{cases} \tag{B9}$$

Combining the two inequalities, we get $1 \leq x + y$, implying that $g(x, y)$ does not possess any stationary point for $x+y < 1$ and $x \geq 0, y \geq 0$. As a result, Eq. (B6) has a maximum value for $x + y = 1$, namely, $P_1^j + P_2^j = 1$ and $P_3^j = 0$.

From Eq. (B8), we get

$$\begin{cases} (1 - x + \epsilon) \leq \frac{n_5(x+y+\epsilon)}{n_1+n_3}; \\ (1 - y + \epsilon) \leq \frac{n_4(x+y+\epsilon)}{n_2+n_3}, \end{cases} \tag{B10}$$

which further implies the following inequality:

$$2 - x - y + 2\epsilon \leq \left(\frac{n_4}{n_2 + n_3} + \frac{n_5}{n_1 + n_3}\right)(x + y + \epsilon) \leq x + y + \epsilon. \tag{B11}$$

Under the condition $n_5/(n_1+n_3)+n_4/(n_2+n_3)\leq 1$, inequality (B11) can be further simplified as: $1+\epsilon/2\leq x+y$, which indicates that $g(x,y)$ attains its maximum value for $x+y=1$. That is, we have $P_1^j+P_2^j=1$ and $P_3^j=0$.

Appendix C: Generalization of theorem 2

The prerequisite of Theorem 2 is that there are only three columns in S_j, so it does not apply to cases where the number of columns is larger than 3. Nonetheless, it is useful to test numerically whether Theorem 2 can be generalized. Taking the Email network as an example, we compare the reduction based on Theorem 1 (denoted as S1) and that based on Theorem 2 (denoted as S2), as shown in Fig. 12. We find that S2 does not reduce the accuracy even though the matrix S^j has many columns. Consequently, we can use Theorem 2 on an empirical basis to further reduce the number of unknown parameters in Eq. (2).

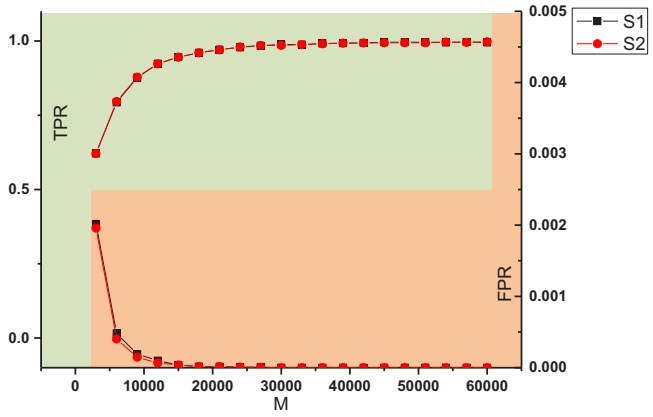

FIG. 12: **Numerical test of generalization of Theorem 2**. For the Email network, we implement SIR dynamics and collect the first-arrival times of signals. The left and right vertical axes represent TPR and FPR, respectively. S1 and S2 denote the reductions based on Theorem 1 and Theorem 2, respectively. The transmission rate β is randomly chosen from the interval $[0, 0.3]$, and the recovery rate is $\mu = 1.0$. The results are averaged over 10 independent runs.

[1] W.-X. Wang, Y.-C. Lai, and C. Grebogi, Phys. Rep. **644**, 1 (2016).
[2] E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman, Nat. Gene. **34**, 166 (2003).
[3] A.-L. Barabasi and Z. N. Oltvai, Nat. Rev. Gene. **5**, 101 (2004).
[4] P. L. van Geert and H. W. Steenbeek, Behav. Brain Sci. **33**, 174 (2010).
[5] S. Gruen, M. Diesmann, and A. Aertsen, Neu. Comp. **14**, 43 (2002).
[6] R. Gütig, A. Aertsen, and S. Rotter, Neu. Comp. **14**, 121 (2002).
[7] T. S. Gardner, D. di Bernardo, D. Lorenz, and J. J. Collins, Science **301**, 102 (2003).
[8] G. Pipa and S. Grün, Neurocomp. **52**, 31 (2003).
[9] A. Brovelli, M. Ding, A. Ledberg, Y. Chen, R. Nakamura, and S. L. Bressler, Proc. Nat. Acad. Sci. (USA) **101**, 9849 (2004).
[10] D. Yu, M. Righero, and L. Kocarev, Phys. Rev. Lett. **97**, 188701 (2006).
[11] J. Bongard and H. Lipson, Proc. Natl. Acad. Sci. (USA) **104**, 9943 (2007).
[12] M. Timme, Phys. Rev. Lett. **98**, 224101 (2007).
[13] W. K.-S. Tang, M. Yu, and L. Kocarev, in *Circuits and Systems, ISCAS 2007. IEEE Inter. Symp.* (IEEE, 2007), pp. 2646–2649.
[14] D. Napoletani and T. D. Sauer, Phys. Rev. E **77**, 026103 (2008).
[15] E. Sontag, Essays Biochem. **45**, 161 (2008).

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

[17] W.-X. Wang, Q.-F. Chen, L. Huang, Y.-C. Lai, and M. Harrison, Phys. Rev. E **80**, 016116 (2009).

[18] J. Donges, Y. Zou, N. Marwan, and J. Kurths, EPL (Europhys. Lett.) **87**, 48007 (2009).

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

[20] J. Chan, A. Holmes, and R. Rabadan, PLoS Comp. Bio. **6**, e1001005 (2010).

[21] Y. Yuan, G.-B. Stan, S. Warnick, and J. Goncalves, in *Decision and Control (CDC), 2010 49th IEEE Conference* (IEEE, 2010), pp. 810–815.

[22] Z. Levnajić and A. Pikovsky, Phys. Rev. Lett. **107**, 034101 (2011).

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

[24] S. G. Shandilya and M. Timme, New J. Phys. **13**, 013004 (2011).

[25] D. Yu and U. Parlitz, PLOS ONE **6**, e24333 (2011).

[26] W.-X. Wang, Y.-C. Lai, C. Grebogi, and J.-P. Ye, Phys. Rev. X **1**, 021021 (2011).

[27] W.-X. Wang, R. Yang, Y.-C. Lai, V. Kovanis, and C. Grebogi, Phys. Rev. Lett. **106**, 154101 (2011).

[28] W.-X. Wang, R. Yang, Y.-C. Lai, V. Kovanis, and M. A. F. Harrison, EPL (Europhys. Lett.) **94**, 48006 (2011).

[29] R. Yang, Y.-C. Lai, and C. Grebogi, Chaos **22**, 033119 (2012).

[30] W. Pan, Y. Yuan, and G.-B. Stan, in *Decision and Control (CDC), 2012 IEEE 51st Annu. Conf.* (IEEE, 2012), pp. 2334–2339.

[31] W.-X. Wang, J. Ren, Y.-C. Lai, and B. Li, Chaos **22**, 033131 (2012).

[32] T. Berry, F. Hamilton, N. Peixoto, and T. Sauer, J. Neurosci. Meth. **209**, 388 (2012).

[33] O. Stetter, D. Battaglia, J. Soriano, and T. Geisel, PLoS Comp. Biol. **8**, e1002653 (2012).

[34] R.-Q. Su, X. Ni, W.-X. Wang, and Y.-C. Lai, Phys. Rev. E **85**, 056220 (2012).

[35] R.-Q. Su, W.-X. Wang, and Y.-C. Lai, Phys. Rev. E **85**, 065201 (2012).

[36] F. Hamilton, T. Berry, N. Peixoto, and T. Sauer, Phys. Rev. E **88**, 052715 (2013).

[37] D. Zhou, Y. Xiao, Y. Zhang, Z. Xu, and D. Cai, Phys. Rev. Lett. **111**, 054102 (2013).

[38] M. Timme and J. Casadiego, J. Phys. A. Math. Theo. **47**, 343001 (2014).

[39] R.-Q. Su, Y.-C. Lai, and X. Wang, Entropy **16**, 3889 (2014).

[40] R.-Q. Su, Y.-C. Lai, X. Wang, and Y.-H. Do, Sci. Rep. **4**, 3944 (2014).

[41] Z. Shen, W.-X. Wang, Y. Fan, Z. Di, and Y.-C. Lai, Nat. Commun. **5**, 4323 (2014).

[42] R.-Q. Su, W.-W. Wang, X. Wang, and Y.-C. Lai, R. Soc. Open Sci. **3**, 150577 (2016).

[43] J. Li, Z. Shen, W.-X. Wang, C. Grebogi, and Y.-C. Lai, Phys. Rev. E **95**, 032303 (2017).

[44] V. M. Eguiluz, D. R. Chialvo, G. A. Cecchi, M. Baliki, and A. V. Apkarian, Phys. Rev. Lett. **94**, 018102 (2005).

[45] D. S. Bassett, A. Meyer-Lindenberg, S. Achard, T. Duke, and E. Bullmore, Proc. Nat. Acad. Sci. (USA) **103**, 19518 (2006).

[46] U. Parlitz, Phys. Rev. Lett. **76**, 1232 (1996).

[47] F. Varela, J.-P. Lachaux, E. Rodriguez, and J. Martinerie, Nat. Rev. Neurosci. **2**, 229 (2001).

[48] D. Zhou, Y. Xiao, Y. Zhang, Z. Xu, and D. Cai, Phys. Rev. Lett. **111**, 054102 (2013).

[49] L. Huang, K. Park, and Y.-C. Lai, Phys. Rev. E **73**, 035103 (2006).

[50] I. J. Myung, J. Math. Psycho. **47**, 90 (2003).

[51] Y. Moreno, R. Pastor-Satorras, and A. Vespignani, Euro. Phys. J. B **26**, 521 (2002).

[52] Y. Moreno, M. Nekovee, and A. F. Pacheco, Phys. Rev. E **69**, 066130 (2004).

[53] J. Borge-Holthoefer and Y. Moreno, Phys. Rev. E **85**, 026116 (2012).

[54] M. Girvan and M. E. Newman, Proc. Nat. Acad. Sci. (USA) **99**, 7821 (2002).

[55] D. Lusseau and M. E. Newman, Proc. Roy. Soc. London B Biol. Sci. **271**, S477 (2004).

[56] M. E. Newman, Phys. Rev. E **69**, 066133 (2004).

[57] R. Guimera, L. Danon, A. Diaz-Guilera, F. Giralt, and A. Arenas, Phys. Rev. E **68**, 065103 (2003).

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

[59] X. Han, Z. Shen, W.-X. Wang, and Z. Di, Phys. Rev. Lett. **114**, 028701 (2015).

[60] D. L. Olson and D. Delen, *Advanced Data Mining Techniques* (Springer Science & Business Media, 2008).

[61] P. Erdös and A. Rényi, Publ. Math. Inst. Hung. Acad. Sci **5**, 43 (1960).

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

[63] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).

[64] D. Centola, Science **329**, 1194 (2010).

[65] I. Z. Kiss, G. Röst, and Z. Vizi, Phys. Rev. Lett. **115**, 078701 (2015).

[66] W. Wang, M. Tang, H.-F. Zhang, and Y.-C. Lai, Phys. Rev. E **92**, 012820 (2015).