Locating the source of diffusion in complex networks by time-reversal backward spreading

Zhesi Shen,[1] Shinan Cao,[2,*] Wen-Xu Wang,[1,3,†] Zengru Di,[1] and H. Eugene Stanley[4]
[1]*School of Systems Science, Beijing Normal University, Beijing, 100875, P. R. China*
[2]*School of Finance, University of International Business and Economics, Beijing, 100029, P. R. China*
[3]*Business School, University of Shanghai for Science and Technology, Shanghai 200093, China*
[4]*Department of Physics, Boston University, Boston, Massachusetts 02215, USA*
(Received 4 October 2015; published 2 March 2016)

Locating the source that triggers a dynamical process is a fundamental but challenging problem in complex networks, ranging from epidemic spreading in society and on the Internet to cancer metastasis in the human body. An accurate localization of the source is inherently limited by our ability to simultaneously access the information of all nodes in a large-scale complex network. This thus raises two critical questions: how do we locate the source from incomplete information and can we achieve full localization of sources at any possible location from a given set of observable nodes. Here we develop a time-reversal backward spreading algorithm to locate the source of a diffusion-like process efficiently and propose a general locatability condition. We test the algorithm by employing epidemic spreading and consensus dynamics as typical dynamical processes and apply it to the H1N1 pandemic in China. We find that the sources can be precisely located in arbitrary networks insofar as the locatability condition is assured. Our tools greatly improve our ability to locate the source of diffusion in complex networks based on limited accessibility of nodal information. Moreover, they have implications for controlling a variety of dynamical processes taking place on complex networks, such as inhibiting epidemics, slowing the spread of rumors, pollution control, and environmental protection.

DOI: 10.1103/PhysRevE.93.032301

I. INTRODUCTION

Many large-scale dynamical processes taking place on complex networks can be triggered from a small number of nodes. Prototypical examples include epidemic spreading on a global scale, rumor propagation through microblogs on the internet, wide-ranging blackouts across North America, and financial crises accompanied by the bankruptcy of a large number of financial institutions. The self-organization theory introduced by Bak and his collaborators [1] has provided a theoretical explanation: when a complex system enters a self-organized criticality state, small perturbations to even single individuals are able to initiate a big event, such as the avalanche of collapses in the sandpile model [2]. Moreover, the development of modern technology considerably facilitates the spreading of disease and information via public traffic systems and the internet, which enables propagation across a large area from a source, such as the worldwide H1N1 pandemic in 2009 [3,4] and the irrational and panicked acquisition of salt in southeast Asian countries caused by a rumor relevant to the nuclear leak in Japan. These phenomena raise a challenging question: how to locate the source in a huge network relying on relatively limited accessibility to nodal states, answers to which are of paramount importance for many aspects of nature and society, such as disease control, antiterrorism, and economic health. Despite some pioneering approaches attempting to locate sources [5–11] and superspreaders [12,13], we still lack a comprehensive understanding of our ability to precisely identify the original source of spreading in a large complex network. The difficulty stems from the lack of a general locatability condition to predict if the source at any possible locations is fully locatable in terms of a given set of observers.

We develop a general locatability framework based on the time reversible characteristic of diffusion-like processes. This allows us to perform a time-reversal backward spreading to accurately locate the source, and offer a locatability condition that guarantees that a source will be fully locatable at any position. The algorithm and locatability condition are applicable in both directed and undirected networks with inherently limited knowledge of nodes and a time delay along links. We validate the tools by using a variety of complex networks in combination with two typical diffusion-like dynamical processes, i.e., epidemic spreading [14–16] and consensus dynamics [17,18]. We have also applied our method to real networked systems by employing empirical data from the 2009 H1N1 pandemic in China, focusing on the Chinese airline and train networks as the epidemic spreading network. The four sources predicted by our tools are in good agreement with empirical findings. Our framework has further potential applications in locating, for example, a spammer who abuses email systems and pollution sources in river networks.

II. TIME-REVERSAL BACKWARD SPREADING

Our goal is to locate the source that initiates a diffusion-like process taking place on an already-known undirected or directed complex network using only the limited time information pertaining to the diffusion observed from a fraction of nodes. This limited information could be the time period during which a person is being invaded by a virus, or the appearance of an abnormal signal at a node. To better mimic a real-world scenario, we assume that we are unable to detect communications between the observable nodes and their neighbors. For example, hospital records tell us when a patient became ill, but do not tell us who passed the

*shinancao@ruc.edu.cn
†wenxuwang@bnu.edu.cn

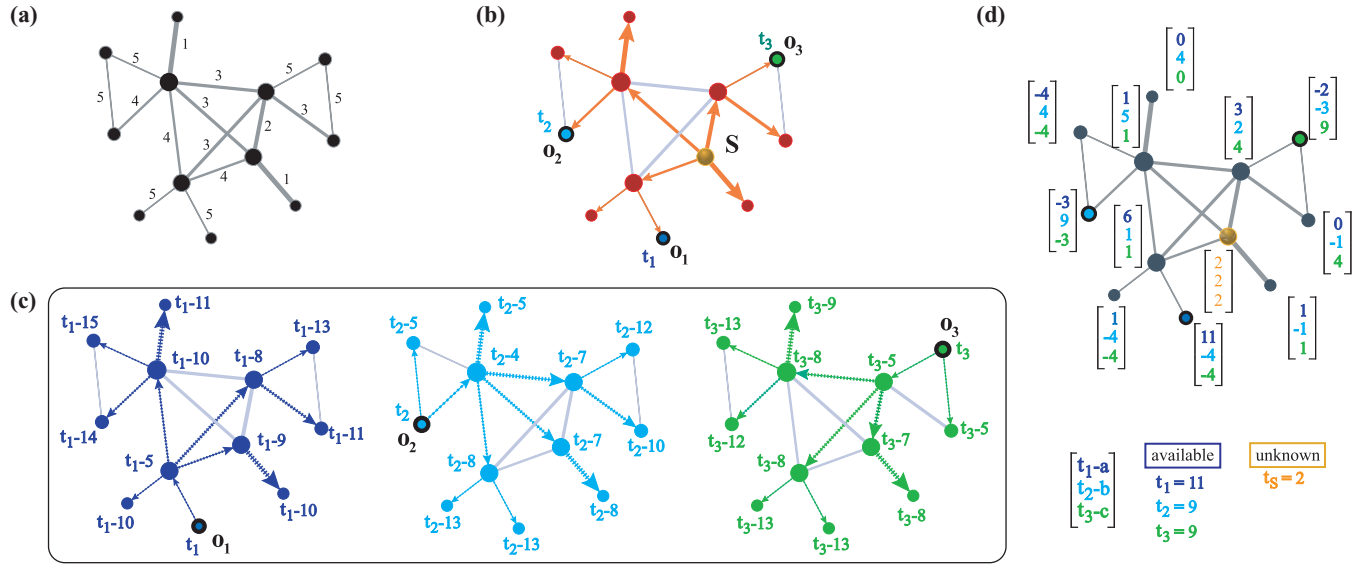

FIG. 1. Time-reversal backward spreading for locating the source. (a) A network topology with link weights (time delay). (b) The diffusion paths from the source S and the observers o_1, o_2, and o_3. The arrival time only at the three observers, namely, t_1, t_2, and t_3 can be accessed. (c) Implement TRBS along weighted shortest paths from o_1, o_2, and o_3, respectively, and the reversed arrival time at each node stems from each observer, respectively. (d) the vector $\mathbf{T}$ consisting of the reversed arrival time from each of the observers. The elements of $\mathbf{T}_s$ of the source are identical, which is the key to distinguishing the source from the other nodes. If the observers provide sufficient information of the source, the revered arrival time from observers are the original time t_s of the diffusion from the source, enabling the recovery of t_s. The source S is in yellow and the three observer nodes are in dark blue, light blue, and green with black boundary. The actual diffusion from S is marked by orange solid lines with arrows and the TRBS from the observers are, respectively, marked by colored dotted lines with arrows. The color of numbers in the vector in (d) corresponds to the observer with the same color.

disease to the patient. Even knowing all of the sick persons with whom the patient has had recent contact does not tell us.

The network and the spreading process are illustrated in Figs. 1(a) and 1(b), respectively. The weights along links are the time delay of passing a signal along links. For an undirected network, the delay along a link is the same for both directions. Figure 1(b) shows that a spreading process starts from source node s and propagates from the source to the whole network along the weighted shortest paths to all nodes (because the shortest paths are associated with the shortest propagation delay).

Our time-reversal backward spreading (TRBS) algorithm for locating sources is based solely on (i) the weighted network structure [Fig. 1(a)] and (ii) the arrival time of certain signals at nodes that we call observers. These accessible observers $o_1, o_2, \ldots, o_m$ receive a signal at time $t_{o_1}, t_{o_2}, \ldots, t_{o_m}$, as shown in Fig. 1(b). We assume the source s, the original time t_s at s, and the diffusion paths from s are unknown. Because of the stochastic effect in real-world networked systems, we may not know the exact propagation delay along a link between two nodes, but we assume that the time delay follows a certain distribution, e.g., the Gaussian or uniform distributions. Insofar as the mean value and variance are finite, which are commonly observed in real scenario, our algorithm is feasible if we use the mean delay. If the distributions of time delay on each link are nonidentical, we can use the mean value of each link to specify the time delay of each link. The TRBS algorithm based on the weighted network and the signal arrival time at some observers is as follows:

(i) Perform the TRBS starting from an observer o_k to all nodes in the networks along the reversed direction of links (for a directed network, TRBS from node i to j is allowed if and only if there is a directed link with direction from j to i, namely the reversed direction of the link; for an undirected network, links are bidirectional with the same time delay on both directions and the reversed direction is the same as the original direction). This yields a reversed arrival time $t_{o_k} - \hat{t}(i,o_k)$ at an arbitrary node i, where $\hat{t}(i,o_k)$ is the shortest time delay from o_k to i [see Fig. 1(c)]. Thus, the set of observers leads to a vector $\mathbf{T}_i = [t_{o_1} - \hat{t}(i,o_1), t_{o_1} - \hat{t}(i,o_2), \cdots, t_{o_1} - \hat{t}(i,o_m)]^{\mathrm{T}}$ for node i [see Fig. 1(d)]. Note that the reversed arrival time is a virtual time for source localization.

(ii) Calculate the variance of the elements in $\mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_N$. The node with the minimum variance is the source [see Fig. 1(d)]. Using our algorithm we can locate the source with computational amount $O(mN \log N)$, and $O(N^2 \log N)$ in the worse case, where m is the number of observers, N is the number of nodes, and $m < N$.

For an idealized scenario in which we know the exact time delay (weight) along each link, the source will have zero variance [see Fig. 1(d)]. Since the diffusion process is reversible, the time-reversal delay from o_k to s is equal to the actual delay from s to o_k, i.e., $t_{o_k} - t_s = \hat{t}(s,o_k)$, which leads to $t_{o_1} - \hat{t}(s,o_1) = t_{o_2} - \hat{t}(s,o_2) = \cdots = t_{o_m} - \hat{t}(s,o_m) = t_s$ with zero variance. In contrast, for a node other than s the paths of TRBS from the observers will not be the same as that of the actual paths of spreading from the source, and node variance will be nonzero.

III. LOCATABILITY CONDITION

We offer a locatability condition to determine if a source at any possible location can be fully localized from the arrival time t_{o_k} $(k = 1, \cdots, m)$ at arbitrary m given observers. Based on the vector $\mathbf{T}_i$ $(i = 1, \cdots, N)$ calculated from m observers, we define the difference between the vector of any two nodes i and j, $\Delta\mathbf{T}_{ij} \equiv \mathbf{T_i} - \mathbf{T_j}$. The locatability condition can then be given: if and only if the elements of $\Delta\mathbf{T}_{ij}$ for any two nodes are not all the same, the source at any location can be exactly identified.

The general locatability condition is equivalent to the statement that if there exist any two nodes, say, i and j, such that the elements of their $\Delta\mathbf{T}_{ij}$ are the same, the source cannot be distinguished between i and j. In the following, we justify this equivalent locatability condition. We first describe the equivalent condition mathematically. Let's denote the shortest time delay from node i to observer o_k by $\hat{t}(i,o_k)$, which is defined as

$$\hat{t}(i,o_k) = \sum_{v \in P(i,o_k)} \theta_v, \tag{1}$$

where θ_v is the time delay along link v and $P(i,o_k)$ denotes the set of shortest weighted path between i and o_k. Since the diffusion process is reversible along reversed links, according to the definition of $\mathbf{T}_i$, we have

$$\Delta\mathbf{T}_{ij} = \mathbf{T}_i - \mathbf{T}_j = \begin{pmatrix} \hat{t}(j,o_1) - \hat{t}(i,o_1) \\ \hat{t}(j,o_2) - \hat{t}(i,o_2) \\ \vdots \\ \hat{t}(j,o_m) - \hat{t}(i,o_m) \end{pmatrix}. \tag{2}$$

If the locatability condition is violated, namely,

$$\begin{aligned} \hat{t}(j,o_1) - \hat{t}(i,o_1) &= \hat{t}(j,o_2) - \hat{t}(i,o_2) = \ldots \\ &= \hat{t}(j,o_m) - \hat{t}(i,o_m), \end{aligned} \tag{3}$$

we cannot identify the source s when $s \in (i,j)$, which is the equivalent locatability condition and can be proved as follows. Assume that i is the actual source with original time t_i^s and node i and j satisfies Eq. (3). The source i gives rise to the arrival time $t_{o_1}, t_{o_2}, \ldots, t_{o_m}$ at observers $o_1, o_2, \ldots, o_m$. Suppose that j is the source and the original time at j is t_j^s, which leads to the arrival time $t'_{o_1}, t'_{o_2}, \ldots, t'_{o_m}$ at the same set of m observers (for the source, origin time is the same as arrival time). Taking the time reversible characteristics of TRBS along reversed links, we can simply have $t_{o_m} = \hat{t}(i,o_m)$ and $t'_{o_m} = \hat{t}(j,o_m)$. According to Eq. (3), we can derive that $t_{o_1} - t'_{o_1} = t_{o_2} - t'_{o_2} = \cdots = t_{o_m} - t'_{o_m} = t_i^s - t_j^s + c$, where c is a constant. Note that if the original time at j is $t_j^s = t_i^s + c$, we have $t_{o_1} - t'_{o_1} = t_{o_2} - t'_{o_2} = \cdots = t_{o_m} - t'_{o_m} = t_i^s - t_j^s + c = 0$, which indicates that source i and source j generate exactly the same arrival time as the actual observed arrival time at all the observers. Thus, the source cannot be distinguished between i and j in principle. In other words, because the actual original time t_s is unknown, if Eq. (3) is satisfied, there exists two possible original time t_i^s and t_j^s with $t_j^s = t_i^s + c$, such that the spreading process starts from node i, and j will generate the same arrival time as the actual arrival time at observers, rendering the source between i and j indistinguishable. Hence, our locatability condition offers a sufficient and necessary criterion for exclusively locating the source. If the locatability condition is satisfied, namely, Eq. (3) is violated, at least one observer is able to provide effective information that is sufficient to distinguish i and j by using, for example, our efficient algorithm. Therefore, the source in a network is said locatable if and only if for any two nodes i and j, the element values in $\Delta\mathbf{T}_{ij}$ are not all the same.

Figure 2 gives an intuitive example to explain the locatability condition. Since the original time t_s at the source is unknown, if we choose a certain original time, e.g., $t_s = 1$ at node i or $t_s = 2$ at node j, both nodes can produce the exact same arrival time at the three observers ($t_1 = 4$, $t_2 = 3$ and $t_3 = 3$), indicating that the source cannot be distinguished between i and j. Thus, the source in the network with respect to the given set of observers is not locatable. This scenario is exactly reflected by $\Delta\mathbf{T}_{ij}$ in which all elements are the same. The locatability condition in principle inhibits the indistinguishable scenario and exclusively locating the source at any location is assured. If the locatability condition is satisfied, namely, there is a single node in which the elements in its vector $\mathbf{T}_s$ are identical, this identical value is the original

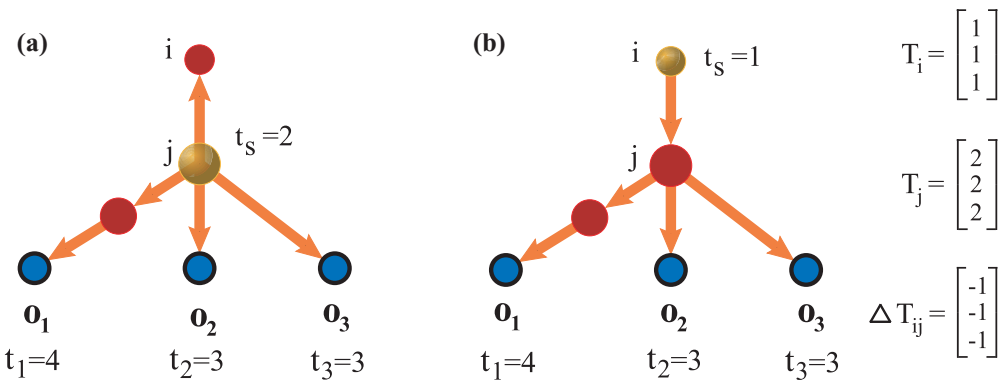

FIG. 2. The uncertainty of source. (a) A diffusion process from the source j at $t_s = 2$ with three observers o_1, o_2, and o_3. (b) A diffusion from the source i at $t_s = 1$ with the same observers as in (a). The source in (a) and (b) produces the same arrival time at the three observers, i.e., t_1, t_2, and t_3. (c) The vector $\mathbf{T}_i$ and $\mathbf{T}_j$ and the difference $\Delta\mathbf{T_{ij}}$ between them. Without loss of generality, we assume the time delay along each link is 1. The original time t_s of the diffusion from a source is known for the locatability problem. The color of nodes and links represents the same meaning as that described in the caption of Fig. 1.

time of the diffusion from the source. This is because of the intrinsic time-reversal characteristic of the TRBS process. When implementing the TRBS, the reversed arrival time at the source is nothing but the original time t_s that is the identical value in the vector $\mathbf{T}_s$ of the source, as shown in Fig. 1(d). Therefore, if the source in a complex network is fully locatable, the original time of diffusion can be inferred as well.

An immediate consequence of the locatability condition is that a node with a single neighbor must be observed to guarantee being fully locatable. This can be easily proved by noting that the node and any one of its neighbors cannot be distinguished for any observers, except the node itself according to Eq. (3). This consequence indicates that for a star graph, all nodes except the star should be observed, and in a tree, we usually need to observe a large fraction of nodes to enable full localization. For a fully connected network with N nodes, we must observe $N-1$ nodes to assure they are fully locatable. For an undirected chain, both ends should be observed for locating a source.

Note that the locatability condition is rigorous for idealized networks in which we know the exact time delay along each link. In practice, if the time delay of a link follows some distribution resulting from the stochastic effect, the locatability condition is violated somewhat. This is analogous to the structural observability [19] of those scenarios in which we lack a complete knowledge of link weights. Despite this lack, it is possible for us to use the locatability condition to identify a source from a pair of nodes. If the element values of $\Delta\mathbf{T}_{ij}$ are sufficiently close, it is likely that nodes i and j will be indistinguishable. If the element values differ greatly, however, it is easier for us to identify which one is more likely to be the source between them.

IV. SOURCE LOCALIZATION PERFORMANCE

To validate our locatability framework we explore two prototypical dynamical processes, diffusion and consensus. Diffusion processes commonly occur in many natural and social network systems, such as epidemic spreading in a population, virus propagation on the internet [20,21], rumor propagation in social networks [22], and risk contagion in financial networks [23]. Some dynamical processes are not subject to diffusion but exhibit diffusion-like behavior, e.g., cascading failures in power grids [24–27] and the spreading of gridlock in urban automobile traffic patterns [28–30]. To be as general as possible, we consider the simplest diffusion model, the one associated with diffusion delay. To simulate a diffusion process, we must first construct a complex network with a node degree distribution that allows the diffusion of a signal, e.g., a virus, a rumor, or a risky behavior in social network. Each link is assigned a time delay (weight) of forwarding the signal and the weights of links can be either the same or follow a distribution. The simulation is carried out as follows. First, a randomly selected source passes the signal to its neighbors. The signal takes some time to reach its neighbor nodes, depending on the link delays. Each node that has received the signal forwards it to its neighbors and this process continues until all the nodes in the network have received the signal. What we can measure and record is the arrival time of the signal at the observer nodes.

Consensus dynamics on complex networks have been investigated since the development of complex network science a decade ago [31–37]. Although most real systems display nonlinear behavior, agreement and synchronization phenomena are in many aspects similar to the consensus of linear systems. We thus use simple canonical linear, time-invariant dynamics with a communication delay [18],

$$\dot{x}_i = \sum_{j=1}^{N} a_{ij}[x_j(t-\tau_{ij}) - x_i(t)], \tag{4}$$

where $x_i(t)$ $(i = 1,\ldots,N)$ is the state of node i at time t, and τ_{ij} is the time delay along the link between node i and node j. We explore the diffusion of a perturbation starting from a single source node in the consensus state. Note that, unlike the standard diffusion process via contact or transportation, the diffusion-like process of perturbation is caused by the node coupling. Specifically, in the absence of external perturbations, all nodes uniformly stay in the consensus state. Thus, the transmission of a signal to other nodes can be discerned when deviation from the consensus state occurs. We record the time at which the state of observable nodes deviates from the consensus state and, using our locatability framework, to locate the source node with original perturbation.

We numerically validate our locatability condition by comparing with the success rate of locating sources when the exact weights of links are known. Figures 3(a) and 3(b) show the success rate of locating sources in small-world and scale-free networks by using our TRBS algorithm. It shows exact agreement with the prediction of the locatability condition for both homogeneous and inhomogeneous networks with a different average node degree $\langle k\rangle$ and fraction of observers n_o. The success rate achieves the upper bound predicted by the locatability condition, indicating that our TRBS algorithm is optimal for locating the source of spreading. Figures 3(c) and 3(d) show the minimum fraction $n_{\mathrm{o}}^{\min}$ of randomly chosen observers to reach 90% success rate affected by $\langle k\rangle$ in random and small-world networks. Note that $n_{\mathrm{o}}^{\min}$ exhibits a w-shaped function of $\langle k\rangle$ with two optimal values of $\langle k\rangle$. This counterintuitive finding can be understood in terms of the change of the maximum betweenness centrality (MBC) and the variance of the shortest path length (SPL). Their joint effects on n_{o} can be heuristically explained based on the locatability condition. On the one hand, let's consider a scenario that node i must be passed in order to reach node j along the shortest path from the observers. In this case, the source between i and j will be indistinguishable (see Fig. 2). If this occurs, the number of the observers is approximately equal to the betweenness centrality of i. Hence highest the probability of encountering this scenario for any two nodes is reflected in the MBC in the network. The larger MBC means that there is a higher probability that the locatability condition will be violated, and this accounts for the requirement of more observers, namely, the higher value of n_{o}. On the other hand, n_{o} is affected by the variance of the shortest path length in the network. If the shortest paths from all the observers to node i and j are the same, based on the locatability condition, the source between i and j will be indistinguishable in the sense that the reversed arrival time at both nodes are exactly the same. An extreme case is the fully connected network with zero variance of SPL

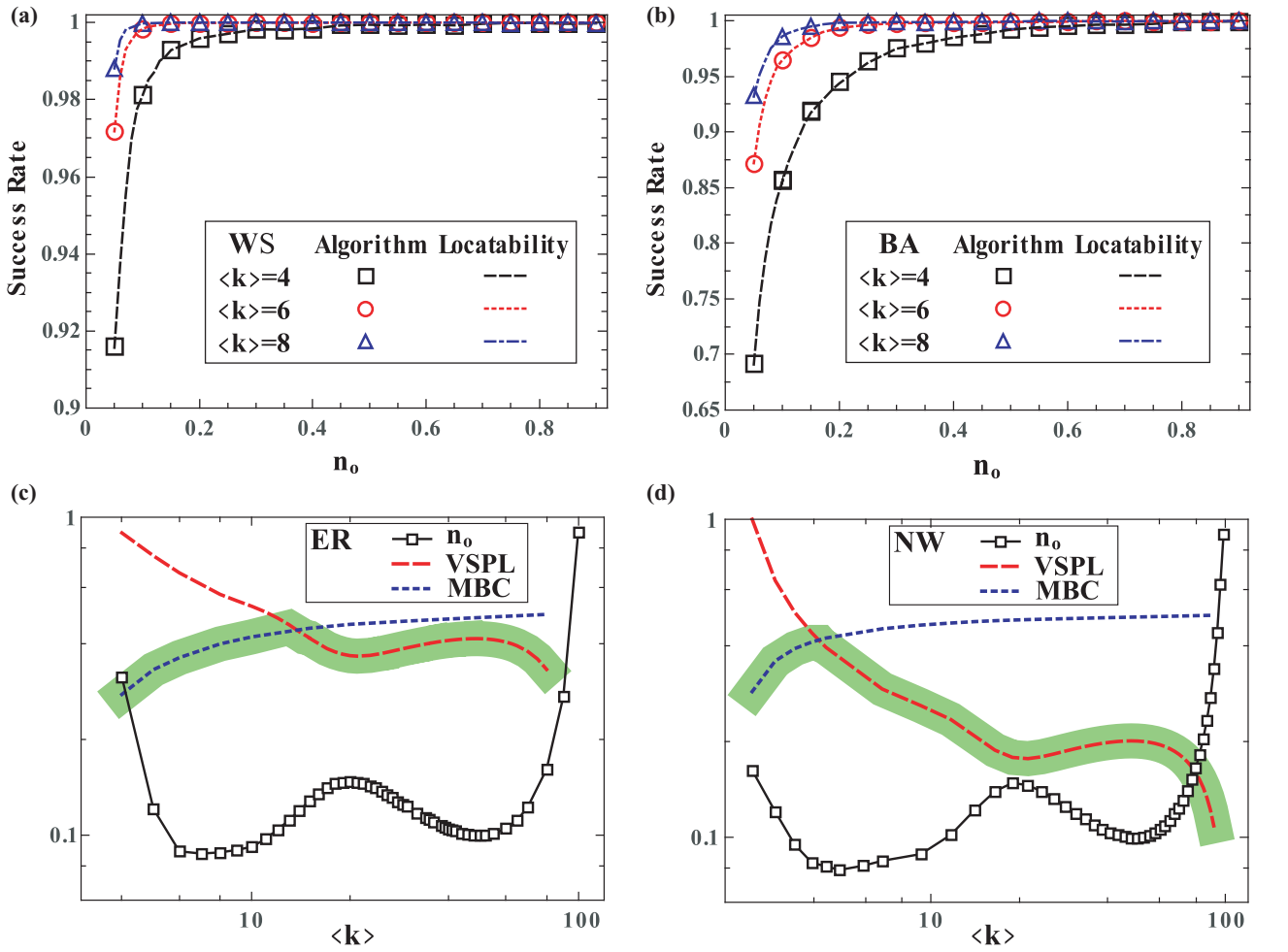

FIG. 3. Locatability of source in model networks. (a, b) Success rate obtained using the efficient algorithm and predicted by the locatability condition in Watts-Strogatz (WS) small-world network (a) and Barabási-Albert (BA) network (b) for different average node degree $\langle k \rangle$. (c, d) The minimum number n_o of observers to reach 90% success rate, the effect of the maximum betweenness centrality (MBC) and the variance of shortest path length (VSPL) as a function of $\langle k \rangle$, respectively, in Erdös-Rényi (ER) random network (c) and Newman-Watts (NW) small-world network. The green belt represents the joint effect of MBC and VSPL on the locatability. The numerical results are obtained by averaging over 400 independent realizations and the network size is 100.

in which $N-1$ observers are needed. Thus, a larger variance of SPL results in lower values of n_o. The joint effect of BC and SP on n_o gives rise to the "w shape" with two optimal average degrees, as shown in the green region in Figs. 3(c) and 3(d).

Table I displays $n_o^{\min}$ for achieving a 90% success rate of locating the source in homogeneous and heterogeneous networks associated with a Gaussian distribution and a uniform distribution of time delay along links, respectively. We assume that only the mean time delay along links rather than the exact time delay along each links is known. We assign the mean time delay to each link, such that the network becomes a weighted network with identical link weights. The results demonstrate that our algorithm is successful based on the mean time delay without exact time delays along links for both spreading and consensus dynamics. The small differences between $n_o^{\min}$ of spreading process and consensus dynamics are resulting from the approximation during the numerical integral of Eq. (4). Figure 4 shows the relations between $n_o^{\min}$ and network size N. As we can see, the fraction of required observers decreases as the network size increases for all the model networks, implying the effectiveness and applicability of our method. We also compares the performance with the Jordan center method [10], which is an topology based method, shown

TABLE I. Minimum fraction of observers. The minimum fraction $n_o^{\min}$ of randomly selected observers that assures 90% success rate of locating the source of spreading process and the propagation of perturbation in consensus dynamics on ER, WS, and BA networks. The time delays of links are assumed to follow Gaussian distributions with mean value 1.0 and variance 0.25 and uniform distributions in the range (0.5, 1.5), respectively. We exclusively use the mean delay of all links to identify sources. The network size N is 100 and the average node degree $\langle k \rangle = 8$. The results are obtained by averaging over 500 independent realizations.

	ER	WS (Gaussian/uniform)	BA
Spreading	0.18/0.23	0.23/0.36	0.29/0.41
Consensus	0.17/0.21	0.21/0.31	0.28/0.36

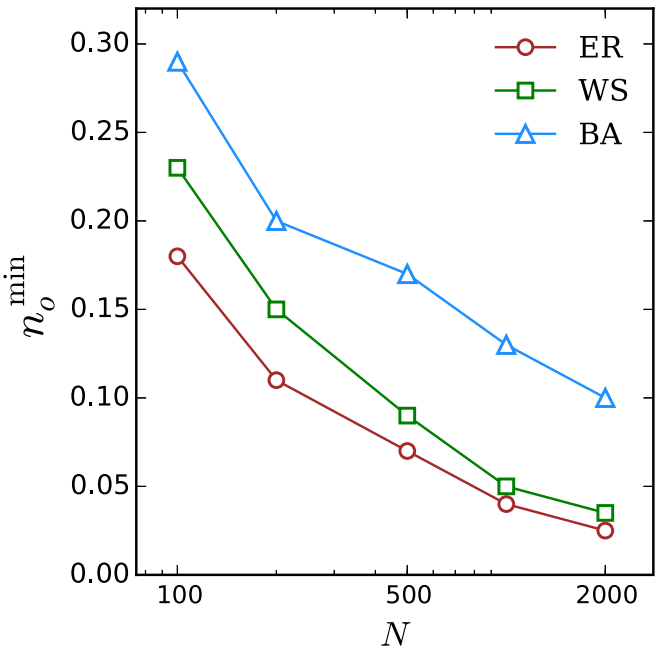

FIG. 4. Minimum fraction of observers for different network size. The minimum fraction $n_o^{\min}$ of randomly selected observers that assures 90% success rate of locating the source of spreading process on ER, WS, and BA networks. The time delays of links are assumed to follow Gaussian distributions with mean value 1.0 and variance 0.25. The average node degree $\langle k \rangle = 8$. The results are obtained by averaging over 500 independent realizations.

in Table II. The average rankings of the real source node in our algorithm approaches 1, which is much smaller than the rankings in Jordan center method. The robustness of our method under conditions of incomplete information and noisy data, and its need for only a small fraction of observers allows it to be generally applicable in real-world networked systems in which conditions of measurement noise and incomplete node information are inevitable.

V. LOCATING THE SOURCE OF H1N1 SPREADING IN CHINA

We apply our locatability framework to the H1N1 pandemic in China in 2009. We use the empirical data to quantify the arrival time of the virus at each major city to discern the source with the earliest arrival time. Note that we assume that only the arrival time of a fraction of major cities are accessible and we aim to locate the source from the arrival time. We use both airline and train networks among provinces to capture the spreading network, in which the total number of vertex is 31. The airports and train stations are usually located at the provincial capital cites, and the bidirectional links between two nodes are weighted and related with the customer flux estimated by the number of flights and trains per day. The time delay τ along each link is estimated from the flux of passengers in unit time by the following formula:

$$\tau_{ij} = \frac{1}{1-(1-\varphi)(1-\xi)^{w_{ij}}}, \tag{5}$$

where i and j represent two major cities, φ characterizes the time scale of the spreading process, ξ is the probability of a single infected passage taking an airplane or a train, w_{ij} is the number of equivalent airplanes per day between i and j. w_{ij} is set according to China airline and train data base, where a train is equivalent to 5 airplanes. φ is set to be $1/4$, due to the fact that the H1N1 pandemic in China lasted for about 4 months with the time unit 1 month. ξ is fixed to be $1/2000$ owing to the fact that on average there are about 300 available seats per airplane and about 1600 available seats per train with the sum is about 2000. We have checked that our results of locating the source is insensitive to the value of ξ. In the range of $1/1800 < \xi < 1/3000$, our algorithm offers approximately the same locating probability of the source. The dominator of Eq. (5) captures the infection probability between i and j, so that the reciprocal of the infection probability corresponds to the time delay.

Figures 5(a)–5(c) show the empirical record of the H1N1 pandemic in China in 2009. Specifically, Fig. 5(a) shows that the disease arises almost simultaneously from Beijing, Shanghai, Fujian, and Guangdong, i.e., these four provinces are the sources. Figure 5(b) shows the outbreak of the disease across China. Figure 5(c) shows the application of medical treatment after the epidemic has spread across the country causes the number of cases to decrease and, some months later, disappear. Figure 5(d) shows both airline and train networks in China with different passenger fluxes along the links. We randomly pick a fraction of nodes to be observers and record the outbreak time in each of them to be the arrival time, and use the combined network of flight and train to locate the disease sources (each province is a node with location

TABLE II. Performance comparison of Jordan center method and Time-reversal backward spreading method. All the nodes are ranked based on Jordan centrality in descending order and reversal time variance in ascending order, respectively. The ranking of the source of spreading process on ER, WS, and BA networks are averaged over 100 independent realizations. The time delays of links are assumed to follow Gaussian distributions with mean value 1.0 and variance 0.25 and uniform distributions in the range (0.5, 1.5), respectively. The fraction of observers is 0.05. The network size N is 1000 and the average node degree $\langle k \rangle = 8$. The mean ranking of source node and its standard deviation are presented.

		ER	WS (Mean ± std)	BA
Gaussian	TRBS	1.01 ± 0.10	1.36 ± 0.88	2.92 ± 8.26
	Jordan center	501.06 ± 285.20	500.95 ± 304.15	446.35 ± 278.48
Uniform	TRBS	1.08 ± 0.36	1.59 ± 1.02	6.11 ± 14.48
	Jordan center	491.75 ± 309.80	478.51 ± 290.18	520.63 ± 317.78

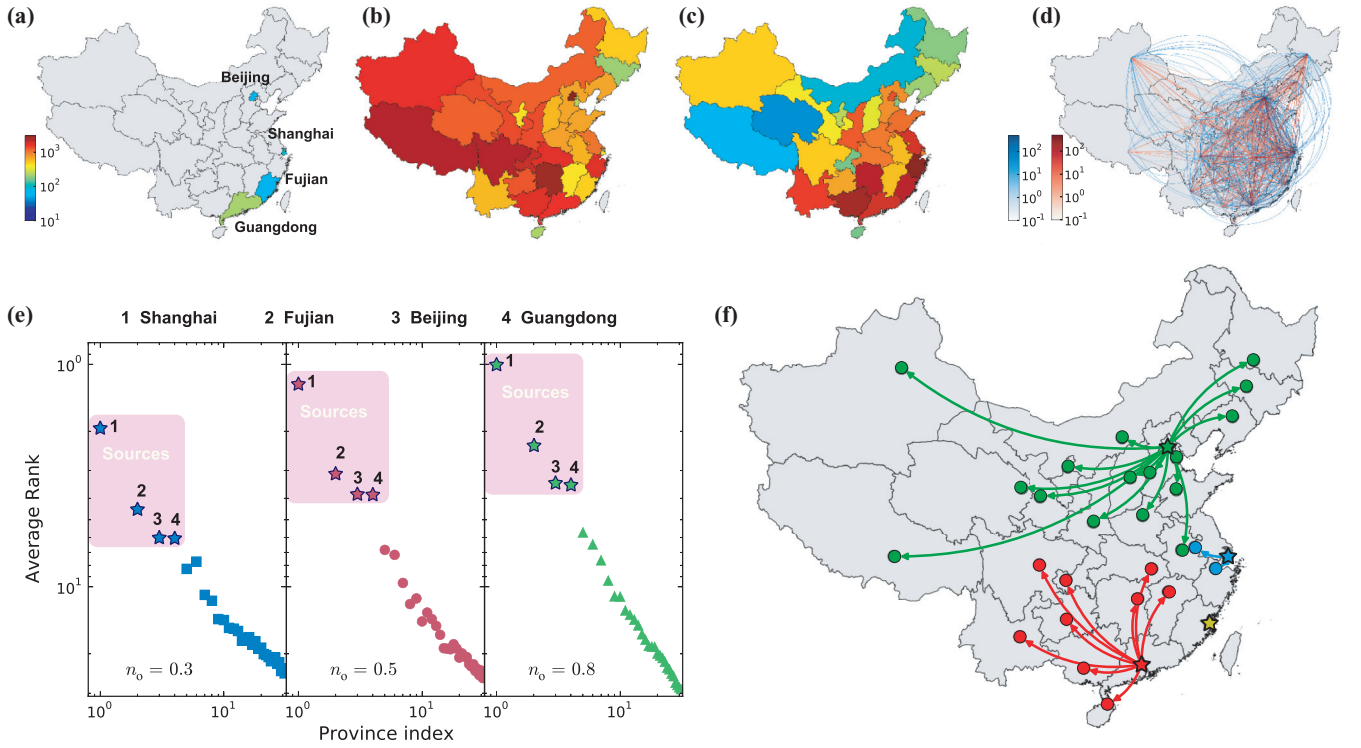

FIG. 5. Locate the sources of H1N1 pandemic in China. (a) The earliest outbreak of H1N1 in June 2009 in four provinces—Beijing, Shanghai, Fujian and Guangdong—which are the sources of the epidemic spreading in China. The epidemic outbreaks occur at the four locations nearly simultaneously. (b) The outbreak of H1N1 all over China in Oct. 2009. (c) The number of patients in China in December 2009. The color bar in (a), (b), and (c) denote the number of patents. (d) China airline and train networks with weighted links. The color bars capture the passenger flux of airlines and trains per day, respectively. The mixture of the airline and train networks is used as the propagation network of the H1N1 virus. (e) The average ranks of different provinces corresponding to the probabilities of being the sources of the epidemic spreading calculated by our algorithm. The four actual sources are of the highest four ranks with respect to different fraction n_o of observers and there is a clear gap between the sources and the other provinces. (f) The most probable paths of spreading from the sources uncovered by using the estimated time delays along links. The results in (e) are obtained by randomly choosing 100 independent configurations of observers with different fractions.

represented by the major city in the province). In particular, for a group of observers, we rank all the provinces according to their probability of being a source as revealed by the variance of the elements in their reversed arrival time vector $\mathbf{T}_i$. A node with smaller variance in $\mathbf{T}_i$ will be of higher probability to be a source. Figure 5(e) shows that the four nodes are found to have the highest average ranks by the independent realizations for different fractions of observers. Note that for $n_o > 0.3$, there is a clear gap between the average rank of the four provinces and that of the other provinces, indicating the presence of four sources. As n_o increases, the gap widens, which is a strong evidence that multiple sources exist. The four sources identified by our method are in exact agreement with the empirical record in Fig. 5(a), validating the practical applicability of our method. From the locations of the sources the most probable spreading paths of the disease can be ascertained based on the estimated time delay, as shown in Fig. 5(f). The spreading paths are obtained by preserving all paths with the shortest time delay from one of the sources in the set of all infection paths. The hidden radial spreading patterns from the sources are then uncovered using our locatability framework.

The fact that the H1N1 virus came from outside China accounts for the four sources that spurs the epidemic spreading in China. The four source provinces have international airports and we suspect that the virus may invade China via international flights from other countries. Despite the challenge of more than one sources, our algorithm still offers quite high accuracy of ascertaining all the sources, demonstrating the general applicability of our approach for addressing real problems.

VI. DISCUSSION AND CONCLUSION

In a huge network often only a subset of nodes are accessible. We thus need an efficient algorithm for locating the sources and ascertaining whether a given set of observers provide sufficient information for source localization. Our locatability framework uses the time-reversal backward spreading process on complex networks to provide tools to address these fundamental questions. Our algorithm uses the arrival time of a signal at the observers, the minimum information required, to locate the source. Our general locatability condition also enables us to determine whether the source in a network is fully locatable from a given set of observer nodes. We have systematically tested our theoretical tools using diffusion processes and consensus dynamics. Among the findings, an interesting result is the presence of two optimal locatabilities

as the link density increases from a very sparse network to a fully connected network. We have also applied our tools to H1N1 pandemic in China in 2009, finding that the four earliest-outbreak provinces identified by our method from a small fraction of observers are in good agreement with real data. Our theoretical tools have implications for many dynamical processes pertaining to disease control, identification of rare events in large networks, protection of the normal functioning of the Internet, and the behavior of economic systems.

Our work still has some limitations. For example, the time delay along each link is assumed to be known, while, in many real situations, we cannot get the time delays. How to accurately approximate the time delays with effective delays or equivalent delays, like the concept of effective distance in Ref. [6], when time delays are unavailable needs further investigation. In addition, our work raises a number of fundamental questions, answers to which could further improve our ability to locate the source of diffusion-like dynamics occurring on complex networks. First, how do we identify a minimum number of observers in an arbitrary network using the locatability condition? Second, how do we locate the sources using current methods if only part of the network structure is accessible? We may overcome this obstacle by using a network reconstruction approach based on the recently developed compressive sensing method [38–41]. Third, how do we rank the observers with respect to the amount of effective information they provide if the resources are limited and only a small fraction of nodes are accessible? Fourth, how to incorporate the information of time-delay variance and improve the performance if the whole time distribution is provided. The ideas in the Ref. [11] may give some hints for better using the information of time delay variance. Taken together, our tools, because of their lower information requirements and solid theoretical supports, could open new avenues for understanding and controlling complex network systems, an extremely important goal in contemporary science.

ACKNOWLEDGMENTS

The authors thank Xiaoyong Yan, Xiao Han, and Yin Fan for valuable discussions and help. This work was supported by NSFC under GrantS No. 61174150, No. 61573064, and No. 61074116, the Fundamental Research Funds for the Central Universities, and Beijing Nova Program.

[1] P. Bak, *How Nature Works* (Oxford University Press, Oxford, 1997).

[2] P. Bak, C. Tang, and K. Wiesenfeld, Self-Organized Criticality: An Explanation of the $1/f$ Noise, Phys. Rev. Lett. **59**, 381 (1987).

[3] C. Fraser *et al.*, Pandemic potential of a strain of influenza a (h1n1): Early findings, Science **324**, 1557 (2009).

[4] G. Neumann, T. Noda, and Y. Kawaoka, Emergence and pandemic potential of swine-origin h1n1 influenza virus, Nature **459**, 931 (2009).

[5] P. C. Pinto, P. Thiran, and M. Vetterli, Locating the Source of Diffusion in Large-Scale Networks, Phys. Rev. Lett. **109**, 068702 (2012).

[6] D. Brockmann and D. Helbing, The hidden geometry of complex, network-driven contagion phenomena, Science **342**, 1337 (2013).

[7] F. Altarelli, A. Braunstein, L. Dall'Asta, A. Lage-Castellanos, and R. Zecchina, Bayesina Inference of Epidemic on Networks via Belief Propagation, Phys. Rev. Lett. **112**, 118701 (2014).

[8] A. Y. Lokhov, M. Mézard, H. Ohta, and L. Zdeborová, Inferring the origin of an epidemy with dynamic message-passing algorithm, Phys. Rev. E **90**, 012801 (2014).

[9] K. Zhu and L. Ying, A robust information source estimator with sparse observations, Comput. Social Netw. **1**, 3 (2014).

[10] K. Zhu and L. Ying, Information source detection in the SIR model: A sample path based approach, *Information Theory and Applications Workshop (ITA)* (IEEE, San Diego, CA, 2013), pp. 1–9.

[11] Nino Antulov-Fantulin, Alen Lančić, Tomislav Šmuc, Hrvoje Štefančić, and Mile Šikić, Identification of Patient Zero in Static and Temporal Networks: Robustness and Limitations, Phys. Rev. Lett. **114**, 248701 (2015).

[12] M. Kitsak *et al.*, Identification of influential spreaders in complex networks, Nat. Phys. **6**, 888 (2010).

[13] S. Pei, L. Muchnik, J. S. Andrade Jr, Z. Zheng, and H. A. Makse, Searching for superspreaders of information in real-world social media, Sci. Rep. **4**, 5547 (2014).

[14] R. M. Anderson and R. M. May, *Infectious Diseases of Humans: Dynamics and Control* (Oxford Science Publications, Oxford, 1991).

[15] V. Colizza and A. Vespignani, Invasion Threshold in Heterogeneous Metapopulation Networks, Phys. Rev. Lett. **99**, 148701 (2007).

[16] D. Balcan *et al.*, Multiscale mobility networks and the spatial spreading of infectious diseases, Proc. Natl. Acad. Sci. U.S.A. **106**, 21484 (2009).

[17] R. O. Saber and R. M. Murray, Consensus protocols for networks of dynamic agents, In *Proceedings of the 2003 American Controls Conference* (IEEE, 2003), pp. 951–956.

[18] R. Olfati-Saber, J. A. Fax, and R. M. Murray, Consensus and cooperation in networked multi-agent systems, Proc IEEE **95**, 215 (2007).

[19] Y.-Y. Liu, J.-J. Slotine, and A.-L. Barabási, Observability of complex systems, Proc. Natl. Acad. Sci. U.S.A. **110**, 2460 (2013).

[20] A. L. Lloyd and R. M. May, How viruses spread among computers and people, Science **292**, 1316 (2001).

[21] J. Kleinberg, Computing: The wireless epidemic, Nature **449**, 287 (2007).

[22] D. H. Zanette, Dynamics of rumor propagation on small-world networks, Phys. Rev. E **65**, 041908 (2002).

[23] P. Gai and S. Kapadia, Contagion in financial networks, Proc. R. Soc. A **466**, 2401 (2010).

[24] A. E. Motter and Y.-C. Lai, Cascade-based attacks on complex networks, Phys. Rev. E **66**, 065102 (2002).

[25] S. V. Buldyrev, R. Parshani, G. Paul, H. E. Stanley, and S. Havlin, Catastrophic cascade of failures in interdependent networks, Nature **464**, 1025 (2010).

[26] C. M Schneider, A. A. Moreira, J. S. Andrade, S. Havlin and H. J. Herrmann, Mitigation of malicious attacks on networks, Proc. Natl. Acad. Sci. U.S.A. **108**, 3838 (2011).

[27] J. Gao, S.V. Buldyrev, H. E. Stanley and S. Havlin, Networks formed from interdependent networks, Nat. Phys. **8**, 40 (2011).

[28] J. Duch and A. Arenas, Scaling of Fluctuations in Traffic on Complex Networks, Phys. Rev. Lett. **96**, 218702 (2006).

[29] Z.-X. Wu, W.-X. Wang, and K.-H. Yeung, Traffic dynamics in scale-free networks with limited buffers and decongestion strategy, New J. Phys. **10**, 023025 (2008).

[30] W.-X. Wang, Z. X. Wu, R. Jiang, G. Chen, and Y.-C. Lai, Abrupt transition to complete congestion on complex networks and control, Chaos **19**, 033106 (2009).

[31] M. Newman, *Networks: An Introduction* (Oxford University Press, Oxford, 2009).

[32] N. A. Lynch, *Distributed Algorithms* (Morgan Kaufmann, Burlington, MA, 1996).

[33] R. Olfati-Saber, Flocking for multi-agent dynamic systems: Algorithms and theory, IEEE Trans. Autom. Control **51**, 401 (2006).

[34] T. Vicsek and A. Zafeiris, Collective motion, Phys. Rep. **517**, 71 (2012).

[35] M. Nagy, G. Vsrhelyi, B. Pettit, I. Roberts-Mariani, T. Vicsek, and D. Biro, Context-dependent hierarchies in pigeons, Proc. Natl. Acad. Sci. U.S.A. **110**, 13049 (2013).

[36] M. Egerstedt and X. Hu, Formation control with virtual leaders and reduced communications, IEEE Trans. Robot Autom. **17**, 947 (2001).

[37] A. V. Savkin, Coordinated collective motion of groups of autonomous mobile robots: Analysis of vicsek's model, IEEE Trans. Autom. Control **49**, 981 (2004).

[38] W.-X. Wang, R. Yang, Y.-C. Lai, V. Kovanis, and C. Grebogi, Predicting Catastrophes in Nonlinear Dynamical Systems by Compressive Sensing, Phys. Rev. Lett. **106**, 154101 (2011).

[39] W.-X. Wang, R. Yang, Y.-C. Lai, V. Kovanis, and M. A. F. Harrison, Time-series-based prediction of complex oscillator networks via compressive sensing, Europhys. Lett. **94**, 48006 (2011).

[40] W.-X. Wang, Y.-C. Lai, C. Grebogi, and J. Ye, Network reconstruction based on evolutionary-game data via compressive sensing, Phys. Rev. X **1**, 021021 (2011).

[41] R.-Q. Su, W.-X. Wang, and Y.-C. Lai, Detecting hidden nodes in complex networks from time series, Phys. Rev. E **85**, 065201 (2012).