


SCIENTIFIC REPORTS



OPEN

Locating multiple diffusion sources in time varying networks from sparse observations

Zhao-Long Hu¹, Zhesi Shen² , Shinan Cao³, Boris Podobnik^{4,5}, Huijie Yang⁶, Wen-Xu Wang^{2,6} & Ying-Cheng Lai^{7,8}

Data based source localization in complex networks has a broad range of applications. Despite recent progress, locating multiple diffusion sources in time varying networks remains to be an outstanding problem. Bridging structural observability and sparse signal reconstruction theories, we develop a general framework to locate diffusion sources in time varying networks based solely on sparse data from a small set of messenger nodes. A general finding is that large degree nodes produce more valuable information than small degree nodes, a result that contrasts that for static networks. Choosing large degree nodes as the messengers, we find that sparse observations from a few such nodes are often sufficient for any number of diffusion sources to be located for a variety of model and empirical networks. Counterintuitively, sources in more rapidly varying networks can be identified more readily with fewer required messenger nodes.

Diffusion and propagation processes taking place in complex networks are ubiquitous in natural and in technological systems^{1,2}. Examples of those processes include air or water pollution diffusion^{3,4}, disease or epidemic spreading in the human society^{5,6}, virus invasion in computer and mobile phone networks^{7,8}, behavior propagation in online social networks⁹. Once a negative diffusion or propagation emerges, it is imperative to locate its sources quickly and precisely to enable timely and appropriate control strategies to prevent and/or inhibit the spreading process. A number of methods have been proposed and tested recently to address the source localization problem of propagation processes in complex networks, which include those based on the maximum likelihood estimation¹⁰, dynamic message passing¹¹, belief propagation¹², hidden geometry of contagion¹³, and inverse spreading^{14,15}. A related problem of practical significance is to identify super spreaders for effective control of spreading^{16,17}. However, most existing approaches are specifically for static networks. In the real world time varying networks are ubiquitous, such as frequently changed social contacts via meetings, emails, phone and online softwares^{18–21}. Recently, a source detection framework was proposed on complex networks from one snapshot observation of the entire network and demonstrated for an empirical temporal network of sexual contacts²².

Those works focus primarily on source localization for propagation processes. However, source localization for diffusion processes is rarely studied. Here we concentrate on diffusion processes, as they constitute a good approximation for different types of dynamical processes (e.g., synchronization and other nonlinear processes amenable of linearization)². Very recently, considering multiple sources may exist (e.g., air or water pollution, rumors), a general framework that locating of multiple sources in static diffusion processes is presented²³. To develop effective frameworks to locate sources in time varying networks is an outstanding problem in network science and engineering. The essential difference between diffusion on a time varying network and on a static network is illustrated in Fig. 1. Specifically, in Fig. 1(a), due to the various time intervals in which different edges are activated, a spreading process starting at node *b* cannot reach node *a* in any time. In contrast, for a static network with the same structure as shown in Fig. 1(a), the spreading process can reach all nodes in the network. To

¹College of Mathematics, Physics and Information Engineering, Zhejiang Normal University, Jinhua, 321004, Zhejiang, China. ²School of Systems Science, Beijing Normal University, Beijing, 100875, China. ³School of Finance, University of International Business and Economics, Beijing, 100029, P. R. China. ⁴Center for Polymer Studies Boston University, Boston Massachusetts, 02215, USA. ⁵Faculty of Civil Engineering, University of Rijeka, 51000, Rijeka, Croatia. ⁶Business School, University of Shanghai for Science and Technology, Shanghai, 200093, China. ⁷School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, Arizona, 85287, USA. ⁸Department of Physics, Arizona State University, Tempe, Arizona, 85287, USA. Zhao-Long Hu and Zhesi Shen contributed equally to this work. Correspondence and requests for materials should be addressed to W.-X.W. (email: wenxuwang@bnu.edu.cn)

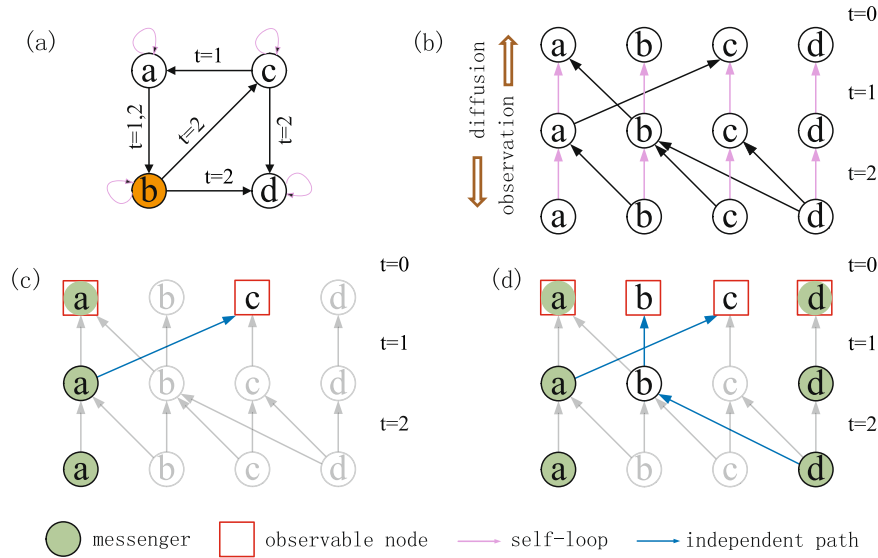


Figure 1. Illustration of proposed framework to locate multiple diffusion sources in time varying networks. (a) A simple directed time varying network where the numbers associated with the edges denote the activation time, where node *b* is the source and self loops are specified by the nonzero diagonal elements of matrix *A*. (b) Static mapping of the network in (a), where each layer corresponds to an activation time. The diffusion direction is from top to bottom (from $t=0$ to $t=2$), while observations occur in the opposite direction. (c) An independent path from observing messenger node *a*, where the observable range is $N_{OR}(\{a\}) = 2$. (d) Two independent paths from observing nodes *a* and *d*. The network is fully observable ($N_{OR}(\{a,d\}) = 4$) and sources are fully locatable.

our knowledge, there has been no solution to the problem of locating multiple diffusion sources associated with general dynamical processes on arbitrary time varying networks from local observations²⁴. The purpose of this paper is to provide an optimal solution. In particular, exploiting a combination of the structural observability and sparse signal reconstruction theories, we develop a general source localization framework that is applicable to arbitrarily time varying networks with any number of sources. We demonstrate that sparse data from a small set of messenger nodes are capable of identifying multiple diffusion sources accurately and efficiently, even in the absence of detailed information about the network structure such as link weights and the presence of measurement noise. The framework is established analytically and validated through extensive numerical tests of model and empirical networks.

Results

Framework of locating multiple sources on time-varying networks. A time-varying network with N nodes is generally defined by a node set $V = \{v_1, v_2, \dots, v_N\}$ with a set E of time varying edges, where $(v_i, v_j, w_{ij}, t) \in E$ denotes a directed edge pointing from nodes v_i to v_j with link weight w_{ij} at activation time t . In this paper, we consider the following class of discrete-time, diffusion processes on such time varying networks:

$$x_i(t + 1) = x_i(t) + \beta \sum_{j=1}^N [w_{ij}(t + 1)x_j(t) - w_{ji}(t + 1)x_i(t)], \tag{1}$$

where $x_i(t)$ is the state of node i at time t capturing the fraction of infected individuals, the concentration of water or air pollutant and etc., at place i . β is the constant diffusion coefficient, and $w_{ij}(t)$ is the link weight at time t , where self loops are a result of the diffusion process². For an undirected network, we have $w_{ij}(t) = w_{ji}(t)$. (Diffusion dynamics in continuous time can be treated similarly - see Sec. S1 in Supplemental Information (SI)). The nodes from which observations are made are the *messenger nodes*. When the outputs from the messenger nodes are taken into account, the system becomes

$$\begin{cases} \mathbf{x}(t + 1) = A(t + 1)\mathbf{x}(t), \\ \mathbf{y}(t) = C\mathbf{x}(t), \end{cases} \tag{2}$$

where the state vector $\mathbf{x}(t) \in \mathbb{R}^N$ comprises all nodes in the network at time t and $A(t + 1) = I + \beta L(t + 1)$. In $A(t + 1)$, $I \in \mathbb{R}^{N \times N}$ is the identity matrix, $L(t) = W(t) - D(t)$ is the network Laplacian matrix at time t , $W(t) \in \mathbb{R}^{N \times N}$ is the weighted adjacency matrix of elements $w_{ij}(t)$, and $D(t) \in \mathbb{R}^{N \times N}$ is a diagonal matrix of elements $d_i(t)$ denoting the total out-weight $\sum_{j \in \Gamma_i(t)} w_{ij}(t)$ of node i with $\Gamma_i(t)$ being the neighboring set of i at time t . The vector $\mathbf{y}(t) = [y_1(t); y_2(t); \dots; y_q(t)]$ represents the q measurable outputs from q messengers at time t , and $C \in \mathbb{R}^{q \times N}$ is the *output matrix*, where $C_{ij} = 1$ if output $y_i(t)$ is measured from node j . The basic difference between

source nodes and passive nodes is that, initially ($t = t_0$), the states of the former and latter are nonzero and zero, respectively. Without loss of generality, we set $t_0 = 0$. Thus, if the initial states of all nodes can be recovered from the measurements of the messenger nodes at a later time ($t > 0$), all sources can be identified. A solution to this problem can be obtained by exploiting the observability condition in canonical control theory. Specifically, we consider instants of time $t = 0, 1, \dots, T$ and rewrite Eq. (2) as

$$\mathbf{Y} = \begin{pmatrix} \mathbf{y}(0) \\ \mathbf{y}(1) \\ \vdots \\ \mathbf{y}(T) \end{pmatrix} = \begin{pmatrix} C \\ CA(1) \\ \vdots \\ CA(T)A(T-1)\cdots A(1) \end{pmatrix} \mathbf{x}(0) \equiv O \cdot \mathbf{x}(0). \tag{3}$$

where $\mathbf{Y} \in \mathbb{R}^{q(T+1)}$, $\mathbf{x}(0) \in \mathbb{R}^N$ is the initial state vector, q is the number of messenger nodes, and $O \in \mathbb{R}^{q(T+1) \times N}$ is the observability matrix. To be able to accurately locate the diffusion sources, a unique solution of Eq. (3) is needed, given the output vector \mathbf{Y} from the set of messenger nodes. The classic observability theory stipulates that, if and only if matrix O has full rank, i.e., $\text{rank}(O) = N$, $\mathbf{x}(0)$ can be fully and uniquely determined.

If we observe only a single node v , matrix O may not have full rank. As a result, only the initial states of a subset of nodes in $\mathbf{x}(0)$ can be reconstructed. The number of nodes whose initial states can be reconstructed is $\text{rank}(O)$, which defines the observable centrality $N_{\text{OR}}(\{v\})$ of v , i.e., $N_{\text{OR}}(\{v\}) = \text{rank}(O)$. Analogously, for a given set Q of nodes, we have an associated matrix C and can obtain $\text{rank}(O)$, which defines the observable range $N_{\text{OR}}(Q)$ of Q , i.e., $N_{\text{OR}}(Q) = \text{rank}(O)$. Note that $N_{\text{OR}}(\{v\}) \leq N$ and $N_{\text{OR}}(Q) \leq N$. Thus we can define a normalized observable centrality $n_{\text{OR}}(\{v\}) \equiv N_{\text{OR}}(\{v\})/N$ and a normalized observable range $n_{\text{OR}}(Q) \equiv N_{\text{OR}}(Q)/N$.

Since information about the link weights may not be available, a direct calculation of $\text{rank}(O)$ is not feasible. A resolution is to analyze the structural observability^{25–28}, which is a highly nontrivial task for time varying networks. Our idea is to exploit the independent paths in static mappings of the underlying network²⁹, as shown in Fig. 1(b). In particular, a mapping from a time varying network to a static network can be obtained by cloning all nodes into different layers that correspond to different time t . If an edge is active at t [as shown in Fig. 1(a)], the two nodes at both ends of the edge in the corresponding layers in Fig. 1(b) will be connected. Note that the direction of links in Fig. 1(b) is reversed with respect to the actual direction of diffusion in Fig. 1(a) - a consequence of the duality relation between structural observability and controllability²⁸.

Figure 1(c) indicates the quantity $N_{\text{OR}}(\{a\})$ when node a is chosen as a messenger node. There is a single independent path, i.e., $a \rightarrow c$, such that $N_{\text{OR}}(\{a\}) = 2$ (one independent path and a itself). If a and d are messengers [Fig. 1(d)], there are two independent paths and $N_{\text{OR}}(\{a, d\}) = 4$ (including the two messengers themselves). In this case, the network is fully observable. The key to source localization is thus to identify all independent paths from messenger nodes in the static mappings of the original time varying network. In this paper, to generate a time-varying network, we propose a uniform activation network model in which random activations are imposed on a static network. Specifically, let z be the number of times (activations) an edge is active in a time interval, which can be randomly selected from a uniform distribution $U(1, z_{\text{max}})$ with z_{max} denoting the maximum number of activations. After z is given for each edge, the active time associated with each activation is uniformly chosen from the distribution $U(1, T)$ under the constraint that a link cannot be activated twice (or more) at one active time.

Estimate of observable range. For a set Q of messenger nodes, $N_{\text{OR}}(Q)$ is exactly the number of independent paths plus the number of the messengers, which can be calculated by using the maximum flux algorithm. Here, we provide a theoretical estimate of the number of independent paths. As shown in Fig. 1, since every node has a self-loop, if there exists a link for a certain layer ($t > 0$), there must exist a path starting from the layer to the top layer ($t = 0$), as shown in Fig. 1(d). Moreover, there exists at most one independent path starting from one node in a given layer ($t > 0$). Thus, for a messenger node v , the maximum number of independent paths from v for all layers is the number of layers in which v has a link that points to other nodes. The number is nothing but the number l_v of distinct activations of v , where each activation (active time) corresponds to a layer with a link going out from v (see Sec. S2 in SI for more details). Thus, since the overlap among independent paths from v is negligible, we have $n_{\text{OR}}(\{v\}) \approx (l_v + 1)/N$, based on which the quantity $n_{\text{OR}}(Q)$ of node set Q can be estimated as

$$n_{\text{OR}}(Q) \approx \sum_{i \in Q} (l_i + 1)/N. \tag{4}$$

The fraction p of messenger nodes is thus $p = q/N$, where q is the number of messengers.

For the uniform activation network model, if the number of distinct activations, l_v , cannot be directly measured, we can use the activation times distribution $U(1, z_{\text{max}})$ and the active time distribution $U(1, T)$ to estimate the average number $\langle l \rangle$ of distinct activations. Specifically, for a node with k edges, we denote their activations by z^1, \dots, z^k . The probability of the number of distinct activations being l for one node with z^1, \dots, z^k is given by (see Sec. S2 in SI)

$$P(l|z^1, \dots, z^k) = c_1 \binom{T}{l} \sum_{j=\max(z^1, \dots, z^k)}^l (-1)^{l-j} \binom{l}{j} \prod_i^k \binom{j}{z^i}, \tag{5}$$

where c_1 is a normalization constant satisfying

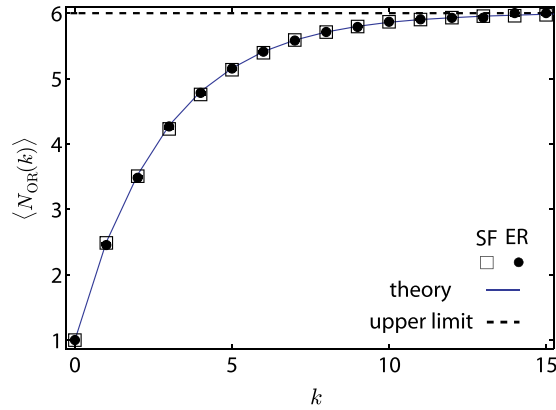


Figure 2. Observable centrality of a single messenger of degree k in ER random and SF networks, where the theoretical prediction is from Eq. (9) and numerical results are obtained from averaging over 10,000 independent realizations. The vertical bars indicate the standard error. Observable centrality increases with the degree and approaches its upper limit $\langle N_{OR} \rangle = T + 1$. Other parameters are $N = 1000$, $\langle k \rangle = 6$, $z_{\max} = 2$ and $T = 5$.

$$\sum_{l=\max(z^1, \dots, z^k)}^{\min(T, \sum z^i)} P(l|z^1, \dots, z^k) = 1. \tag{6}$$

Therefore, for one node associated with z^1, \dots, z^k , the average number of distinct activations is

$$\langle l \rangle_{\{z^1, \dots, z^k\}} = \sum_{l=\max(z^1, \dots, z^k)}^{\min(T, \sum z^i)} l P(l|z^1, \dots, z^k). \tag{7}$$

For a node of degree of k , the average number of distinct activations is

$$\langle l \rangle = c_2 \sum_{z^1=1}^{z_{\max}} \dots \sum_{z^k=1}^{z_{\max}} \langle l \rangle_{\{z^1, \dots, z^k\}}, \tag{8}$$

where $c_2 = (z_{\max})^{-k}$. Given $\langle l \rangle$ for each node, for the entire messenger set Q , the normalized observable range can be approximated as

$$n_{OR}(Q) \approx \sum_{i \in Q} (\langle l_i \rangle + 1) / N. \tag{9}$$

Messenger selection. Considering the cost of allocating messengers for monitoring the state of the whole network, finding a minimum set of messengers through independent paths represents the most efficient way to locate sources. Moreover, the set can be used to characterize the source locatability of the network. The difficulty is that this task is NP-complete³⁰. We employ an alternative approach by exploiting a greedy optimization algorithm to maximize the observable range n_{OR} through selection of the messenger set (see Sec. S3 in SI). In addition, sub-modularity^{31,32} is exploited to reduce the computational cost and provides guaranteed performance at least $(1 - 1/e) \approx 0.63$ compared to the global optima.

We test our framework using model and empirical networks. Figure 2 shows the observable centrality of nodes for Erdős-Rényi (ER)³³ random and scale-free (SF)³⁴ networks. Three features are found, which do not occur for static networks³⁵. First, nodes of larger degree k have a higher observable centrality N_{OR} , in sharp contrast to what happens in a static network where both driver and messenger nodes tend to avoid large degree nodes due to their small controllable and observable range. Second, N_{OR} gradually approaches the upper limit $T + 1$ as k increases. Third, N_{OR} is nearly independent of the network structure and depends mainly on T and z_{\max} . The theoretical prediction [Eq. (9)] and numerical results agree well with each other.

The results in Fig. 2 suggest that large-degree nodes be chosen as the messengers (denoted as the max-deg strategy). To validate this strategy, we compare it with the more elaborative strategy of greedy optimization. As shown in Fig. 3, n_{OR} resulting from the max-deg strategy is quite close to that from the greedy strategy, especially for relatively larger values of z_{\max} . The great advantage of the max-deg strategy is that it is based on *local information only* whereas the greedy strategy requires global information about the network. Another remarkable finding is that a very small fraction p of messenger nodes are sufficient to fully locate multiple sources ($n_{OR} = 1$) for both ER and SF networks. We also test our framework using three empirical time varying networks, as shown in Fig. 4. It should be noted that the number of distinct activations l of every node is available. We see that a quite small value of p can ensure a complete localization of diffusion sources in all the empirical networks. For both model and empirical networks, numerical calculations are in good agreement with theoretical predictions (see Sec. S3 in SI for more details).

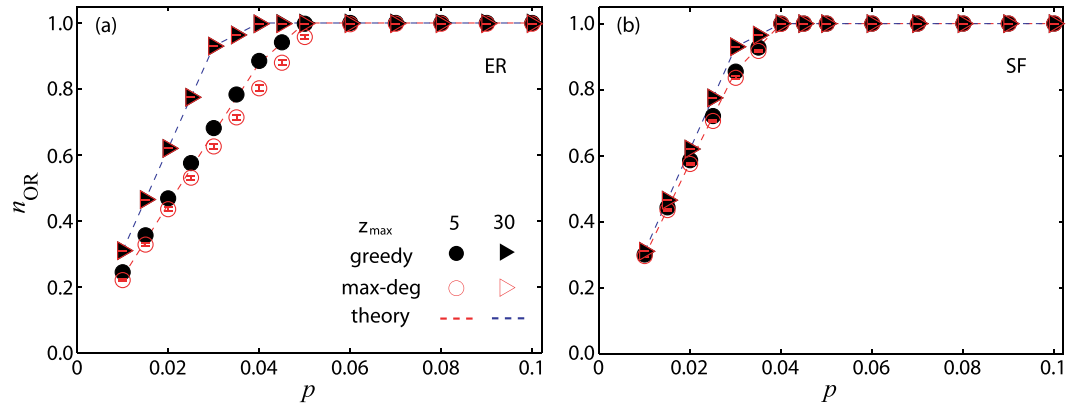


Figure 3. Normalized observable range n_{OR} as a function of the fraction p of messenger nodes for (a) ER and (b) SF networks, using the greedy algorithm and max-deg strategy for different values of z_{max} , for $T = 30$. The analytical results (dashed curves) from the max-deg strategy are from Eq. (9). Network parameters are $N = 100$ and $\langle k \rangle = 6$. All results are obtained by averaging over 50 independent realizations and the vertical bars indicate the standard error.

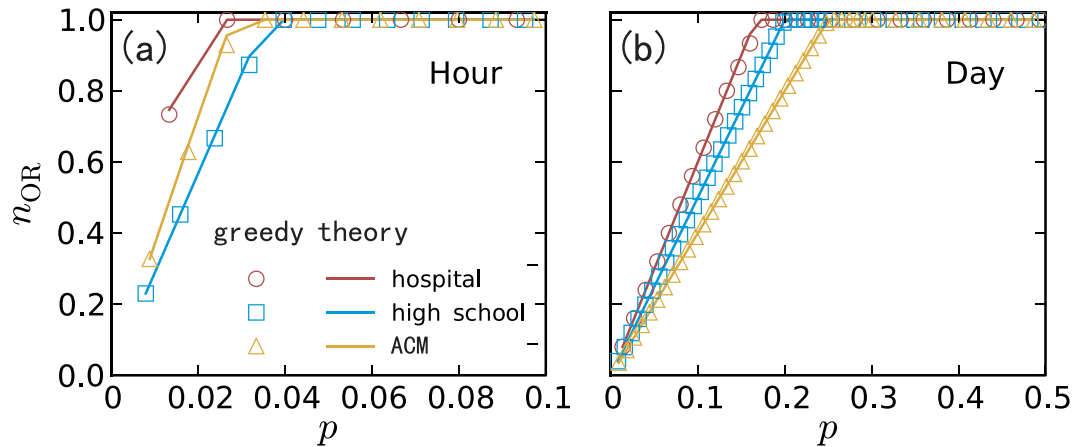


Figure 4. Normalized observable range n_{OR} as a function of the fraction p of messenger nodes for three empirical networks: Hospital, High School, and ACM. The time windows for (a) and (b) are one hour and one day, respectively. A greedy algorithm for finding the messenger nodes is used. The theoretical predictions (the solid curves) are from $n_{OR}(Q) \approx \sum_{i \in Q} (l_i + 1) / N$. Details of the empirical networks and the meaning of the time window can be found in Table S1 and Sec. S4 in SI.

A counterintuitive phenomenon is that, in both model and real networks, it is relatively easier to locate diffusion sources in more rapidly changing (more frequently updating) networks as the set of required messenger nodes is smaller (e.g., comparing $z_{max} = 5$ with $z_{max} = 30$ in Fig. 3 and hour with day in Fig. 4). A heuristic explanation is that more rapid changes in the network structure in fact limit the spreading patterns from sources, facilitating source localization from a relatively smaller number of messenger nodes.

Actual localization of multiple diffusion sources. We articulate an efficient and robust method to actually locate the sources based on the already identified messenger set. In a realistic situation, the number of sources is much smaller than the network size, so the vector $\mathbf{x}(0)$ in Eq. (3) has many zero elements. The sparsity of $\mathbf{x}(0)$ can be exploited to greatly reduce the required measurement from messengers by using the compressive sensing (CS) paradigm for sparse signal reconstruction^{36,37}. Specifically, Eq. (3) can be solved and accurate reconstruction of $\mathbf{x}(0)$ can be achieved through solutions of the following convex-optimization problem:

$$\min \|\mathbf{x}(0)\|_1 \text{ subject to } \mathbf{Y} = \mathbf{O} \cdot \mathbf{x}(0), \tag{10}$$

where $\|\mathbf{x}(0)\|_1 = \sum_{i=1}^N |\mathbf{x}_i(0)|$ is the L_1 norm of $\mathbf{x}(0)$, while $\mathbf{Y} \in \mathbb{R}^{qM}$, and $\mathbf{O} \in \mathbb{R}^{qM \times N}$. Here M is the number of continuous measurements made by messengers. Because of the linear independence of the rows in matrix \mathbf{O} and the sparsity of $\mathbf{x}(0)$, it is feasible to reconstruct $\mathbf{x}(0)$ as M is much smaller than $T + 1$. We define $n_M \equiv M / (T + 1)$ to compare with the data amount $T + 1$ required by conventional solution to $\mathbf{x}(0)$. To be more realistic, we include both measurement noise and uncertainties in the link weights in Eq. (2), which is reformulated as

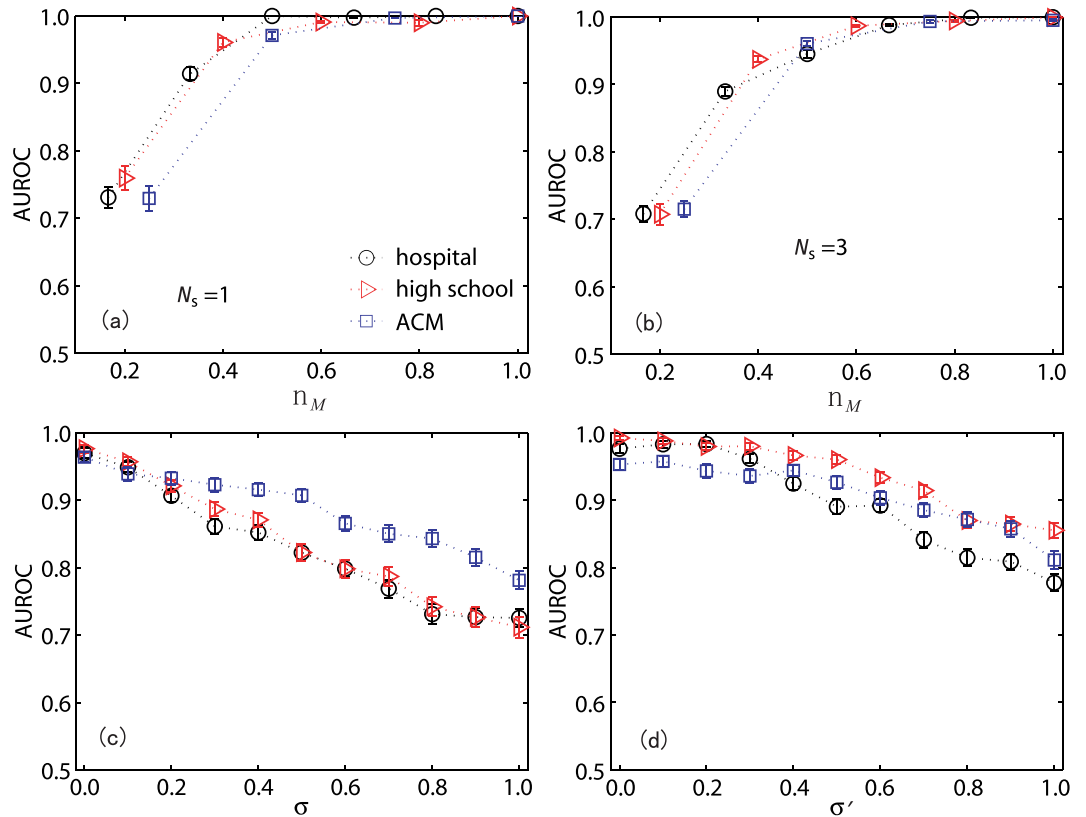


Figure 5. Performance of source localization for empirical networks. **(a,b)** AUROC as a function of n_M without noise for a single source ($N_s = 1$) and three sources ($N_s = 3$), respectively. **(c,d)** AUROC versus the measurement noise standard deviation σ and link weight standard deviation σ' , respectively. Parameters are $p = 0.15$ and $\beta = 0.05$. In **(c,d)**, the values of n_M are 0.5, 0.6 and 0.5 for hospital, high school and ACM, respectively. The time window is a day and there is a single source. All results are obtained by averaging over 500 independent realizations and the vertical bars indicate the standard error.

$$\begin{cases} \mathbf{x}(t) = \hat{A}(t)\mathbf{x}(t - 1), \\ \hat{\mathbf{y}}(t) = \mathbf{C}\mathbf{x}(t) \cdot (\mathbf{1} + \varepsilon), \end{cases} \quad (11)$$

where the measurement $\mathbf{y}(t)$ is contaminated by white truncated Gaussian noise of zero mean and variance σ^2 : $\varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{1})$, where $\mathbf{0} \in \mathbb{R}^q$ is zero vector and $\mathbf{1} \in \mathbb{R}^q$ is the one vector. We assume that the uncertainties in the link weights W are also truncated Gaussian: $\hat{w}_{ij}(t) = w_{ij}(t)(1 + \varepsilon')$, where $\varepsilon' \sim \mathcal{N}(0, \sigma'^2)$. The random noise is restricted to positive values to make sure that the values of measurements and link weights are nonnegative. Here we use multiplicative noise to ensure that, on average, the ratio of the measurements remains the same with or without noise during the dynamics. To quantify the performance of source localization, we use the standard AUROC (area under a receiver operating characteristic) metric³⁷, where $AUROC = 1$ indicates the existence of a threshold to fully distinguish between sources and passive nodes whereas $AUROC = 0.5$ indicates that the two types of nodes cannot be distinguished (Sec. S7 in SI).

We use empirical networks (as in Fig. 4) to test the performance of our CS based source localization method. As shown in Fig. 5(a) and (b), AUROC increases with n_M . When n_M is small, AUROC shows large deviation indicating that the location of sources largely affects the accuracy of source localization for given selected messengers; once n_M exceeds some value, say 0.5, AUROC is close to 1 and the standard deviation reduces a lot implying that all sources at any locations can be accurately located. We also compared the performance of source localization for different messenger selection strategies (See Sec. S4 and Fig. S5 in SI). Figure 5(c) and (d) show the localization accuracy versus measurement noise σ and weight uncertainty σ' . We see that relatively high accuracy can still be achieved even when the noise variance approaches unity. Nonetheless, in some simulations the AUROC is small (See Sec. S5 and Fig. S6 in SI for the distributions of AUROC) and we may improve these performances by increasing the number of messengers or the length of observation time. Further efforts are still needed to see how to balance the cost of adding more messengers or increasing observation time.

In real systems, we cannot know the time-varying network structure in advance, which prevents us from selecting the optimal messengers. However, if the network structure evolves with periodicity or follows some patterns, e.g., the activation dynamic of each edge remains stable for a long period, we can construct a rough network based on the past interactions and select messengers using its structural properties, e.g., nodal degree

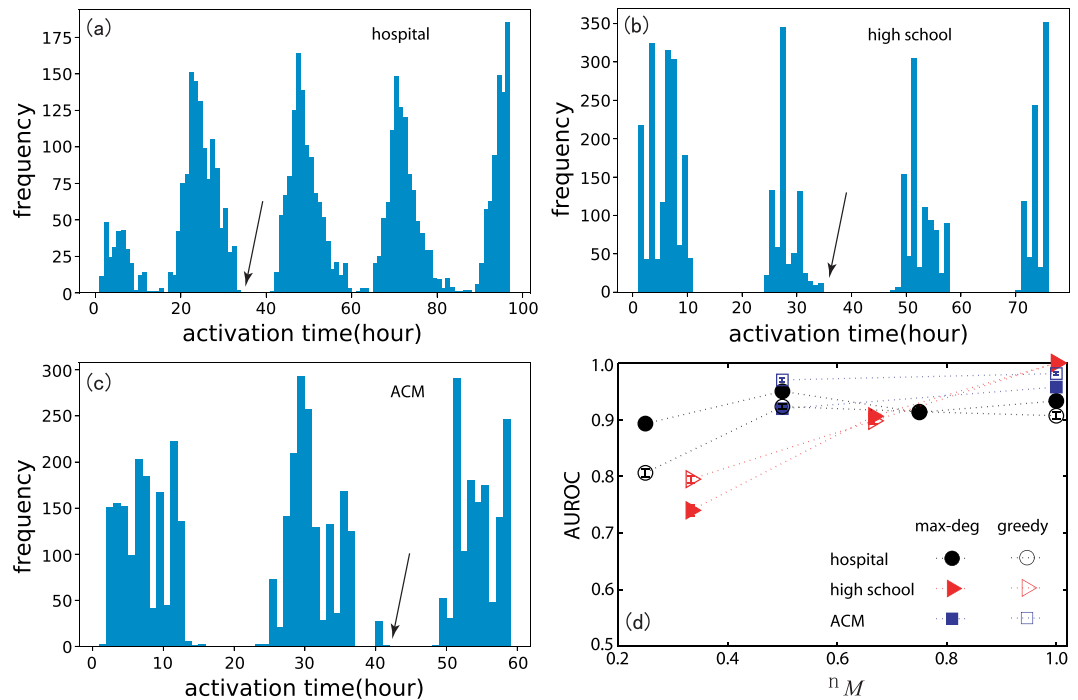


Figure 6. (a–c) Activation time distributions for empirical networks. The dividing point of the dataset is shown in black arrow. (d) Localization performance as a function of n_M for three sources ($N_s = 3$) without noise for greedy algorithm and max-deg strategy. $\beta = 0.05$. The results are obtained by averaging over 500 independent realizations and the vertical bars indicate the standard error. Here we use the first part of the data to choose messengers and locate the sources on the networks constructed with the second part. The time window is a day.

and estimated observable range. To test the effectiveness of our method under such situation, we divide the time-varying network into two parts according to the order of each edge's activation time: the first part with which a rough network is constructed and a set of messengers is selected, and the second part within which the source localization is applied. Figure 6(a–c) display the activation time distributions of the three empirical networks, which indicates circadian rhythms, and illustrate the dividing time point used in the simulation. Messengers are selected using greedy algorithm and max-deg strategy ensuring full observable of the first part network, and are further used to locate the sources on the second part network. As shown in Fig. 6(d), our sources localization method shows a good performance for both strategies on the empirical networks.

Discussions

Source localization is significant for preventing negative diffusion processes and reducing damages. Combining structural observability theory with sparse signal reconstruction, we succeed in developing a general framework for locating multiple diffusion sources in time varying networks, an extremely challenging problem in complex dynamical systems. The framework allows us to define an observable centrality for each node and to locate any number of sources by observing a small number of messenger nodes with larger values of observable centrality and exploiting the natural sparsity of sources. Appealing features of our framework include requirement of only small amounts of measurements and robustness against noise and uncertainties in system parameters. We offer analytic formulas for the observable centrality and the minimum number of messenger nodes, which are validated using model and empirical networks. A general finding based on our framework is that large degree nodes produce more valuable information than small degree nodes, an opposite result to that for static networks based on structural observability theory. As a result, choosing larger degree nodes as messenger nodes is more efficient to locate multiple sources in time varying nodes; in contrast, small degree nodes are often selected as messenger nodes in static networks. A counterintuitive finding is that sources in a more rapid varying network can be located more readily than in a slowly changed network. A heuristic explanation for this phenomenon is that frequent changes of the network structure in general produce more independent path in the static mapping of the original time varying network. As a result, the number of necessary messenger nodes is reduced and the sources become relatively easier to be localized. When dealing with time-varying networks, forward-planning problem is an unavoidable issue, because in many real systems the future structure of the time-varying network cannot be obtained in advance. While if the network structure evolves periodically or following some patterns, we can select messengers by fully exploiting the structural information embedded in the past interactions; If the evolution of time-varying network is totally random, then selecting messengers randomly may be the only way. In this paper, multiplicative noise is considered to test the robustness of our method, although the average performance is still satisfied, the worst cases are even worse than that of random guess ($AUROC < 0.5$) when the noise is strong. Therefore, it is very important to develop a more robust and efficient inference framework that can

deal with different noise settings. One possible improvement is relaxing the object function $Y = O \cdot X$ to $\|Y - O \cdot X(0)\|_2 + \lambda \|X(0)\|_1$ in the cost of adding a tuning parameter λ . Another possible way is to develop a probabilistic approach which can utilize the distribution of noise to give a maximum likelihood estimation of the sources.

Our framework has potential applications in addressing many problems relevant to source localization, such as consensus, synchronization on power grid networks, locating the sources of epidemic spreading and rumor spreading in society, online social communities and computer networks. Moreover, our work has implications in disease diagnosis and therapy, such as identify focus sources of epilepsy and tumors in human body. Because of the significance and broad application potential of the source localization problem, we expect that the theory and practical algorithms presented in this work will stimulate further efforts, e.g., a more efficient and accurate algorithm to identify a minimum set of messenger nodes and a new framework available for systems with strong nonlinear properties.

Data availability statement. Data can be accessed at <http://www.sociopatterns.org/datasets>.

References

- Vespignani, A. Modelling dynamical processes in complex socio-technical systems. *Nat. Phys.* **8**, 32–39 (2012).
- Gomez, S. *et al.* Diffusion dynamics on multiplex networks. *Phys. Rev. Lett.* **110**, 028701 (2013).
- Pope, C. A. III *et al.* Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *Jama* **287**, 1132–1141 (2002).
- Shao, M., Tang, X., Zhang, Y. & Li, W. City clusters in china: air and surface water pollution. *Front. Ecol. Environ.* **4**, 353–361 (2006).
- Neuman TM, G., Noda, T. & Kawaoka, Y. Emergence and pandemic potential of swine-origin h1n1 influenza virus. *Nature* **459**, 931–939 (2009).
- Hvistendahl, M., Normile, D. & Cohen, J. Despite large research effort, h7n9 continues to baffle. *Science* **340**, 414–415 (2013).
- Lloyd, A. L. & May, R. M. How viruses spread among computers and people. *Science* **292**, 1316–1317 (2001).
- Wang, P., González, M. C., Hidalgo, C. A. & Barabási, A.-L. Understanding the spreading patterns of mobile phone viruses. *Science* **324**, 1071–1076 (2009).
- Centola, D. The spread of behavior in an online social network experiment. *Science* **329**, 1194–1197 (2010).
- Pinto, P. C., Thiran, P. & Vetterli, M. Locating the source of diffusion in large-scale networks. *Phys. Rev. Lett.* **109**, 068702 (2012).
- Lokhov, A. Y., Mézard, M., Ohta, H. & Zdeborová, L. Inferring the origin of an epidemic with a dynamic message-passing algorithm. *Phys. Rev. E* **90**, 012801 (2014).
- Altarelli, F., Braunstein, A., Dall'Asta, L., Lage-Castellanos, A. & Zecchina, R. Bayesian inference of epidemics on networks via belief propagation. *Phys. Rev. Lett.* **112**, 118701 (2014).
- Brockmann, D. & Helbing, D. The hidden geometry of complex, network-driven contagion phenomena. *Science* **342**, 1337–1342 (2013).
- Zhu, K. & Ying, L. Information source detection in the sir model: A sample-path-based approach. *IEEE/ACM Trans. Netw.* **24**, 408–421 (2016).
- Shen, Z., Cao, S., Wang, W.-X., Di, Z. & Stanley, H. E. Locating the source of diffusion in complex networks by time-reversal backward spreading. *Phys. Rev. E* **93**, 032301 (2016).
- Kitsak, M. *et al.* Identification of influential spreaders in complex networks. *Nat. Phys.* **6**, 888–893 (2010).
- Morone, F. & Makse, H. A. Influence maximization in complex networks through optimal percolation. *Nature* **524**, 65–68 (2015).
- Fu, L., Shen, Z., Wang, W.-X., Fan, Y. & Di, Z. Multi-source localization on complex networks with limited observers. *Europhys. Lett.* **113** (2016).
- Zejniliovic, S., Gomes, J. & Sinopoli, B. In 51st Annual Allerton Conference on. IEEE, 2013.
- Zhu, K., Chen, Z. & Ying, L. Locating the contagion source in networks with partial time stamps. *Data Mining and Knowledge Discovery* **30**, 1217–1248 (2015).
- Holme, P. & Saramäki, J. Temporal networks. *Phys. Rep.* **519**, 97–125 (2012).
- Antulov-Fantulin, N., Lancic, A., Smuc, T., Stefancic, H. & Sikic, M. Identification of patient zero in static and temporal networks: Robustness and limitations. *Phys. Rev. Lett.* **114**, 248701 (2015).
- Hu, Z.-L., Han, X., Lai, Y.-C. & Wang, W.-X. Optimal localization of diffusion sources in complex networks. *Royal Society Open Science* **4**, 170091 (2017).
- Wang, W.-X., Lai, Y.-C. & Grebogi, C. Data based identification and prediction of nonlinear and complex dynamical systems. *Phys. Rep.* **644**, 1–76 (2016).
- Shields, R. W. & Pearson, J. B. Structural controllability of multi-input linear systems. *Rice University ECE Technical Report* (1975).
- Mayeda, H. On structural controllability theorem. *IEEE Trans. Automat. Contr.* **26**, 795–798 (1981).
- Liu, Y.-Y., Slotine, J.-J. & Barabási, A.-L. Observability of complex systems. *Proc. Natl. Acad. Sci.* **110**, 2460–2465 (2013).
- Pósfai, M. & Hövel, P. Structural controllability of temporal networks. *New J. Phys.* **16**, 123055 (2014).
- Pósfai, M. *Structure and controllability of complex networks*. Ph.D. thesis, Eötvös Loránd University, Budapest (2014).
- Nemhauser, G. L., Wolsey, L. A. & Fisher, M. L. An analysis of approximations for maximizing submodular set functions I. *Math. Program.* **14**, 265–294 (1978).
- Golovin, D. & Krause, A. Adaptive submodularity: Theory and applications in active learning and stochastic optimization. *J. Artif. Intell. Res.* **42**, 427–486 (2011).
- Erds, P. & Rényi, A. On the evolution of random graphs. *Publ. Math. Inst. Hungar. Acad. Sci.* **5**, 17–61 (1960).
- Barabasi, A. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509–512 (1999).
- Liu, Y.-Y., Slotine, J.-J. & Barabási, A.-L. Controllability of complex networks. *Nature* **473**, 167–173 (2011).
- Candès, E. J., Romberg, J. & Tao, T. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Theory* **52**, 489–509 (2006).
- Donoho, D. L. Compressed sensing. *IEEE Trans. Inf. Theory* **52**, 1289–1306 (2006).
- Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology* **143**, 29–36 (1982).

Acknowledgements

We thank M. Pósfai for valuable discussion. Z.-L.H and Z.S contributed equally to this work. This work was supported by NSFC under Grant Nos. 61174150, 61573064, 71401037 and 61074116, the Fundamental Research Funds for the Central Universities and the Beijing Nova Program. 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 ONR through Grant No. N00014-16-1-2828.

Author Contributions

Conceived and designed the research: W.W.X., P.B., L.Y.C., C.S. Performed the research: H.Z.L., S.Z., W.W.X., C.S. Wrote the paper: P.B., Y.H., W.W.X., L.Y.C.

Additional Information

Supplementary information accompanies this paper at <https://doi.org/10.1038/s41598-018-20033-9>.

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2018