ARTICLE

OPEN

Reconstructing propagation networks with natural diversity and identifying hidden sources

Zhesi Shen[1], Wen-Xu Wang[1,2], Ying Fan[1], Zengru Di[1] & Ying-Cheng Lai[2,3]

Our ability to uncover complex network structure and dynamics from data is fundamental to understanding and controlling collective dynamics in complex systems. Despite recent progress in this area, reconstructing networks with stochastic dynamical processes from limited time series remains to be an outstanding problem. Here we develop a framework based on compressed sensing to reconstruct complex networks on which stochastic spreading dynamics take place. We apply the methodology to a large number of model and real networks, finding that a full reconstruction of inhomogeneous interactions can be achieved from small amounts of polarized (binary) data, a virtue of compressed sensing. Further, we demonstrate that a hidden source that triggers the spreading process but is externally inaccessible can be ascertained and located with high confidence in the absence of direct routes of propagation from it. Our approach thus establishes a paradigm for tracing and controlling epidemic invasion and information diffusion in complex networked systems.

[1] School of Systems Science, Beijing Normal University, Beijing 100875, China. [2] School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, Arizona 85287, USA. [3] Department of Physics, Arizona State University, Tempe, Arizona 85287, USA. Correspondence and requests for materials should be addressed to W.-X.W. (email: wenxuwang@bnu.edu.cn).

1

One of the outstanding problems in interdisciplinary science is nonlinear and complex systems identification, prediction and control. Given a complex dynamical system, the various types of dynamical processes are of great interest. The ultimate goal in the study of complex systems is to devise practically implementable strategies to control the collective dynamics. A great challenge is that the network structure and the nodal dynamics are often unknown but only limited measured time series are available. To control the system dynamics, it is imperative to be able to map out the system details from data. Reconstructing complex network structure and dynamics from data, the inverse problem, has thus become a central issue in contemporary network science and engineering[1–8]. There are broad applications of the solutions of the network reconstruction problem, due to the ubiquity of complex interacting patterns arising from many systems in a variety of disciplines[9–12].

An important class of collective dynamics is epidemic spreading and information diffusion in the human society or on computer networks[13–20]. The past decades have witnessed severe epidemic outbreaks at the global scale due to the mutation of virus, including SARS[21,22], H5N1 (refs 23,24), H1N1 (refs 25,26) and the recent invasion of H7N9 in eastern China[27,28]. Our goal is to reconstruct the networks hosting the spreading process and identify the source of spreading using limited measurements. This is especially challenging due to (1) difficulty in predicting and monitoring mutations of deadly virus and (2) absence of epidemic threshold in heterogeneous networks[29–32]. Another example is rumour propagation in the online virtual communities, which can cause financial loss or even social instabilities, such as the 2011 irrational and panicked acquisition of salt in southeast Asian countries caused by the nuclear leak in Japan. In this regard, identifying the propagation network for controlling the dynamics is of great interest. Another significant challenge in reconstructing a spreading network lies in the nature of the available time series: they are polarized, despite stochastic spreading among nodes. Indeed, the link pattern and the probability of infection are encrypted in the binary status of individuals, infected or not, analogous to the collapse of wave function to one associated with some discrete quantum state induced by observation in quantum mechanics.

There have been recent efforts in addressing the inverse problem of some special types of complex propagation networks[33,34]. In particular, for diffusion process originated from a single source, the routes of diffusion from the source constitute a tree-like structure. If information about the early stage of the spreading dynamics is available, it would be feasible to decode all branches that reveal the connections from the source to its neighbours, and then to their neighbours and so on. Taking into account the time delays in the diffusion process enables a straightforward inference of the source in a complex network through enumerating all possible hierarchical trees[33,34]. However, if no immediate information about the diffusion is available, the tree-structure-based inference method is inapplicable, and the problem of network reconstruction and locating the source becomes extremely challenging, hindering control of diffusion and delivery of immunization. The loss of knowledge about the source is common in real situations. For example, passengers on an international flight can carry a highly contagious disease, making certain airports the immediate neighbours of the hidden source, which would be difficult to trace. In another example, the source could be migratory birds coming from other countries or continents. A general data-driven approach, applicable in such scenarios, is still lacking.

In this paper, we develop a general theoretical framework to reconstruct complex propagation networks from time series based on the compressed sensing theory (CST)[35–40], a novel optimization paradigm for sparse-signal reconstruction with broad applications in signal and image processing. Owing to the striking characteristics of CST such as the extremely low data requirement and rigorous guarantee of convergence to optimal solutions, our framework is highly efficient and accurate. However, casting the inverse problem into the CST framework is highly non-trivial. Although CST has been used to uncover the nodal interaction patterns for coupled oscillator networks or evolutionary games from time series[41–43], the dynamics of epidemic propagation is typically highly stochastic with, for example, binary time series, rendering inapplicable the existing CST-based formulation. Further, despite the alternative sparsity enforcing regularizers and convex optimization used in ref. 44 to infer networks, CST has not been applied to reconstructing propagation networks, especially when the available time series are binary. The main accomplishment of this work is then the development of a scheme to implement the highly non-trivial transformation associated with the spreading dynamics in the paradigm of CST. Without loss of generality, we employ two prototypical models of epidemic spreading: classic susceptible-infected-susceptible (SIS) dynamics[13] and contact processes (CPs)[45,46], on both model and real-world (empirical) networks. Inhomogeneous infection and recovery rates as representative characteristics of the natural diversity are incorporated into the diffusion dynamics to better mimic the real-world situation. We assume that only binary time series can be measured, which characterize the status of any node, infected or susceptible, at any time after the outbreak of the epidemic. The source that triggers the spreading process is assumed to be externally inaccessible (hidden). In fact, one may not even realize its existence from available time series. Our method enables, based on relatively small amounts of data, a full reconstruction of the epidemic spreading network with nodal diversity and successful identification of the immediate neighbouring nodes of the hidden source (thereby ascertaining its existence and uniquely specifying its connections to nodes in the network). The framework is validated with respect to different amounts of data generated from various combinations of the network structures and dynamical processes. High accuracy, high efficiency and applicability in a strongly stochastic environment with measurement noise and missing information are the most striking characteristics of our framework. Thus, broad applications can be expected in addressing significant problems such as targeted control of disease and rumour spreading.

Results

Compressed sensing. The general problem that CST addresses is to reconstruct a vector $\mathbf{X} \in R^N$ from linear measurements $\mathbf{Y}$ about $\mathbf{X}$ in the form

$$\mathbf{Y} = \Phi \cdot \mathbf{X}, \tag{1}$$

where $\mathbf{Y} \in R^M$ and Φ is an $M \times N$ matrix. The striking feature of CS is that the number of measurements can be much less than the number of components of the unknown vector, that is, $M \ll N$, insofar as $\mathbf{X}$ is sparse and the number of non-zero components in it is less than M. Accurate reconstruction can be achieved by solving the following convex-optimization problem[35]:

$$\min \| \mathbf{X} \|_1 \quad \text{subject to} \quad \mathbf{Y} = \Phi \cdot \mathbf{X}, \tag{2}$$

where $\| \mathbf{X} \|_1 = \sum_{i=1}^{N} | \mathbf{X}_i |$ is the L_1 norm of $\mathbf{X}$ and the matrix Φ satisfies the restricted isometry property. Solutions to the convex optimization are now standard[35–40]. (More details of the CST can be found in Supplementary Note 1.) Our goal is to develop a framework to cast the problem of reconstructing propagation networks into form (1).

Reconstruction framework. To present our framework in a transparent manner, we first consider the relatively simple case where there is no hidden source. Further, we assume that the disease starts to propagate from a fraction of the infected nodes. As we will see, based on this framework, it is feasible to locate any hidden source based solely on time series after outbreak of infection. The state of an arbitrary node i is denoted as S_i, where

$$S_i = \begin{cases} 0, & \text{susceptible}; \\ 1, & \text{infected}. \end{cases} \quad (3)$$

Owing to the characteristic difference between the SIS dynamics and CP, we treat them separately (see Methods).

For the SIS dynamics, the probability $P_i^{01}(t)$ of an arbitrary node i being infected by its neighbors at time t is

$$P_i^{01}(t) = 1 - (1 - \lambda_i)^{\sum_{j=1, j \neq i}^{N} a_{ij} S_j(t)}, \quad (4)$$

where λ_i is the infection rate of i, a_{ij} stands for the elements of the adjacency matrix ($a_{ij} = 1$ if i connects to j and $a_{ij} = 0$ otherwise), $S_j(t)$ is the state of node j at t, and the superscript 01 denotes the change from susceptible state (0) to infected state (1). At the same time, the recovery probability of i is $P_i^{10}(t) = \delta_i$, where δ_i is the recovery rate of node i and the superscript 10 denotes the

transition from infected state to susceptible state. Equation (4) can be rewritten as

$$\ln[1 - P_i^{01}(t)] = \ln(1 - \lambda_i) \cdot \sum_{j=1, j \neq i}^{N} a_{ij} S_j(t). \quad (5)$$

If measurements at different times $t = t_1, t_2, \cdots, t_m$ are available, equation (5) can be written in the matrix form $\mathbf{Y}_{m \times 1} = \mathbf{\Phi}_{m \times (N-1)} \cdot \mathbf{X}_{(N-1) \times 1}$, where $\mathbf{Y}$ contains $\ln[1 - P_i^{01}(t)]$ at different t, $\mathbf{\Phi}$ is determined by the state $S_j(t)$ of nodes except i, and $\mathbf{X}$ comprising the links and infection rates of i is sparse for a general network (see Methods). The main challenge here is that the infection probabilities $P_i^{01}(t)$ at different times are not given directly from the time series of the nodal state.

To develop a method to estimate the probability from the nodal states, we set a threshold Δ pertaining to the normalized Hamming distance between strings composed of $S_j(t)$ ($j \neq i$) at different t to identify a base string at $\hat{t}_\alpha$ and a set of strings subject to the base. According to the law of large numbers, the probability $P_i^{01}(\hat{t}_\alpha)$ can be estimated by the average over the state $S_i(t+1)$ at all proper time. By setting another threshold Θ associated with the normalized Hamming distance, we can identify a set of base strings. This process finally gives rise to a set of reconstruction
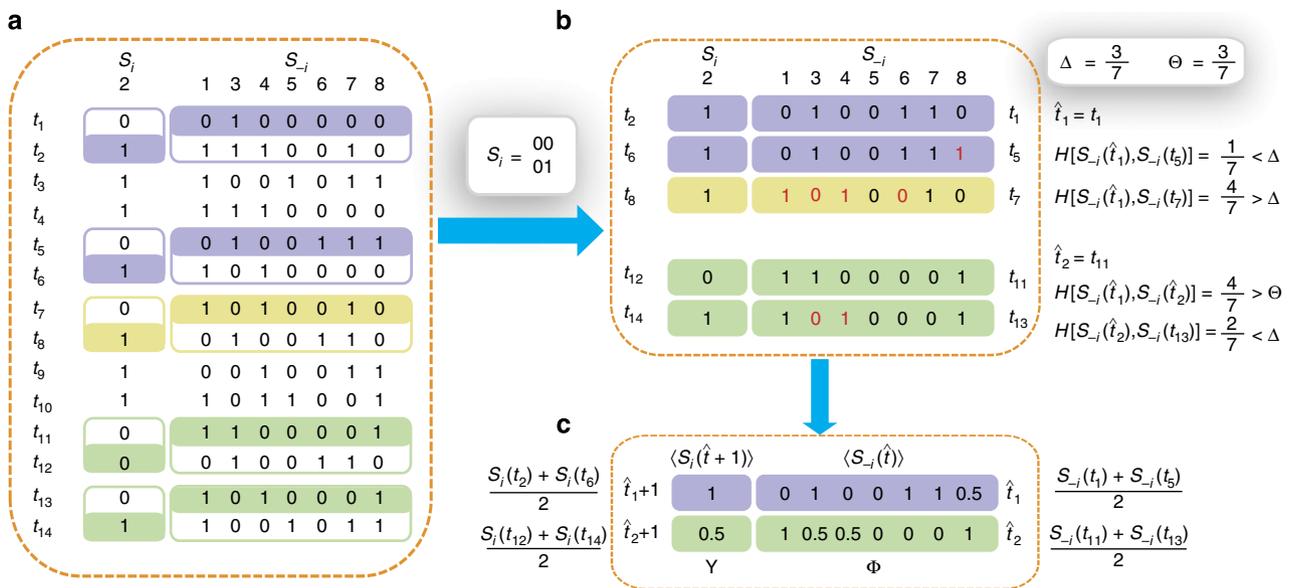

Figure 1 | Schematic illustration of building up Y and Φ from binary time series. (a) Fourteen snapshots of data at time instants $t_1 - t_{14}$ of eight nodes in a sample network, where S_i is the time series of node 2 and S_{-i} denotes the strings of other nodes at different times. The neighbourhood of node 2 is to be reconstructed. Only the pairs 00 and 01 in the time series of S_i ($i = 2$) and the corresponding S_{-i} contain useful information about the network, as marked by different colours. **(b)** As $S_i(t+1)$ is determined by the neighbours of i and $S_{-i}(t)$, we sort out $S_i(t+1)$ and $S_{-i}(t)$ in the coloured sections of the time series in **a**. According to the threshold parameters $\Delta = 3/7$ and $\Theta = 3/7$, we calculate the normalized Hamming distance between each pair of strings $S_{-i}(t)$, finding two base strings at $\hat{t}_1 = t_1$ and $\hat{t}_2 = t_{11}$ with $H[S_{-i}(\hat{t}_1), S_{-i}(\hat{t}_2)] > \Theta$. We separate the coloured strings into two groups that are led by the two base strings, respectively. In each group, the normalized Hamming distance $H[S_{-i}(\hat{t}_\alpha), S_{-i}(t_v)]$ between the base string and other strings is calculated and the difference from $S_{-i}(\hat{t}_\alpha)$ in each string is marked by red. Using parameter Δ, in the group led by $S_{-i}(\hat{t}_1)$, $S_{-i}(t_5)$ and $S_i(t_6)$ are preserved, because of $H[S_{-i}(\hat{t}_1), S_{-i}(t_5)] < \Delta$. In contrast, $S_{-i}(t_7)$ and $S_i(t_8)$ are disregarded because $H[S_{-i}(\hat{t}_1), S_{-i}(t_7)] > \Delta$. In the group led by $S_{-i}(\hat{t}_2)$, due to $H[S_{-i}(\hat{t}_2), S_{-i}(t_{13})] < \Delta$, the string is preserved. The two sets of remaining strings marked by purple and green can be used to yield the quantities required by the reconstruction formula. (Note that different base strings are allowed to share some strings, but for simplicity, this situation is not illustrated here. See Supplementary Note 3 for a detailed discussion.) **(c)** The average values $\langle S_i(\hat{t}_\alpha + 1) \rangle$ and $\langle S_{-i}(\hat{t}_\alpha) \rangle$ used to extract the vector $\mathbf{Y}$ and the matrix $\mathbf{\Phi}$ in the reconstruction formula, where $\langle S_{-i}(\hat{t}_1) \rangle = [S_{-i}(t_1) + S_{-i}(t_5)]/2$, $\langle S_{-i}(\hat{t}_2) \rangle = [S_{-i}(t_{11}) + S_{-i}(t_{13})]/2$, $\langle S_i(\hat{t}_1 + 1) \rangle = [S_i(t_2) + S_i(t_6)]/2$ and $\langle S_i(\hat{t}_2 + 1) \rangle = [S_i(t_{12}) + S_i(t_{14})]/2$ based on the remaining strings marked in different colours (see Methods for more details). CST can be used to reconstruct the neighbouring vector $\mathbf{X}$ of node 2 from $\mathbf{Y}$ and $\mathbf{\Phi}$ from $\mathbf{Y} = \mathbf{\Phi} \cdot \mathbf{X}$.

equations in the matrix form:

$$
\begin{bmatrix}
\ln[1 - \langle S_i(\hat{t}_1 + 1)\rangle] \\
\ln[1 - \langle S_i(\hat{t}_2 + 1)\rangle] \\
\vdots \\
\ln[1 - \langle S_i(\hat{t}_m + 1)\rangle]
\end{bmatrix}
$$

$$
=
\begin{bmatrix}
\langle S_1(\hat{t}_1)\rangle & \cdots & \langle S_{i-1}(\hat{t}_1)\rangle & \langle S_{i+1}(\hat{t}_1)\rangle & \cdots & \langle S_N(\hat{t}_1)\rangle \\
\langle S_1(\hat{t}_2)\rangle & \cdots & \langle S_{i-1}(\hat{t}_2)\rangle & \langle S_{i+1}(\hat{t}_2)\rangle & \cdots & \langle S_N(\hat{t}_2)\rangle \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\langle S_1(\hat{t}_m)\rangle & \cdots & \langle S_{i-1}(\hat{t}_m)\rangle & \langle S_{i+1}(\hat{t}_m)\rangle & \cdots & \langle S_N(\hat{t}_m)\rangle
\end{bmatrix}
$$

$$
\times
\begin{bmatrix}
\ln(1 - \lambda_i)a_{i1} \\
\vdots \\
\ln(1 - \lambda_i)a_{i,i-1} \\
\ln(1 - \lambda_i)a_{i,i+1} \\
\vdots \\
\ln(1 - \lambda_i)a_{iN}
\end{bmatrix},
$$

$$(6)$$

where $\hat{t}_1, \hat{t}_2, \cdots, \hat{t}_m$ correspond to the time associated with m base strings and $\langle \cdot \rangle$ denote the average over all satisfied t (see Methods). The vector $\mathbf{Y}_{m \times 1}$ and the matrix $\mathbf{\Phi}_{m \times (N-1)}$ can then be obtained based solely on time series of nodal states and the vector $\mathbf{X}_{(N-1) \times 1}$ to be reconstructed is sparse, rendering applicable the CS framework. As a result, we can achieve exact

reconstruction of all neighbours of node i from relatively small amounts of observation. In a similar manner the neighbouring vectors of all other nodes can be uncovered from time series, enabling a full reconstruction of the whole network by matching the neighboring sets of all nodes.

For the CP dynamics, the infection probability of an arbitrary node i is given by

$$
P_i^{01}(t) = \lambda_i \sum_{j=1, j \neq i}^{N} a_{ij} S_j(t)/k_i, \tag{7}
$$

where k_i is the degree of the node i, and the recovery probability is $P_i^{10}(t) = \delta_i$ (see Methods). In close analogy to the SIS dynamics, we have

$$
\langle S_i(\hat{t}_\alpha + 1)\rangle \simeq \langle P_i^{01}(\hat{t}_\alpha)\rangle = \frac{\lambda_i \sum a_{ij} \langle S_j(\hat{t}_\alpha)\rangle}{k_i}. \tag{8}
$$

We then choose a series of base strings using a proper threshold Θ to establish a set of equations, expressed in the matrix form $\mathbf{Y}_{m \times 1} = \mathbf{\Phi}_{m \times (N-1)} \cdot \mathbf{X}_{(N-1) \times 1}$ (see Supplementary Note 2), where $\mathbf{\Phi}$ has the same form as in equation (6), but $\mathbf{Y}$ and $\mathbf{X}$ are given by

$$
\mathbf{Y} = \left[\langle S_i(\hat{t}_1 + 1)\rangle, \langle S_i(\hat{t}_2 + 1)\rangle, \cdots, \langle S_i(\hat{t}_m + 1)\rangle\right]^{\mathrm{T}},
$$
$$
\mathbf{X} = \left[\frac{\lambda_i}{k_i} a_{i1}, \cdots, \frac{\lambda_i}{k_i} a_{i,i-1}, \frac{\lambda_i}{k_i} a_{i,i+1}, \cdots, \frac{\lambda_i}{k_i} a_{iN}\right]^{\mathrm{T}}. \tag{9}
$$

Our reconstruction framework based on establishing the vector $\mathbf{Y}$ and the matrix $\mathbf{\Phi}$ is schematically illustrated in Fig. 1. It is noteworthy that our framework can be extended to directed
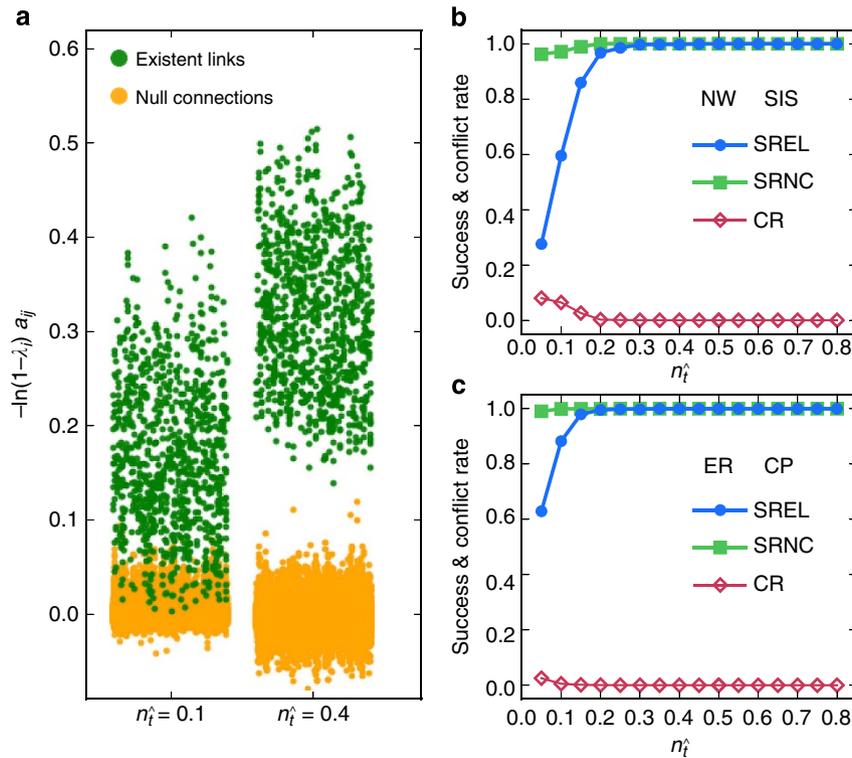

Figure 2 | Network reconstruction performance. (**a**) Element values $\ln(1 - \lambda_i)a_{ij}$ of vector $\mathbf{X}$ times -1 for different fraction $n_{\hat{t}}$ of base strings for SIS dynamics. (**b,c**) Success rate (SREL and SRNC) and conflict rate (CR) of reconstruction as a function of $n_{\hat{t}}$ for SIS dynamics on Newman–Watts (NW) small-world networks (**b**) and CP dynamics on Erdös–Rényi (ER) random networks (**c**). For the SIS dynamics, the parameters are $\Theta = 0.25$, $\Delta = 0.45$ and the infection and recovery rates λ_i and δ_i are randomly distributed in the ranges (0.2, 0.4) and (0.4, 0.6), respectively. For the CP dynamics, the parameters are $\Theta = 0.35$, $\Delta = 0.45$ and λ_i and δ_i are randomly distributed in the ranges (0.7, 0.9) and (0.2, 0.4), respectively. The network size N is 200 with average node degree $\langle k \rangle = 4$. The results are obtained by ensemble averaging over ten independent realizations. The success rate is determined by setting a cut-off according to Supplementary Fig. 1a and the method described in Supplementary Note 4.

networks in a straightforward manner due to the feature that the neighbouring set of each node can be independently reconstructed. For instance, the neighbouring vector $\mathbf{X}$ can be defined to represent a unique link direction, for example, incoming links. Inference of the directed links of all nodes yields the full topology of the entire directed network.

Reconstructing networks and infection and recovery rates. To quantify the performance of our method in terms of the number of base strings (equations) for a variety of diffusion dynamics and network structures, we study the success rates for existent links (SRELs) and null connections (SRNCs), corresponding to non-zero and zero element values in the adjacency matrix, respectively. We impose the strict criterion that the network is regarded to have been fully reconstructed if and only if both success rates reach 100%. The sparsity of links makes it necessary to define SREL and SRNC separately. As the reconstruction method is implemented for each node in the network, we define SREL and SRNC on the basis of each individual node and the two success rates for the entire network are the respective averaged values over all nodes. We also consider the issue of trade-off in terms of the true positive rate (for correctly inferred links) and the false positive rate (for incorrectly inferred links).

Here we assume that there is no hidden source and the spreading process starts from a fraction of infected nodes, and record the binary time series. Figure 2a shows the reconstructed values of the components of the neighbouring vector $\mathbf{X}$ of all nodes. Let $n_{\hat{t}}$ be the number of base strings normalized by the network size N. For small values of $n_{\hat{t}}$, for example, $n_{\hat{t}} = 0.1$, the values of elements associated with links and that associated with null connections (actual zeros in the adjacency matrix) overlap, leading to ambiguities in the identification of links. In contrast, for larger values of $n_{\hat{t}}$, for example, $n_{\hat{t}} = 0.4$, an apparent gap emerges between the two groups of element values, enabling us to correctly identify all links by simply setting a cut-off within the gap (see Supplementary Fig. 1a and Supplementary Note 4 for a method to set the cut-off). The success rates (SREL and SRNC) as a function of $n_{\hat{t}}$ for SIS and CP on both homogeneous and heterogeneous networks are shown in Fig. 2b,c, where we observe nearly perfect reconstruction of links insofar as $n_{\hat{t}}$ exceeds a relatively small value—an advantage of compressed sensing. The exact reconstruction is robust in the sense that a wide range of $n_{\hat{t}}$ values can yield nearly 100% success rates. Our reconstruction method is then effective for tackling real networks in the absence of any *a priori* knowledge about its topology. In particular, the existence of a clear gap in the reconstructed vector $\mathbf{X}$ represents a successful reconstruction for a real network.

Note that a network is reconstructed through the union of all neighbourhoods, which may encounter 'conflicts' with respect to the presence/absence of a link between two nodes as generated by reconstruction centred at the two nodes. The conflicts would reduce the accuracy in the reconstruction of the entire network. To characterize the effects of edge conflicts, we study the consistency of mutual assessment of the presence or absence of link between each pair of nodes, as shown in Fig. 2b,c. We see that inconsistency arises for small values of $n_{\hat{t}}$ but vanishes completely when the success rates reach 100%, indicating complete consistency among the mutual inferences of nodes and consequently guaranteeing accurate reconstruction of the entire network. Detailed results of success rates and trade-off measures with respect to a variety of model and real networks are displayed in Table 1, Supplementary Figs 2 and 3. and Supplementary Note 5.

Although the number of base strings is relatively small compared with the network size, we need a set of strings at different time with respect to a base string to formulate the

mathematical framework for reconstruction. We study how the length of time series affects the accuracy of reconstruction. Figure 3a,b show the success rate as a function of the relative length n_t of time series for SIS and CP dynamics on both homogeneous and heterogeneous networks, where n_t is the total length of time series from the beginning of the spreading process divided by the network size N. The results demonstrate that even for very small values of n_t, most links can already be identified, as reflected by the high values of the success rate shown. Figure 3c,d show the minimum length $n_t^{\min}$ required to achieve at least 95% success rate for different network sizes. For both SIS and CP dynamics on different networks, $n_t^{\min}$ decreases considerably as N is increased. This seemingly counterintuitive result is due to the fact that different base strings can share strings at different times to enable reconstruction. In general, as N is increased, $n_{\hat{t}}$ will increase accordingly. However, a particular string can belong to different base strings with respect to the threshold Δ, accounting for the slight increase in the absolute length of the time series (see Supplementary Fig. 4 and Supplementary Note 5) and the reduction in $n_t^{\min}$ (see Supplementary Note 3 on the method to choose base and subordinate strings). The dependence of the success rate on the average node degree $\langle k \rangle$ for SIS and CP on different networks has been investigated as well (see Supplementary Fig. 5 and Supplementary Note 5). The results in Figs 2 and 3, Supplementary Figs 2–5 and Table 1 demonstrate the high accuracy and efficiency of our reconstruction method based on small amounts of data.

In practice, noise is present and it is also common for time series from certain nodes to be missing, and it is necessary to test the applicability of our method in more realistic situations. Figure 4a,b show the dependence of the success rate on the fraction n_f of states in the time series that flip due to noise for SIS and CP dynamics on two types of networks. We observe that the success rates are hardly affected, providing strong evidence for the applicability of our reconstruction method. For example, even when 25% of the nodal states flip, we can still achieve about 80% success rates for both dynamical processes and different network topologies. Figure 4c,d present the success rate versus the fraction n_m of unobservable nodes, the states of which are externally inaccessible. We find that the high success rate remains mostly unchanged as n_m is increased from 0 to 25%, a somewhat counterintuitive but striking result. The high degree of robustness against the limit to accessing nodal states is elaborated further in Supplementary Fig. 6 and Supplementary Note 5. We find that, in general, missing information can affect the reconstruction of the neighbouring vector, as reflected by the reduction of the gap between the reconstructed values associated with actual links and null connections. However, even for high values of n_m, for example, $n_m = 0.3$, there is still a clear gap, indicating that a full recovery of all links is achievable. We have also found that our method is robust against inaccurately specified diffusion processes with fluctuation in infection rates (see Supplementary Fig. 7 and Supplementary Note 5). Taken together, the high accuracy, efficiency and robustness against noise, missing information and inaccurately modelling of real dynamical processes provide strong credence for the validity and power of our framework for binary time-series-based network reconstruction.

Having reconstructed the network structure, we can estimate the infection and recovery rates of individuals to uncover their diversity in immunity. This is an essential step to implement target vaccination strategy in a population or on a computer network to effectively suppress/prevent the spreading of virus at low cost, as a large body of literature indicates that knowledge about the network structure and individual characteristics is sufficient for controlling the spreading dynamics[47–50]. Here we

Table 1 | Performances of reconstruction and locating hidden source.

	Network reconstruction				Error of λ			LHS	
	SREL	SRNC	TPR	FPR	Mean	Min	Max	TPR	FPR
SIS									
WS	1.0	1.0	1.0	0.0	0.008	0.0	0.042	1.0	0.001
ER	0.992	0.991	0.992	0.009	0.014	0.0	0.192	0.977	0.026
BA	0.977	0.986	0.977	0.014	0.046	0.0	1.881	0.99	0.066
prison	0.994	0.993	0.994	0.007	0.032	0.0	0.311	1.0	0.017
Santafe	0.978	0.991	0.978	0.009	0.063	0.0	1.122	1.0	0.041
netscience	0.959	0.992	0.959	0.008	0.172	0.0	2.878	0.949	0.144
NW	1.0	0.999	1.0	0.001	0.015	0.0	0.15	1.0	0.001
ZK	0.981	0.993	0.981	0.007	0.079	0.001	0.712	0.97	0.078
Polbooks	0.932	0.984	0.932	0.016	0.133	0.0	1.1875	0.732	0.234
Football	0.970	0.989	0.970	0.011	0.166	0.001	0.651	0.612	0.021
Dolphin	0.952	0.967	0.952	0.033	0.022	0.001	0.123	0.925	0.083
Leadership	0.975	0.986	0.975	0.014	0.05	0.0	0.706	0.869	0.057
CP									
WS	1.0	1.0	1.0	0.0	0.009	0.0	0.048	0.936	0.068
ER	0.999	1.0	1.0	0.0	0.01	0.0	0.073	0.925	0.327
BA	0.997	1.0	1.0	0.0	0.008	0.0	0.043	0.943	0.08
Prison	0.995	0.996	0.996	0.004	0.005	0.0	0.018	0.911	0.012
Santafe	0.984	0.996	0.996	0.004	0.006	0.0	0.036	0.929	0.034
Netscience	0.996	0.999	0.996	0.001	0.007	0.0	0.166	1.0	0.050
NW	1.0	1.0	1.0	0.0	0.009	0.0	0.052	0.98	0.034
ZK	0.992	0.992	0.992	0.008	0.007	0.001	0.022	0.977	0.028
Polbooks	0.973	0.995	0.973	0.005	0.008	0.0	0.042	0.829	0.386
Football	0.995	0.997	0.995	0.003	0.006	0.0	0.028	0.517	0.015
Dolphin	0.952	0.971	0.971	0.029	0.006	0.0	0.026	0.786	0.059
Leadership	0.996	0.993	0.993	0.007	0.006	0.0	0.013	0.857	0.075

FPR, false positive rate; LHS, locating hidden source; SREL, success rates for existent link; SRNC, success rates for null connection; TPR, true positive rate.
The accuracy of network reconstruction is quantified by the success rates SREL and SRNC, as well as the trade-off measures TPR and FPR. The accuracy in determining the values of the infection rate λ is characterized by the relative mean errors, the minimum and maximum errors. The accuracy of LHS is characterized by the trade-off measures TPR and FPR. The results of network reconstruction and error in λ are obtained from 30 independent realizations. The results of LHS is obtained from ten dynamical realizations and ten configurations of the hidden source. Other parameters are the same as in Fig. 2. For data sources, reference and network models, see Supplementary Table 1 and Supplementary Note 10.

offer an effective method to infer the individuals' infection rates λ_i based solely on the binary time series of the nodal states after an outbreak of contamination. (To our knowledge, there was no prior work addressing this critical issue.) In particular, after all links have been successfully predicted, λ_i can be deduced from the infection probabilities that can be approximated by the corresponding infection frequencies (see Methods). These probabilities depend on both λ_i and the number of infected neighbours. The reproduced infection rates λ_i of individuals for both SIS and CP dynamics on different networks are in quite good agreement with the true values with small prediction errors (see Supplementary Fig. 8 and Supplementary Note 6). Results from a comprehensive error analysis are listed in Table 1, where the uniformly high accuracy validates our method. The inhomogeneous recovery rates δ_i of nodes can be predicted from the binary time series in a more straightforward way, because δ_i do not depend on the nodal connections (see Supplementary Fig. 9 and Supplementary Note 6). Thus, our framework is capable of predicting characteristics of nodal diversity in terms of degrees, and infection and recovery rates based solely on binary time series of nodal states.

Locating the hidden source of propagation. We assume that a hidden source exists outside the network but there are connections between it and some nodes in the network. In practice, the source can be modelled as a special node that is always infected. Starting from the neighbourhood of the source, the infection originates from the source and spreads all over the network. We collect a set of time series of the nodal states, except the hidden source (see Methods). The basic idea of ascertaining and locating the hidden source is based on missing information from the

hidden source when attempting to reconstruct the network. In particular, to reconstruct the connections belonging to the immediate neighbourhood of the source accurately, time series from the source are needed to generate the matrix $\mathbf{\Phi}$ and the vector $\mathbf{Y}$. However, as the source is hidden, no time series from it are available, leading to reconstruction inaccuracy and, consequently, anomalies in the predicted link patterns of the neighbouring nodes. It is then possible to detect the neighbourhood of the hidden source by identifying any abnormal connection patterns[51], which can be accomplished by using different data segments. If the inferred links of a node are stable with respect to different data segments, the node can be deemed to have no connection with the hidden source; otherwise, if the result of inferring a node's links varies significantly with respect to different data segments, the node is likely to be connected to the hidden source. The s.d. of the predicted results with respect to different data segments can be used as a quantitative criterion for the anomaly. Once the neighbouring set of the source is determined, the source is then precisely located topologically.

Figure 5 presents an example, where a hidden source is connected with four nodes in the network (Fig. 5a), as reflected in the network adjacency matrix (Fig. 5b). We implement our reconstruction framework on each accessible node by using different sets of data in the time series. For each data set, we predict the neighbours of all nodes and generate an adjacency matrix. Averaging over the elements corresponding to each location in all the reconstructed adjacency matrices, we obtain Fig. 5c, in which each row corresponds to the mean number of links in a node's neighbourhood. The inferred links of the immediate neighbours of the hidden source exhibit anomalies. To quantify the anomalies, we calculate the structural s.d. σ from
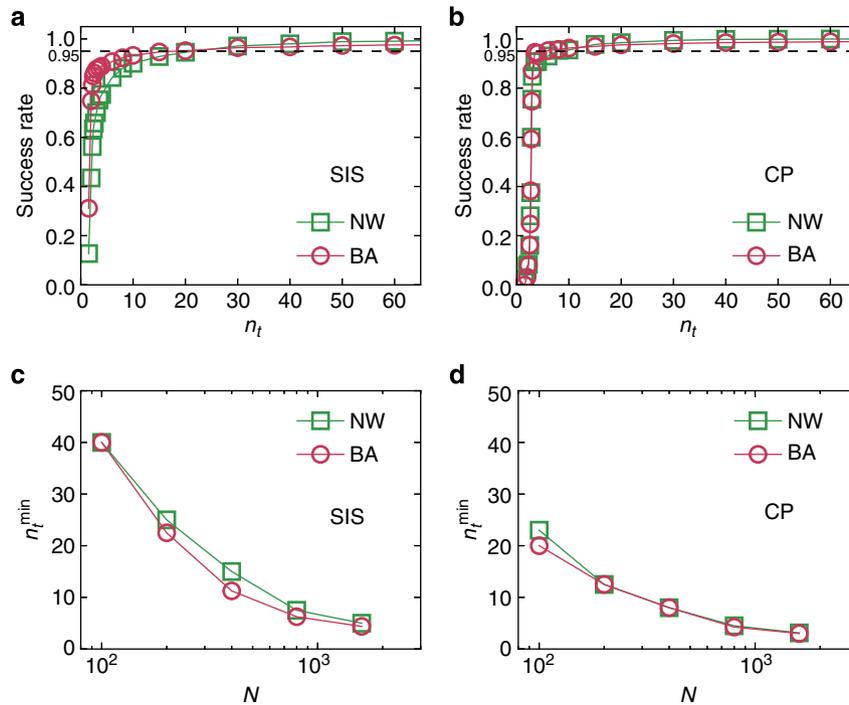
Figure 3 | Effect of the length of time series and network size. (**a,b**) Success rate as a function of the relative length n_t of time series for SIS (**a**) and CP (**b**) in combination with Newman–Watts (NW) and Barabási–Albert (BA) networks. The dashed lines represent 95% success rate. (**c,d**) the minimum relative length n_t^{min} that assures at least 95% success rate as a function of the network size N for SIS (**c**) and CP (**d**) dynamics on NW and BA networks. Here the success rate is the geometric average over SREL and SRNC. For SIS dynamics, $\lambda \in (0.1, 0.3)$ and $\delta \in (0.2, 0.4)$. For CP dynamics, $\lambda \in (0.7, 0.9)$ and $\delta \in (0.3, 0.5)$. In **a** and **b**, the network size N is 500. The other parameters are the same as in Fig. 2. Note that $n_t \equiv t/N$, where t is the absolute length of time series, and $n_t^{min} \equiv t_{min}/N$, where t_{min} is the minimum absolute length of time series required for at least 95% success rate.
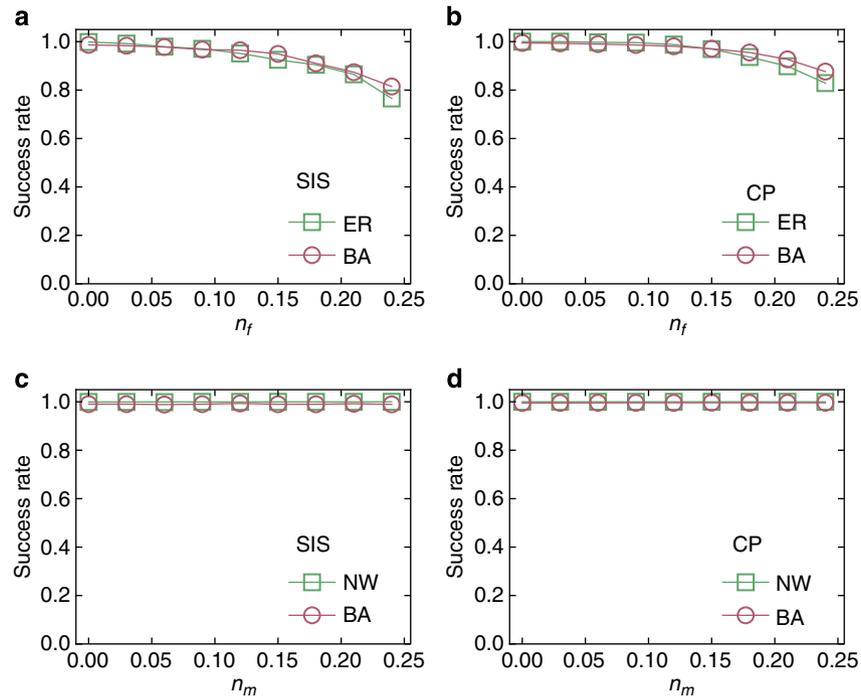

Figure 4 | Reconstruction against noise and inaccessible nodes. (**a,b**) Success rate as a function of the fraction n_f of the flipped states induced by noise in time series for SIS (**a**) and CP (**b**) dynamics on Erdös–Rényi (ER) and Barabási–Albert (BA) networks. (**c,d**) Success rate as a function of the fraction n_m of nodes that are externally inaccessible for SIS (**c**) and CP (**d**) dynamics on NW and BA networks. The network size N is 500 and $\langle k \rangle = 4$. The other parameters are the same as in Fig. 2.
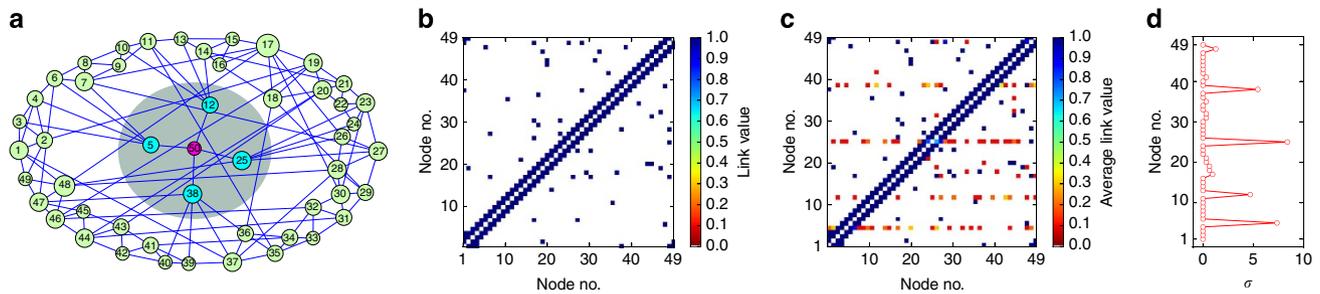
7

Figure 5 | Locating an external hidden source from time series. (**a**) Hidden source treated as a special node (in red) is connected to four nodes in the network (blue). The time series of other nodes except the source (number 50) after the outbreak of an epidemic are assumed to be available. (**b**) True adjacency matrix of the NW network with identical link weights to facilitate a comparison with the reconstructed adjacency matrix. (**c**) Reconstructed adjacency matrix from a number of segments in time series. The four neighbouring nodes of the source are predicted to be densely linked to other nodes, as indicated by the average value of the elements in the four rows corresponding to these nodes. (**d**) Structural variance σ of each node. The four neighbouring nodes of the source exhibit much larger values of σ than those from the other nodes, providing unequivocal evidence that they belong to the immediate neighbourhood of the hidden source.

different data segments, where σ associated with node i is defined through the ith row in the adjacency matrix as

$$\sigma_i = \frac{1}{N} \sum_{j=1}^{N} \sqrt{\frac{1}{g} \sum_{k=1}^{g} (a_{ij}^{(k)} - \langle a_{ij} \rangle)^2}, \qquad (10)$$

where j denotes the column, $a_{ij}^{(k)}$ represents the element value in the adjacency matrix inferred from the kth group of the data, $\langle a_{ij} \rangle = (1/g) \sum_{k=1}^{g} a_{ij}^{k}$ is the mean value of a_{ij} and g is the number of data segments. Applying equation (10) to the reconstructed adjacency matrices gives the results in Fig. 5d, where the values of σ associated with the immediate neighbouring nodes of the hidden source are much larger than those from others (which are essentially zero). A cut-off value can be set in the distribution of σ_i to identify the immediate neighbours of the hidden source (see Supplementary Fig. 1b and Supplementary Note 4). The performance of locating hidden source by means of the trade-off measures (true positive rate versus false positive rate) are displayed in Table 1.

Discussion

We have developed a general framework to reconstruct complex propagation networks on which epidemic spreading takes place from binary time series. Our paradigm is based on compressed sensing, is completely data-driven and practically significant for controlling epidemic spreading through targeted vaccination. Both theoretically and practically, our framework can be used to address the extremely challenging problem of reconstructing the intrinsic interacting patterns of complex stochastic systems based on small amounts of polarized time series. The key to success of our method lies in our development of a novel class of transformation, allowing the network inference problem to be converted to the problem of sparse signal reconstruction, which can then be solved by the standard compressed-sensing algorithm. The accuracy and efficiency of our framework in uncovering the network structure, the natural diversity in the nodal characteristics, and any hidden source are guaranteed by the CST with rigorous proof for low-data requirement and convergence to optimal solution. The feasibility of our framework has been demonstrated using a large number of combinations of epidemic processes and network structures, where in all cases highly accurate reconstruction is achieved. Our approach opens up a new avenue towards fully addressing the inverse problem in complex stochastic systems in a highly efficient manner, a fundamental stepping stone towards understanding and controlling complex dynamical systems in general.

We have focused on two types of spreading dynamics, SIS and CP, where an infected individual can recover and becomes susceptible again. In this regard, even if an outbreak occurs, control strategy such as targeted vaccination or quarantine can be helpful to eliminate the virus eventually. A main purpose of our work is to identify the key individuals in the network to implement target control and to locate the source of infection to isolate it so as to prevent recurrent infection in the future. Although for any spreading dynamics, the most effective way to prevent a large-scale outbreak is to implement control during the early stage, this may be impractical in many situations. If we miss the early stage, which is possible especially in complex networks where the epidemic threshold can be near zero, to be able to reconstruct the spreading network is of tremendous value. Besides disease spreading, our framework is applicable to rumour or information spreading. In this case, identifying the source of rumour is important, a problem that our framework is capable of solving.

Our work raises a number of questions to further and perfect the theoretical and algorithmic development of reconstructing complex dynamical systems. For example, if partial knowledge about the network structure is available, the information can be incorporated into our framework to further reduce the required data amount. Moreover, for non-Markovian spreading processes, our current reconstruction framework may fail. This raises the need to develop new and more general approaches. Nevertheless, our theory, due to its generality and applicability to various types of inhomogeneous interactions, can be applied to networks of networks or interdependent networks, in which there may be different spreading patterns associated with distinct layers or components. Taken together, our results provide strong credence to the proposition that complex networks can be fully decrypted from measurements, even when stochastic disturbance and hidden sources are present. This can offer a deeper understanding of complex systems in general and significantly enhance our ability to control them based on, for example, the recently developed controllability theory of complex networks[52–58].

Methods

Spreading processes. The SIS model is a classic epidemic model that has been used frequently to study a variety of spreading behaviours in social and computer networks. Each node of the network represents an individual and links are connections along which infection can propagate to others with certain probability. At each time step, a susceptible node i in state 0 is infected with rate λ_i if it is connected to an infected node in state 1. If i connects to more than one infected neighbour, the infection probability P^{01} is given by equation (4). At the same time, infected nodes are continuously recovered to be susceptible at the rates δ_i. The CP model has been used extensively to describe, for example, the spreading of infection

and competition of animals over a territory, where λ_i is determined by equation (7). The main difference between SIS and CP dynamics lies in the influence on a node's state from its vicinity. In both SIS and CP dynamics, λ_i and δ_i depend on the individuals' immune systems and are selected from a uniform distribution characterizing the natural diversity (see Supplementary Note 7 for details of numerical simulations). Moreover, a hidden source is regarded as infected for all time.

Mathematical formulation of reconstruction based on CST. For SIS dynamics, suppose measurements at a sequence of times $t = t_1, t_2, \cdots, t_m$ are available. equation (5) leads to the following matrix form $\mathbf{Y}_{m \times 1} = \mathbf{\Phi}_{m \times (N-1)} \cdot \mathbf{X}_{(N-1) \times 1}$:

$$
\begin{bmatrix}
\ln[1 - P_i^{01}(t_1)] \\
\ln[1 - P_i^{01}(t_2)] \\
\vdots \\
\ln[1 - P_i^{01}(t_m)]
\end{bmatrix}
$$

$$
=
\begin{bmatrix}
S_1(t_1) & \cdots & S_{i-1}(t_1) & S_{i+1}(t_1) & \cdots & S_N(t_1) \\
S_1(t_2) & \cdots & S_{i-1}(t_2) & S_{i+1}(t_2) & \cdots & S_N(t_2) \\
\vdots & & \vdots & \vdots & & \vdots \\
S_1(t_m) & \cdots & S_{i-1}(t_m) & S_{i+1}(t_m) & \cdots & S_N(t_m)
\end{bmatrix}
\begin{bmatrix}
\ln(1-\lambda_i)a_{i1} \\
\vdots \\
\ln(1-\lambda_i)a_{i,i-1} \\
\ln(1-\lambda_i)a_{i,i+1} \\
\vdots \\
\ln(1-\lambda_i)a_{iN}
\end{bmatrix},
$$

where the vector $\mathbf{X}_{(N-1) \times 1}$ contains all possible connections between node i and all other nodes, and it is sparse for a general complex network. We see that if the vector $\mathbf{Y}_{m \times 1}$ and the matrix $\mathbf{\Phi}_{m \times (N-1)}$ can be constructed from time series, $\mathbf{X}_{(N-1) \times 1}$ can then be solved by using CST. The main challenge here is that the infection probabilities $P_i^{01}(t)$ at different times are not given directly by the time series of the nodal states. To devise a heuristic method to estimate the probabilities, we assume that the neighbouring set Γ_i of the node i is known. The number of such neighbouring nodes is given by k_i, the degree of node i and their states at time t can be denoted as

$$S_{\Gamma_i}(t) \equiv \{S_1(t), S_2(t), \cdots, S_{k_i}(t)\}. \tag{11}$$

To approximate the infection probability, we use $S_i(t) = 0$ so that at $t+1$, the node i can be infected with certain probability. In contrast, if $S_i(t) = 1$, $S_i(t+1)$ is only related with the recovery probability δ_i. Hence, we focus on the $S_i(t) = 0$ case to derive $P_i^{01}(t)$. If we can find two time instants: $t_1, t_2 \in T$ (T is the length of time series), such that $S_i(t_1) = 0$ and $S_i(t_2) = 0$, we can then calculate the normalized Hamming distance $H[S_{\Gamma_i}(t_1), S_{\Gamma_i}(t_2)]$ between $S_{\Gamma_i}(t_1)$ and $S_{\Gamma_i}(t_2)$, where the normalized Hamming distance between two strings of equal length is defined as ratio of the number of positions with different symbols between them and the length of string. If $H[S_{\Gamma_i}(t_1), S_{\Gamma_i}(t_2)] = 0$, we can regard the states at the next time step, $S_i(t_1 + 1)$ and $S_i(t_2 + 1)$, as i.i.d Bernoulli trials. In this case, using the law of large numbers, we have

$$\lim_{l \to \infty} \frac{1}{l} \sum_{v=1}^{l} S_i(t_v + 1) \to P_i^{01}(\hat{t}_\alpha), \quad \forall\, t_v, S_i(t_v) = 0, \; H[S_{\Gamma_i}(\hat{t}_\alpha), S_{\Gamma_i}(t_v)] = 0. \tag{12}$$

A more intuitive understanding of (equation (12)) is that if the states of i's neighbours are unchanged, the fraction of times of i being infected by its neighbours over the entire time period will approach the actual infection probability P_i^{01}. Note, however, that the neighbouring set of i is unknown and to be inferred. A strategy is then to artificially enlarge the neighbouring set $S_{\Gamma_i}(t)$ to include all nodes in the network, except i. In particular, we denote

$$S_{-i}(t) \equiv \{S_1(t), S_2(t), \dots, S_{i-1}(t), S_{i+1}(t), \dots, S_N(t)\}. \tag{13}$$

If $H[S_{-i}(t_1), S_{-i}(t_2)] = 0$, the condition $H[S_{\Gamma_i}(t_1), S_{\Gamma_i}(t_2)] = 0$ will be ensured. Consequently, due to the nature of i.i.d Bernoulli trials, from the law of large numbers, we have

$$\lim_{l \to \infty} \frac{1}{l} \sum_{v=1}^{l} S_i(t_v + 1) \to P_i^{01}(\hat{t}_\alpha), \quad \forall\; t_v, S_i(t_v) = 0, \; H[S_{-i}(\hat{t}_\alpha), S_{-i}(t_v)] = 0. $$

Hence, the infection probability $P_i^{01}(\hat{t}_\alpha)$ of a node at $\hat{t}_\alpha$ can be evaluated by averaging over its states associated with zero-normalized Hamming distance between the strings of other nodes at some time associated with $\hat{t}_\alpha$. In practice, to find two strings with absolute zero-normalized Hamming distance is unlikely. We thus set a threshold Δ so as to pick the suitable strings to approximate the law of large numbers, that is

$$\frac{1}{l} \sum_{v=1}^{l \gg 1} S_i(t_v + 1) \simeq \frac{1}{l} \sum_{v=1}^{l \gg 1} P_i^{01}(t_v), \quad \forall\; t_v, S_i(t_v) = 0, \; H[S_{-i}(\hat{t}_\alpha), S_{-i}(t_v)] < \Delta, \tag{14}$$

where $S_{-i}(\hat{t}_\alpha)$ serves as a base for comparison with $S_{-i}(t)$ at all other times and $\frac{1}{l} \sum_{v=1}^{l \gg 1} P_i^{01}(t_v) \simeq P_i^{01}(\hat{t}_\alpha)$. As $H[S_{-i}(\hat{t}_\alpha), S_{-i}(t_v)]$ is not exactly zero, there is a small difference between $P_i^{01}(\hat{t}_\alpha)$ and $P_i^{01}(t_v)$ ($v = 1, \cdots, l$). We thus consider the average of $P_i^{01}(t_v)$ for all t_v to obtain $P_i^{01}(\hat{t}_\alpha)$, leading to the right-hand side of equation (14). We denote $\langle S_i(\hat{t}_\alpha + 1) \rangle = \frac{1}{l} \sum_{v=1}^{l \gg 1} S_i(t_v + 1)$ and $\langle P_i^{01}(\hat{t}_\alpha) \rangle = \frac{1}{l} \sum_{v=1}^{l \gg 1} P_i^{01}(t_v)$. To reduce the error in the estimation, we implement the average on

$S_{-i}(t)$ over all selected strings through equation (14). The averaging process is with respect to the nodal states $S_{j, j \neq i}(t)$ on the right-hand side of the modified dynamical equation (5). Specifically, averaging over time t restricted by equation (14) on both sides of equation (5), we obtain $\langle \ln[1 - P_i^{01}(t)] \rangle = \ln(1 - \lambda_i) \sum_{j=1, j \neq i}^{N} a_{ij} \langle S_j(t) \rangle$. If λ_i is small with insignificant fluctuations, we can approximately have $\ln[1 - \langle P_i^{01}(t) \rangle] \simeq \langle \ln[1 - P_i^{01}(t)] \rangle$ (see Supplementary Fig. 10 and Supplementary Note 8), which leads to $\ln[1 - \langle P_i^{01}(t) \rangle] \simeq \ln(1 - \lambda_i) \sum_{j=1, j \neq i}^{N} a_{ij} \langle S_j(t) \rangle$. Substituting $\langle P_i^{01}(\hat{t}_\alpha) \rangle$ by $\langle S_i(\hat{t}_\alpha + 1) \rangle$, we finally get

$$\ln[1 - \langle S_i(\hat{t}_\alpha + 1) \rangle] \simeq \ln(1 - \lambda_i) \cdot \sum_{j=1, j \neq i}^{N} a_{ij} \langle S_j(\hat{t}_\alpha) \rangle. \tag{15}$$

Although the above procedure yields an equation that bridges the links of an arbitrary node i with the observable states of the nodes, a single equation does not contain sufficient structural information about the network. Our second step is then to derive a sufficient number of linearly independent equations required by CST to reconstruct the local connection structure. To achieve this, we choose a series of base strings at a number of time instants from a set denoted by T_{base}, in which each pair of strings satisfy

$$H[S_{-i}(\hat{t}_\beta), S_{-i}(\hat{t}_\alpha)] > \Theta, \quad \forall \hat{t}_\alpha, \hat{t}_\beta \in T_{\text{base}}, \tag{16}$$

where $\hat{t}_\alpha$ and $\hat{t}_\beta$ correspond to the time instants of two base strings in the time series and Θ is a threshold. For each string, we repeat the process of establishing the relationship between the nodal states and connections, leading to a set of equations at different values of $\hat{t}_\alpha$ in equation (15), as described in the matrix form (equation (6)). See Supplementary Fig. 11, 12 and Supplementary Note 8 for the dependence of success rate on threshold Δ and Θ for SIS and CP dynamics in combination with four types of networks.

Inferring inhomogeneous infection rates. The values of the infection rate λ_i of nodes can be inferred after the neighbourhood of each node has been successfully reconstructed. The idea roots in the fact that the infection probability of a node approximated by the frequency of being infected calculated from time series is determined both by its infection rate and by the number of infected nodes in its neighbourhood. To provide an intuitive picture, we consider the following simple scenario in which the number of infected neighbours of node i does not change with time. In this case, the probability of i being infected at each time step is fixed. We can thus count the frequency of the 01 and 00 pairs embedded in the time series of i. The ratio of the number of 01 pairs over the total number of 01 and 00 pairs gives approximately the infection probability. The infection rate can then be calculated by using equations (4) and (7) for the SIS and CP dynamics, respectively. In a real-world situation, however, the number of infected neighbours varies with time. The time-varying factor can be taken into account by sorting out the time instants corresponding to different numbers of the infected neighbours, and the infection probability can be obtained at the corresponding time instants, leading to a set of values for the infection rate whose average represents an accurate estimate of the true infection rate for each node.

To be concrete, considering all the time instants t_v associated with k_I infected neighbors, we denote $S_i^{(k_I)} = (1/l) \sum_{v=1}^{l} S_i(t_v + 1)$, $\forall\, t_v$, $\sum_{j \in \Gamma_i} S_j(t_v) = k_I$ and $S_i(t_v) = 0$, where Γ_i is the neighbouring set of node i, k_I is the number of infected neighbours and $S_i^{(k_I)}$ represents the average infected fraction of node i with k_I infected neighbours. Given $S_i^{(k_I)}$, we can rewrite equation (4) by substituting $S_i^{(k_I)}$ for $P_i^{01}(t)$ and $\lambda_i^{(k_I)}$ for λ_i, which yields $\lambda_i^{(k_I)} = 1 - \exp[\ln(1 - S_i^{(k_I)})/k_I]$. To reduce the estimation error, we average $\lambda_i^{(k_I)}$ with respect to different values of k_I, as follows:

$$\lambda_i^{\text{true}}(\text{SIS}) \approx \langle \lambda_i^{(k_I)} \rangle = \frac{1}{N_{\Lambda_i}} \sum_{k_I \in \Lambda_i} \lambda_i^{(k_I)}, \tag{17}$$

where Λ_i denotes the set of all possible infected neighbours during the epidemic process and N_{Λ_i} denotes the number of different values of k_I in the set. Analogously, for CP, we can evaluate λ_i^{true} from equation (7) by

$$\lambda_i^{\text{true}}(\text{CP}) \approx \langle \lambda_i^{(k_I)} \rangle = \frac{1}{N_{\Lambda_i}} \sum_{k_I \in \Lambda_i} \frac{S_i^{(k_I)} k_i}{k_I} \tag{18}$$

where $k_i = \sum_{j=1}^{N} a_{ij}$ is the node degree of i. Insofar as all the links of i have been successfully reconstructed, $S_i^{(k_I)}$ can be obtained from the time series in terms of the satisfied $S_i(t_v + 1)$, allowing us to infer λ_i^{true} via equations (17) and (18).

Note that the method is applicable to any type of networks insofar as the network structure has been successfully reconstructed.

Networks analysed. Model networks and real networks we used are described in Supplementary Note 10 and Table 1.

References

1. Caldarelli, G., Chessa, A., Pammolli, F., Gabrielli, A. & Puliga, M. Reconstructing a credit network. *Nat. Phys.* **9,** 125–126 (2013).
2. Gardner, T. S., di Bernardo, D., Lorenz, D. & Collins, J. J. Inferring genetic networks and identifying compound mode of action via expression profiling. *Science* **301,** 102–105 (2003).

3. Timme, M. Revealing network connectivity from response dynamics. *Phys. Rev. Lett.* **98,** 224101 (2007).

4. Bongard, J. & Lipson, H. Automated reverse engineering of nonlinear dynamical systems. *Proc. Natl Acad. Sci.* **104,** 9943–9948 (2007).

5. Clauset, A., Moore, C. & Newman, M. E. Hierarchical structure and the prediction of missing links in networks. *Nature* **453,** 98–101 (2008).

6. Ren, J., Wang, W.-X., Li, B. & Lai, Y.-C. Noise bridges dynamical correlation and topology in coupled oscillator networks. *Phys. Rev. Lett.* **104,** 058701 (2010).

7. Levnajić, Z. & Pikovsky, A. Network reconstruction from random phase resetting. *Phys. Rev. Lett.* **107,** 034101 (2011).

8. Hempel, S., Koseska, A., Kurths, J. & Nikoloski, Z. Inner composition alignment for inferring directed networks from short time series. *Phys. Rev. Lett.* **107,** 054101 (2011).

9. Albert, R. & Barabási, A.-L. Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74,** 47 (2002).

10. Newman, M. E. The structure and function of complex networks. *SIAM Rev.* **45,** 167–256 (2003).

11. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M. & Hwang, D. -U. Complex networks: Structure and dynamics. *Phys. Rep.* **424,** 175–308 (2006).

12. Newman, M. *Networks: An Introduction* (OUP Oxford, 2009).

13. Pastor-Satorras, R. & Vespignani, A. Epidemic spreading in scale-free networks. *Phys. Rev. Lett.* **86,** 3200–3203 (2001).

14. Eames, K. T. & Keeling, M. J. Modeling dynamic and network heterogeneities in the spread of sexually transmitted diseases. *Proc. Natl Acad. Sci.* **99,** 13330–13335 (2002).

15. Watts, D. J., Muhamad, R., Medina, D. C. & Dodds, P. S. Multiscale, resurgent epidemics in a hierarchical metapopulation model. *Proc. Natl Acad. Sci. USA* **102,** 11157–11162 (2005).

16. Colizza, V., Barrat, A., Barthélemy, M. & Vespignani, A. The role of the airline transportation network in the prediction and predictability of global epidemics. *Proc. Natl Acad. Sci. USA* **103,** 2015–2020 (2006).

17. Gómez-Gardeñes, J., Latora, V., Moreno, Y. & Profumo, E. Spreading of sexually transmitted diseases in heterosexual populations. *Proc. Natl Acad. Sci* **105,** 1399–1404 (2008).

18. Wang, P., González, M. C., Hidalgo, C. A. & Barabási, A.-L. Understanding the spreading patterns of mobile phone viruses. *Science* **324,** 1071–1076 (2009).

19. Merler, S. & Ajelli, M. The role of population heterogeneity and human mobility in the spread of pandemic influenza. *Proc. R. Soc. B* **277,** 557–565 (2010).

20. Balcan, D. & Vespignani, A. Phase transitions in contagion processes mediated by recurrent mobility patterns. *Nat. Phys.* **7,** 581–586 (2011).

21. Riley, S. *et al.* Transmission dynamics of the etiological agent of sars in hong kong: impact of public health interventions. *Science* **300,** 1961–1966 (2003).

22. Marra, M. A. *et al.* The genome sequence of the sars-associated coronavirus. *Science* **300,** 1399–1404 (2003).

23. Ferguson, N. M. *et al.* Strategies for containing an emerging influenza pandemic in southeast asia. *Nature* **437,** 209–214 (2005).

24. Zhang, Y. *et al.* H5n1 hybrid viruses bearing 2009/h1n1 virus genes transmit in guinea pigs by respiratory droplet. *Science* **340,** 1459–1463 (2013).

25. Neumann, G., Noda, T. & Kawaoka, Y. Emergence and pandemic potential of swine-origin h1n1 influenza virus. *Nature* **459,** 931–939 (2009).

26. Smith, G. J. *et al.* Origins and evolutionary genomics of the 2009 swine-origin h1n1 influenza a epidemic. *Nature* **459,** 1122–1125 (2009).

27. Hvistendahl, M., Normile, D. & Cohen, J. Despite large research effort, h7n9 continues to baffle. *Science* **340,** 414–415 (2013).

28. Horby, P. H7n9 is a virus worth worrying about. *Nature* **496,** 399–399 (2013).

29. Boguná, M., Pastor-Satorras, R. & Vespignani, A. Absence of epidemic threshold in scale-free networks with degree correlations. *Phys. Rev. Lett.* **90,** 028701 (2003).

30. Parshani, R., Carmi, S. & Havlin, S. Epidemic threshold for the susceptible-infectious-susceptible model on random networks. *Phys. Rev. Lett.* **104,** 258701 (2010).

31. Castellano, C. & Pastor-Satorras, R. Thresholds for epidemic spreading in networks. *Phys. Rev. Lett.* **105,** 218701 (2010).

32. Gleeson, J. P. High-accuracy approximation of binary-state dynamics on networks. *Phys. Rev. Lett.* **107,** 068701 (2011).

33. Gomez Rodriguez, M., Leskovec, J. & Krause, A. in: *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1019–1028 (ACM, 2010).

34. Pinto, P. C., Thiran, P. & Vetterli, M. Locating the source of diffusion in large-scale networks. *Phys. Rev. Lett.* **109,** 068702 (2012).

35. 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).

36. Candes, E. J., Romberg, J. K. & Tao, T. Stable signal recovery from incomplete and inaccurate measurements. *Commun. Pure Appl. Math.* **59,** 1207–1223 (2006).

37. Donoho, D. L. Compressed sensing. *IEEE Trans. Inf. Theory* **52,** 1289–1306 (2006).

38. Baraniuk, R. G. Compressive sensing [lecture notes]. *Sig. Proc. Mag. IEEE* **24,** 118–121 (2007).

39. Candès, E. J. & Wakin, M. B. An introduction to compressive sampling. *Sig. Proc. Mag. IEEE* **25,** 21–30 (2008).

40. Romberg, J. Imaging via compressive sampling. *Sig. Proc. Mag. IEEE* **25,** 14–20 (2008).

41. Wang, W.-X., Yang, R., Lai, Y. -C., Kovanis, V. & Grebogi, C. Predicting catastrophes in nonlinear dynamical systems by compressive sensing. *Phys. Rev. Lett.* **106,** 154101 (2011).

42. Wang, W. -X., Yang, R., Lai, Y. -C., Kovanis, V. & Harrison, M. A. F. Time-series–based prediction of complex oscillator networks via compressive sensing. *EPL* **94,** 48006 (2011).

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

44. Myers, S. A. & Leskovec, J. On the Convexity of Latent Social Network Inference. *Adv. Neural Inf. Process Syst.* **23,** 1741–1749 (2010).

45. Castellano, C. & Pastor-Satorras, R. Non-mean-field behavior of the contact process on scale-free networks. *Phys. Rev. Lett.* **96,** 038701 (2006).

46. Volz, E. & Meyers, L. A. Epidemic thresholds in dynamic contact networks. *J. R. Soc. Interface* **6,** 233–241 (2009).

47. Cohen, R., Havlin, S. & ben Avraham, D. Efficient immunization strategies for computer networks and populations. *Phys. Rev. Lett.* **91,** 247901 (2003).

48. Forster, G. A. & Gilligan, C. A. Optimizing the control of disease infestations at the landscape scale. *Proc. Natl Acad. Sci* **104,** 4984–4989 (2007).

49. Klepac, P., Laxminarayan, R. & Grenfell, B. T. Synthesizing epidemiological and economic optima for control of immunizing infections. *Proc. Natl Acad. Sci* **108,** 14366–14370 (2011).

50. Kleczkowski, A., Ole, K., Gudowska-Nowak, E. & Gilligan, C. A. Searching for the most cost-effective strategy for controlling epidemics spreading on regular and small-world networks. *J. R. Soc. Interface* **9,** 158–169 (2012).

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

52. Slotine, J.-J. E. *et al. Applied Nonlinear Control* vol. 1 (Prentice hall New Jersey, 1991).

53. Liu, Y.-Y., Slotine, J.-J. & Barabási, A.-L. Controllability of complex networks. *Nature* **473,** 167–173 (2011).

54. Nepusz, T. & Vicsek, T. Controlling edge dynamics in complex networks. *Nat. Phys.* **8,** 568–573 (2012).

55. Yan, G., Ren, J., Lai, Y. -C., Lai, C. -H. & Li, B. Controlling complex networks: How much energy is needed? *Phys. Rev. Lett.* **108,** 218703 (2012).

56. Liu, Y. -Y., Slotine, J. -J. & Barabási, A. -L. Observability of complex systems. *Proc. Natl Acad. Sci* **110,** 2460–2465 (2013).

57. Galbiati, M., Delpini, D. & Battiston, S. The power to control. *Nat. Phys.* **9,** 126–128 (2013).

58. Yuan, Z., Zhao, C., Di, Z., Wang, W. -W. & Lai, Y. -C. Exact controllability of complex networks. *Nat. Commun.* **4,** 2447–1–9 (2013).

Acknowledgements

Author contributions

W.-X.W., Z.S., Y.F., Z.D. and Y.-C.L. designed research; Z.S and W.-X.W. performed research; Y.F. and Z.D. contributed analytic tools; Z.S., W.-X.W., Y.F., Z.D. and Y.-C.L. analysed data; and W.-X.W. and Y.-C.L. wrote the paper.

Additional information

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

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

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Shen, Z. *et al.* Reconstructing propagation networks with natural diversity and identifying hidden sources. *Nat. Commun.* 5:4323 doi: 10.1038/ncomms5323 (2014).