ARTICLES
https://doi.org/10.1038/s43588-022-00217-0

nature computational science



# Autonomous inference of complex network dynamics from incomplete and noisy data

Ting-Ting Gao [1,2] and Gang Yan [1,2,3] ✉

**The availability of empirical data that capture the structure and behaviour of complex networked systems has been greatly increased in recent years; however, a versatile computational toolbox for unveiling a complex system's nodal and interaction dynamics from data remains elusive. Here we develop a two-phase approach for the autonomous inference of complex network dynamics, and its effectiveness is demonstrated by the tests of inferring neuronal, genetic, social and coupled oscillator dynamics on various synthetic and real networks. Importantly, the approach is robust to incompleteness and noises, including low resolution, observational and dynamical noises, missing and spurious links, and dynamical heterogeneity. We apply the two-phase approach to infer the early spreading dynamics of influenza A flu on the worldwide airline network, and the inferred dynamical equation can also capture the spread of severe acute respiratory syndrome and coronavirus disease 2019. These findings together offer an avenue to discover the hidden microscopic mechanisms of a broad array of real networked systems.**

From two-photon calcium imaging of neuronal activities[1,2] and high-throughput genetic experiments[3,4] to digital recordings of human mobility[5–7], our ability to observe the dynamic behaviour of nodes in complex biological, social and technological systems has advanced spectacularly in the past years. The collected observations, often in the form of time-series data, allow us to extract the dynamic patterns of a system's individual nodes. To gain meaningful insights into the system, however, such a reductionist approach of tracking all the individual nodes is insufficient. Indeed, complex system behaviour emerges not just from the single nodes but rather from the dynamic interactions between the nodes[6,8–18]. This requires us to infer complex network dynamics, that is, to retrieve both self-nodal dynamics and interaction dynamics from the accumulating data of network topological structure and nodes' activities.

The balance of self versus interaction dynamics is the most naturally captured by a general equation that tracks the activities of all the nodes via[9]

$$\frac{\mathrm{d}\,\mathbf{x}_i(t)}{\mathrm{d}\,t} = \mathbf{F}\left(\mathbf{x}_i(t)\right) + \sum_{j=1}^{n} A_{ij}\mathbf{G}(\mathbf{x}_i(t), \mathbf{x}_j(t)), \qquad (1)$$

where $\mathbf{x}_i(t) \equiv (x_{i,1}(t),\ldots,x_{i,d}(t))^{\mathrm{T}}$ is node $i$'s $d$-dimensional activity, representing, for example, the membrane potential of a neuron in a brain network[9,12], the proportion of infected people in a country or region[5–7], or the state of a component in an oscillator network[19]. These activities are driven by the self-regulation function $\mathbf{F}(\mathbf{x}_i) \equiv (F_1(\mathbf{x}_i),\ldots,F_d(\mathbf{x}_i))^{\mathrm{T}}$ (designed to describe the dynamics of all the nodes in isolation) and the pairwise function $\mathbf{G}(\mathbf{x}_i(t), \mathbf{x}_j(t)) \equiv (G_1(\mathbf{x}_i,\mathbf{x}_j),\ldots,G_d(\mathbf{x}_i,\mathbf{x}_j))^{\mathrm{T}}$ (which captures the dynamic mechanisms of interaction between the nodes). Finally, the network $A_{ij}$, an $n \times n$ adjacency matrix, denotes the influence or flow from node $j$ to $i$, where $n$ is the number of nodes in the system. As shown in another study, with appropriate choices of nonlinear functions $\mathbf{F}$ and $\mathbf{G}$, equation (1) is able to describe a broad range of complex systems[9]. However, for most real systems, the functions $\mathbf{F}$ and $\mathbf{G}$ are unknown. Hence, a pressing lacuna in the study of complex systems is a versatile computational toolbox for automatically inferring equation (1) from the observed data of network topology $A_{ij}$ and nodes' activities $\mathbf{x}_i(t)$.

Complex biological, social or technological systems lack the fundamental physical rules that govern particle systems; therefore, we do not have a priori knowledge of their internal microscopic mechanisms[20]. Therefore, the goal is not to only identify the model's parameters but rather to retrieve the forms of $\mathbf{F}$ and $\mathbf{G}$ and infer the explicit model itself. Despite the recent important progress in developing methods to infer the governing equations of single- or few-body dynamics[21–27], the task of inferring network dynamics poses particular challenges. For example, $\mathbf{F}$ and $\mathbf{G}$ are usually of different types; hence, one cannot obtain their compact forms when only using orthogonal basis functions[22,23,28,29]. Nodes' activities data are noisy and the mappings of network topologies are usually incomplete[30,31]. Collective behaviour, such as synchronization and consensus[19], can conceal the specific forms of microscopic mechanisms in interaction dynamics. To overcome these challenges, we propose here a two-phase inference approach. Our analysis indicates that the two-phase strategy allows us to achieve efficient and—most importantly—highly accurate inference, even in the face of unfavourable scenarios, such as noisy or low-resolution data or an only partially mapped topology (Fig. 1a).

## Results

**Overview of the two-phase inference approach.** Lacking a priori knowledge of the structures of $\mathbf{F}$ and $\mathbf{G}$, a natural approach is to pre-construct two extensive libraries $L_F$ and $L_G$ that contain a variety of elementary functions. The combinations of these elementary functions can potentially generate the true network dynamics. In this work, the libraries contain not only orthogonal basis functions but include polynomial, trigonometric, exponential, fractional, rescaling, sigmoid and other activation functions frequently used in various domains (Supplementary Tables 1 and 2). Large libraries are helpful for finding a compact and optimal model to capture network

[1]MOE Key Laboratory of Advanced Micro-Structured Materials and School of Physics Science and Engineering, Tongji University, Shanghai, People's Republic of China. [2]Frontiers Science Center for Intelligent Autonomous Systems, Tongji University, Shanghai, People's Republic of China. [3]Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Shanghai, People's Republic of China. ✉e-mail: gyan@tongji.edu.cn

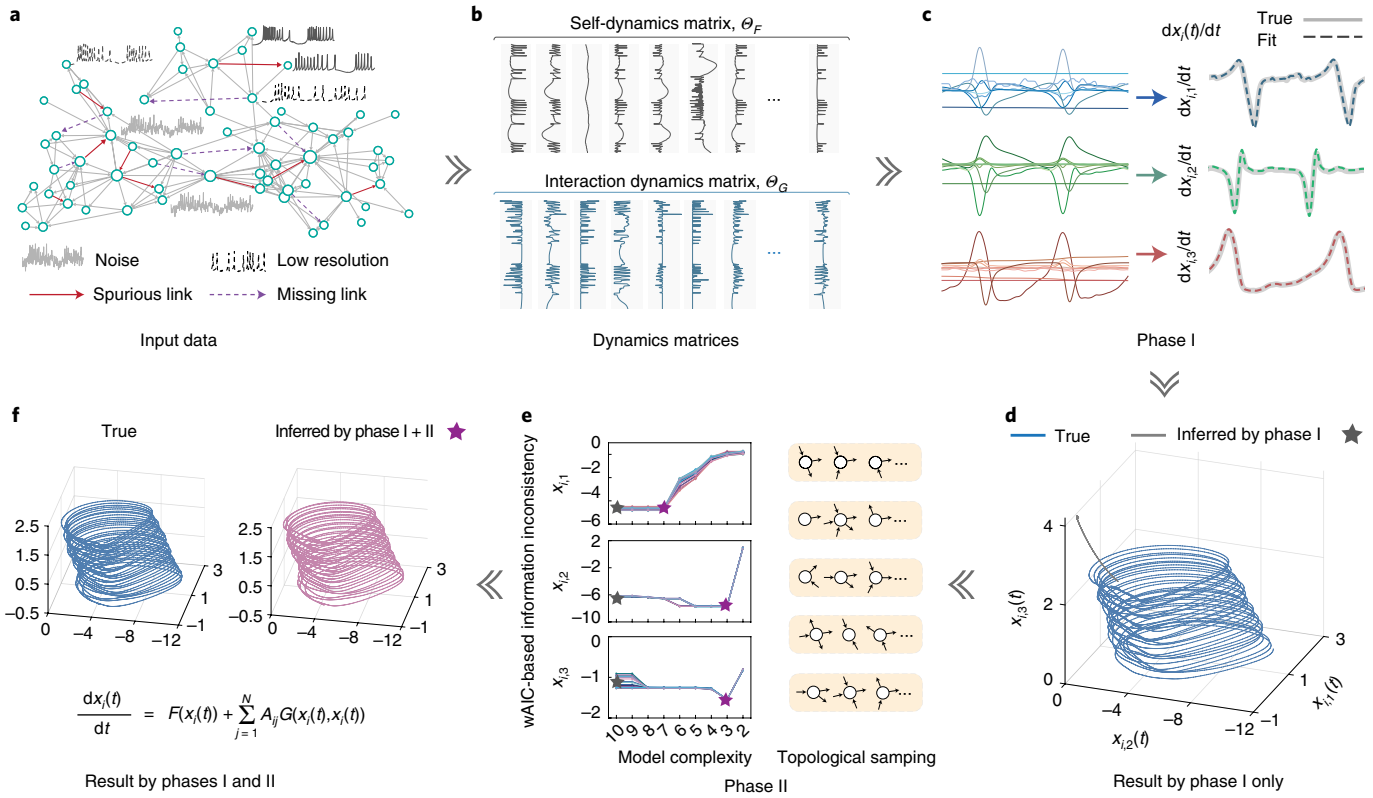


**Fig. 1 | Overview of the two-phase inference approach. a**, Observation data of network topology $A_{ij}$, including spurious and missing links, and low-resolution and noisy data of nodal activities $\mathbf{x}_i(t)$. **b**, Mapping the normalized observation data into two matrices $\Theta_F$ and $\Theta_G$ that represent the time-varying patterns of elementary functions. **c**, Phase I that narrows down the model space by identifying several leading elementary functions through global regression for each dimension of $\dot{\mathbf{x}}_i(t)$. **d**, Comparison of trajectories generated by the true network dynamics and the dynamical equation inferred by phase I alone. **e**, Phase II that performs local fine-tuning, by using topological sampling and wAIC, to further determine the optimal number (indicated by purple stars) of elementary functions for $\hat{\mathbf{F}}(\mathbf{x}_i(t))$ and $\hat{\mathbf{G}}(\mathbf{x}_i(t), \mathbf{x}_j(t))$. **f**, Comparison of trajectories generated by the true and inferred dynamical equations. The example illustrated in **c**–**f** is HR neuronal dynamics on a directed Barabási-Albert (BA) network with size $n=100$ and average degree $\langle k\rangle=5$.

dynamics but they also make the inference problem more difficult; due to the lack of orthogonality, the elementary functions can be similar with each other and thus less discriminative.

By introducing the time-series data $\mathbf{x}_i(t)$ (where $i=1,2,\ldots,n$) into $L_F$ and $L_G$, we obtain two time-varying matrices $\Theta_F(t)\equiv L_F(\mathbf{x}_i(t))$ and $\Theta_G(t)\equiv L_G(\mathbf{x}_i(t),\mathbf{x}_j(t))$ that encode the patterns of nodes' activities imposed by the elementary functions in $L_F$ and $L_G$ (Fig. 1b). Then, the inference problem can be recast to the selection of appropriate patterns in $\Theta_F(t)$ and $\Theta_G(t)$ that best match the evolution of observed system state $\dot{\mathbf{x}}(t)$, that is, to inferring the sparse coefficients $\boldsymbol{\xi}_F$ and $\boldsymbol{\xi}_G$ that best solve

$$\dot{\mathbf{x}}(t) = \widetilde{\Theta}_F(t)\boldsymbol{\xi}_F + \widetilde{A}\widetilde{\Theta}_G(t)\boldsymbol{\xi}_G, \tag{2}$$

where $\widetilde{A}\equiv A\otimes I_d$, $\widetilde{\Theta}_F\equiv\Theta_F\otimes I_d$ and $\widetilde{\Theta}_G\equiv\Theta_G\otimes I_d$, where the symbol $\otimes$ denotes the Kronecker product, and $I_d$ is the $d$-dimensional identity matrix. Here we consider the general setting where each node state is $d$ dimensional and the network is directed and heterogeneous. Consequently, the problem of inferring complex network dynamics is high dimensional and irreducible. Indeed, the number of elementary functions in $L_F$ and $L_G$ is approximately 25, 80 or 140 when the node activity itself has one, two or three dimensions, respectively, in the simulation validations below (Supplementary Tables 1 and 2).

Our approach is a two-phase procedure consisting of global regression and local fine-tuning. In phase I, we approximate the derivatives $\dot{\mathbf{x}}(t)$ (Methods) and calculate the matrices $\widetilde{\Theta}_F(t)$ and $\widetilde{\Theta}_G(t)$ and then normalize each of their columns (Fig. 1b). These normalized data are used to identify, through regression, the leading elementary functions that are most probably constituents of true $\mathbf{F}$ and $\mathbf{G}$ (Fig. 1c and Methods). Phase I is able to narrow down the model space, but the dynamical equation inferred by such regression alone lacks generative power (Fig. 1d). Next, in phase II, we perform fine-tuning with the original values of $\dot{\mathbf{x}}(t)$, $\widetilde{\Theta}_F(t)$ and $\widetilde{\Theta}_G(t)$, that is, without normalization. We use topological samplings (Methods) and the weighted Akaike's information criterion (wAIC; Methods) to sequentially remove the elementary functions with the smallest inferred coefficients (Fig. 1e). The final sets of elementary functions and their coefficients $\hat{\boldsymbol{\xi}}_F$ and $\hat{\boldsymbol{\xi}}_G$ compose $\hat{\mathbf{F}}$ and $\hat{\mathbf{G}}$, leading to the inferred dynamics of complex networks (Fig. 1f).

**Inferring complex network dynamics.** To validate the effectiveness of our approach, we apply it to infer five network dynamics, including the Hindmarsh–Rose[32] (HR, $d=3$) and FitzHugh–Nagumo[32] (FHN, $d=2$) neuronal systems, social balance dynamics[33] (SB, $d=1$), Kuramoto dynamics[34] ($d=1$) and coupled heterogeneous Rössler oscillators[35] ($d=3$); here $d$ is the dimension of each node activity. To obtain the nodes' activities data, we simulate these dynamics (Supplementary Table 4) on a variety of topologies, including Erdős–Rényi (ER) and scale-free (SF) synthetic networks and five empirical networks—cellular-level brain networks of *Caenorhabditis elegans* and *Drosophila*, Advogato social network, and power grids of Northern Europe and United States. The time series of node activities and each network topology are the input

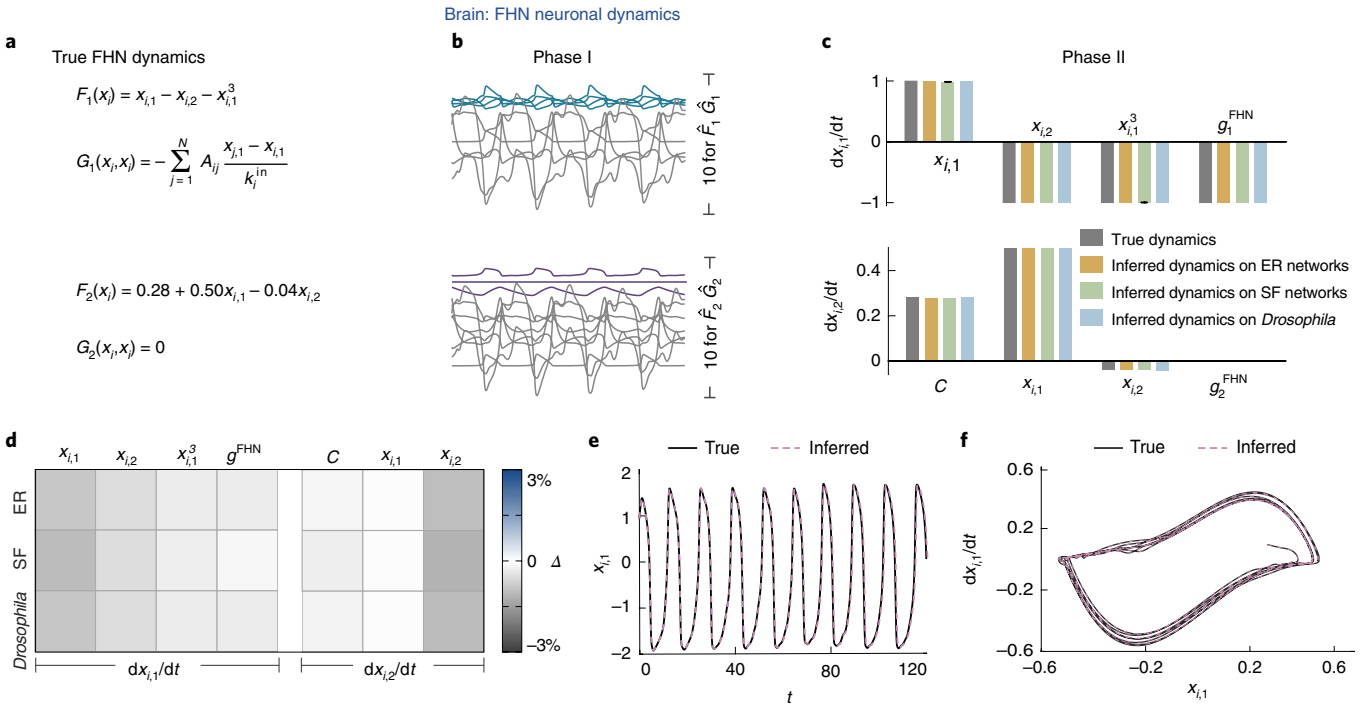


**Fig. 2 | Inferring FHN neuronal network dynamics on synthetic and real topologies. a**, True FHN dynamics used to simulate nodes' activities data on various topologies. $F_d$ and $G_d$ are self- and interaction dynamics of the $d$th dimension, respectively; $x_{i,d}$ is the $d$th dimension's state of node $i$, and $x_{i,d}^p$ is the polynomial with order $p$. **b**, Ten leading elementary functions identified by phase I for each dimension. **c**, Necessary elementary functions and their coefficients further inferred through phase II on two synthetic networks (directed ER and undirected SF) and one empirical network (*Drosophila* mushroom body), where $g^{FHN}$ denotes the term $(x_j - x_i)/k_i^{in}$. **d**, Relative errors $\Delta$ of the inferred elementary functions and their coefficients. Note that the elementary functions ruled out from $\Theta_F$ and $\Theta_G$ by our approach (whose coefficients are inferred as zero) are not shown. **e**,**f**, Nodes' activities (**e**) and trajectories (**f**) generated by the true and inferred equations.

data to our approach. The five specific equations governing these dynamics are the ground truths that we aim to infer. These dynamical models and networks are widely used in various domains and exhibit different properties (Supplementary Sections II and III), which accounted for the diversity of our tests.

Figure 2 illustrates the procedure of inferring FHN neuronal network dynamics. Through global regression, phase I identifies the ten most relevant elementary functions for each dimension of FHN (Fig. 2b); then, by local fine-tuning, phase II autonomously learns the compact and optimal form of the dynamical equation as well as the most appropriate coefficient for each of the necessary elementary functions (Fig. 2c). The form of the inferred equation in Fig. 2c perfectly matches the ground truth in Fig. 2a, and the learnt coefficients are also highly accurate. Indeed, the relative errors $\Delta = (\xi - \hat{\xi})/\xi$, where $\xi$ and $\hat{\xi}$ are the true and learnt coefficients, respectively, are smaller than 3% (Fig. 2d). The dynamical equation inferred by our approach exhibits generative power, being able to generate nodes' activities and trajectories that agree well with the observation data (Fig. 2e,f).

Our approach also successfully infers the equations governing the other four network dynamics. Regarding the accuracy of learnt coefficients, the relative errors $|\Delta|$ are less than 3% for the HR (Fig. 3a) and edge (Fig. 3c) dynamics on both synthetic and empirical networks. In Kuramoto dynamics and coupled heterogeneous Rössler oscillators, the self-dynamics are non-identical, that is, each node's dynamics has its own form (Supplementary Section III). Hence, we aim to infer an effective form of equation (1) that minimizes the inconsistency between the inferred and true nodes' activities. Even for these more challenging cases, the two-phase approach still succeeds with relative coefficient errors $|\Delta| < 5\%$ or $|\Delta| < 20\%$ (Fig. 3e,g). Both activities and trajectories generated by the effective equations exhibit high agreement with the true averaging dynamics (Fig. 3f,h,i).

**Inferrability of network dynamics.** Whether a network dynamics is inferrable depends on several factors. Here we explore three key factors, namely, synchronized dynamics, dynamical heterogeneity and deficient libraries.

Synchronized dynamics: if a network is completely synchronized, that is, all its nodes behave in the same manner[19,34,35], distinguishing the activities of a node and its neighbours becomes impossible, and the microscopic interacting mechanism $\mathbf{G}(\mathbf{x}_i, \mathbf{x}_j)$ between the nodes will be cloaked and undiscoverable. In other words, the more synchronized a network, more difficult it is to infer its dynamics. Here we tune the coupling strength between the nodes to change the degree of network synchronization (that is, order parameter $\langle R \rangle$; Supplementary Section IV), and test the capability of our two-phase approach in inferring partially synchronized network dynamics. As shown in Fig. 4a, although the inference inaccuracy increases when the system becomes more synchronized, our approach can still infer the true FHN equation even when the network is highly synchronized ($\langle R \rangle \approx 0.7$). The inference inaccuracy is quantified by a symmetric mean absolute percentage error (sMAPE; Methods). The more accurate the inference result, the closer the sMAPE value is to zero.

Dynamical heterogeneity: equation (1) assumes that nodes have the same form $\mathbf{F}$ of self-dynamics; yet this is not always true. For instance, although the self-dynamics of the Kuramoto model is simply one elementary function $\omega$ representing the natural frequency of a node, different nodes can have different values of $\omega$. For such non-identical self-dynamics, it is difficult—if not impossible—to infer a specific form $\mathbf{F}_i(\mathbf{x}_i)$ for each node $i$ due to an $n$-fold increase

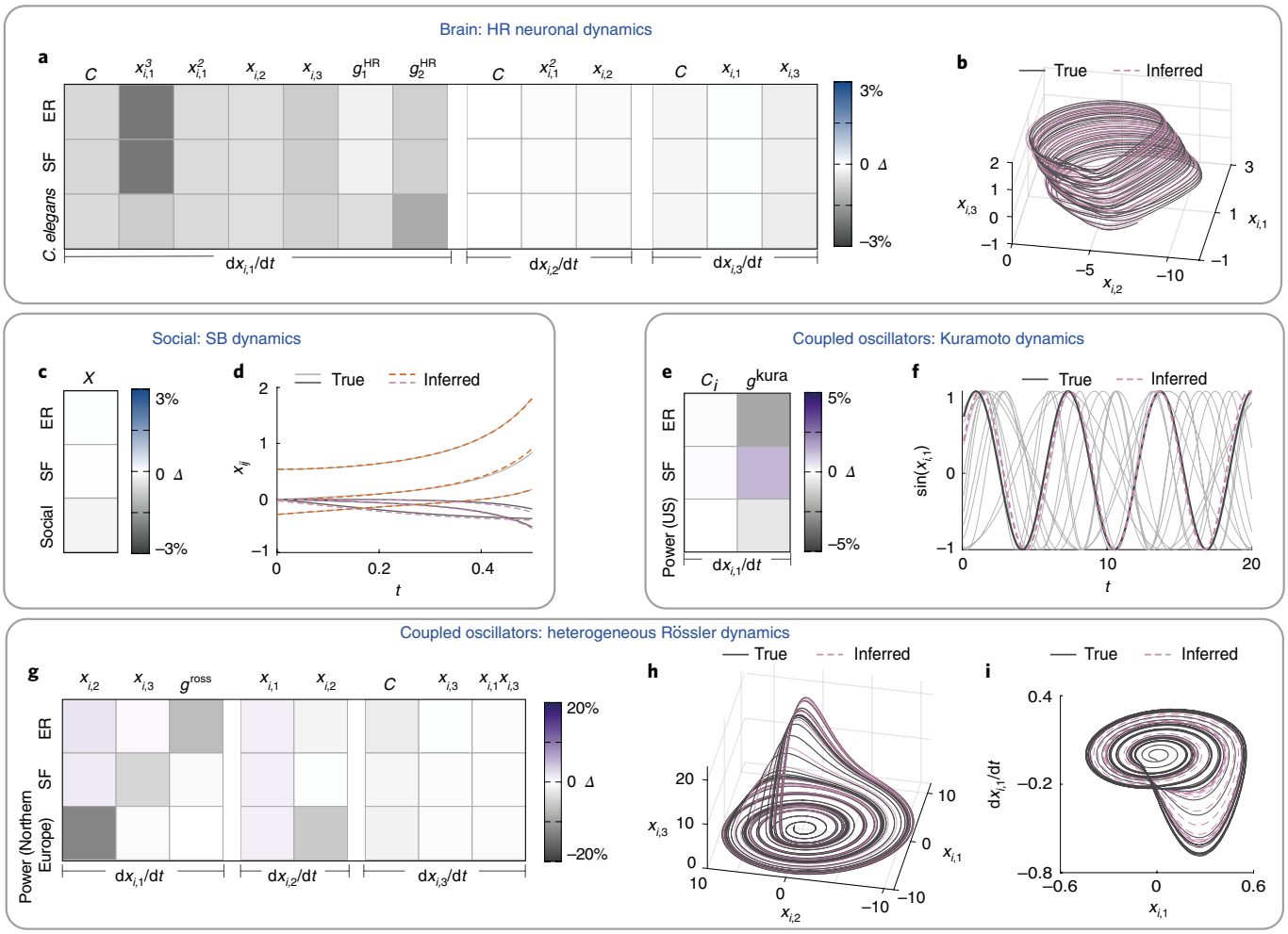


**Fig. 3 | Inference accuracy for other four typical nonlinear network dynamics. a**,**b**, Similar to Fig. 2 but for inferring HR neuronal dynamics, where the interaction dynamics $G(\mathbf{x}_i, \mathbf{x}_j)$ are composed of $g_1^{\mathrm{HR}} \equiv 1/(1+\mathrm{e}^{10(x_j-1)})$ and $g_2^{\mathrm{HR}} \equiv x_i/(1+\mathrm{e}^{10(x_j-1)})$. **c**,**d**, Relative errors (**c**) and six edges' activities (**d**) of the inferred edge dynamics of social balance. **e**–**i**, Relative errors of the inferred effective equations for network dynamics of the Kuramoto model and coupled Rössler oscillators. In both cases, the self-dynamics are heterogeneous, that is, the intrinsic frequency of each node is not identical but follows a normal distribution $\mathcal{N}(1, \sigma)$ with $\sigma = 0.1$. The grey curves represent the activity of individual nodes and the black curves represent the averaging activity of systems. Symbols $g^{\mathrm{kura}}$ and $g^{\mathrm{ross}}$ denote the terms $\sin(x_j - x_i)$ and $(x_j - x_i)$, respectively. The details of these dynamics and empirical networks are shown in Supplementary Tables 3 and 4.

in the dimensionality of potential model space ($n$ is the network size). Therefore, we aim to infer an effective equation that best captures the averaging dynamics (Fig. 3e,g). Here we further explore the extent of dynamical heterogeneity that our approach can tolerate. To do so, we assign each node a value of $\omega$ randomly drawn from a normal distribution $\mathcal{N}(0, \sigma)$ and increase the standard deviation $\sigma$. The inference inaccuracy indeed increases when $\sigma$ becomes larger, and the two-phase approach can tolerate dynamical heterogeneity $\sigma \lesssim 0.5$ (Fig. 4b).

Deficient libraries: although two rather comprehensive libraries of elementary functions are built, it is still possible that some elementary functions of the true unknown dynamics are missing. Another possibility is that the compact form of true dynamics cannot be composed by these elementary functions. For these cases, our two-phase approach will infer an alternative equation to capture the system behaviours. We test such capability in gene regulation and HR neuronal dynamics whose true coupling functions are intentionally removed from $L_G$. As shown in Fig. 4c, the trajectories generated by the inferred and true equations are close to each other, and the discrepancy is small for all the nodes (Methods and Supplementary Section IVB).

**Inferring from incomplete and noisy data.** The incompleteness of mapped network topology and noises of observed nodes' activities are inevitable in real data[30,31]. Hence, here we validate the robustness of our two-phase approach against low resolution, dynamical and observational noises, and spurious and missing links, as well as through comparisons with previous methods[23,36,37].

Low resolution: experimental and digital recording technologies often have limited measurement frequencies, inducing low resolution of the observed time series. To validate our approach's robustness against low resolution, we numerically simulate the five nonlinear network dynamics in Figs. 2 and 3 with a step size of 0.01, and then regularly downsample the activity data. We calculate the failure ratios in inferring the form of true equations (Supplementary Fig. 14a) as well as the inference inaccuracies (Fig. 5a). The results show that the two-phase strategy requires only a proportion of 5% to 50% data for the inference.

Observational and dynamical noises: observational noises are induced by the measuring process and dynamical noises represent the intrinsic stochasticity in dynamics. To produce the former, we add Gaussian noises to the nodes' activity data and quantify the intensity of observational noise with the signal-to-noise ratio

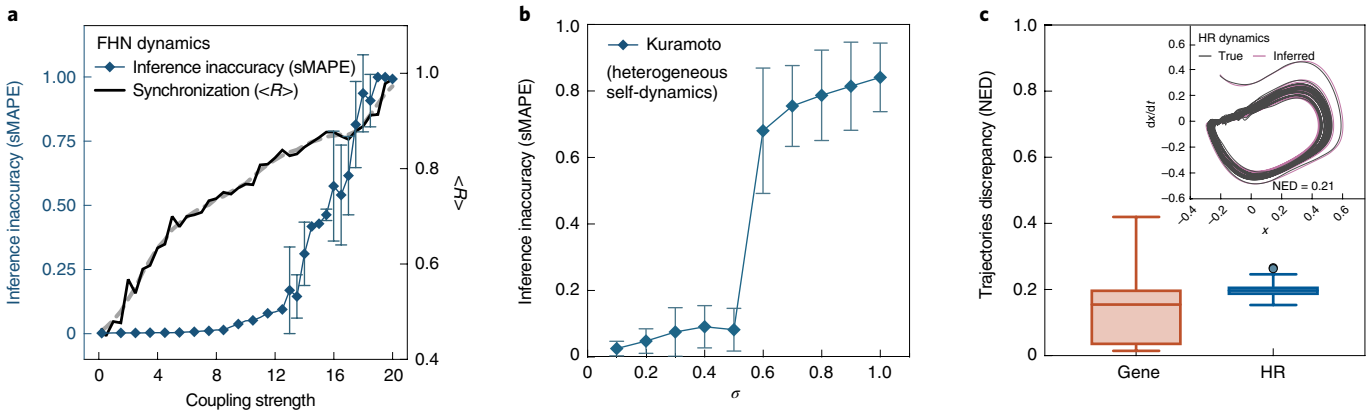


**Fig. 4 | Inferrability of network dynamics. a**, Inference inaccuracy represented by sMAPE and synchronization represented by order parameter $\langle R\rangle$ versus coupling strength between the nodes. **b**, Inaccuracy of inferred effective equation for Kuramoto network dynamics where the natural frequency $\omega$ of each node follows a normal distribution $\mathcal{N}(1,\sigma)$. Larger $\sigma$ indicates higher dynamical heterogeneity. **c**, NED (Methods) when some true elementary functions were deliberately removed from libraries $L_F$ and $L_G$. The box-whisker plots are visualized with the Tukey method (the box represents the interquartile range (IQR) and the line in the box indicates the median, with whiskers that extend 1.5 times the IQR from the box edges; the outliers are also shown) and the sample size is 100. The networks are SF with size $n=100$ and average degree $\langle k\rangle=5.0$. The simulation details are shown in Supplementary Table 4.

(Supplementary Section VA). To imitate the latter, we add a stochastic term of Gaussian white noise with intensity $\eta$ into the true dynamical equations and generate the nodes' activities data by the numerical simulations of these stochastic differential equations (Supplementary Section VA). We test the impact of these two types of noise on the performance of the two-phase inference approach, without any denoising pre-process. As shown in Fig. 5b and Supplementary Fig. 14b, the approach can tolerate dynamical noise with $\eta \leq 0.15$, meaning that it successfully reconstructs the hidden equations when the stochastic intensity is not higher than 15% of the average amplitude of true deterministic dynamics. Moreover, the approach can tolerate 30 dB observational noise (Fig. 5c and Supplementary Fig. 14c).

Spurious and missing links: spurious and missing links in real data induce an incomplete network topology $A_{ij}$, which further leads to an inaccurate interaction matrix $\Theta_G$. To test the impact of these erroneous links, we randomly add or remove a fraction of links from the true network topology that was used to simulate the nodes' activities. Owing to the topological sampling in phase II, our approach is able to tolerate 25% spurious and 30% missing links (Fig. 5d,e and Supplementary Fig. 14d,e).

Comparison with previous methods: the two most illuminating and effective methods for dynamics inference are Sparse Identification of Nonlinear Dynamics (SINDy)[23] and Algorithm for Revealing Network Interactions (ARNI)[37]. Note that ARNI originally aimed at inferring network topology but can be transferred to infer network dynamics by minor modification (Supplementary Section VC). Here we compare our approach with SINDy and ARNI from different aspects, including the amount of required data (Fig. 5f), robustness against observational noise (Fig. 5g), correlated dynamical noise (Fig. 5h and Supplementary Section VA), missing links (Fig. 5i) and different network sizes (Fig. 5j). Although ARNI needs fewer data points if network topologies are complete and nodal activities do not have any noise (Fig. 5f), the two-phase approach outperforms both SINDy and ARNI in inferring complex network dynamics from incomplete and noisy data (Fig. 5g–j). We also perform comparisons with SINDy's variant[36] regarding partially synchronized or heterogeneous dynamics (Supplementary Figs. 13 and 17). These results indicate that our approach can better handle high-dimensional networked systems and better cope with incompleteness and noises in data.

Ablation studies: besides the two-phase strategy, our approach also involves three important components, namely, normalization in the first phase yet non-normalization in the second for solving the issue raised by highly skewed observations at different nodes, topological sampling for imitating the feature of observed incomplete topologies and optimal selection by wAIC for determining the most appropriate complexity of inferred dynamics. The essentiality of the two-phase strategy and the three abovementioned components is demonstrated by ablation studies. Specifically, we ablate each phase or component and then assess the performance of the degenerated approaches. As shown in Fig. 5k,l and Supplementary Section VB, the inference inaccuracy (sMAPE) indeed increases if the phases or components are individually ablated.

**Inference of empirical systems.** To demonstrate the approach's ability of handling empirical systems, we apply it to infer the spreading dynamics of the infectious disease influenza A (H1N1). The network underlying this diffusion system is the worldwide airline network, which captures human mobility between different countries or regions and plays a dominant role for global disease spreading[5,6]. Each entry $A_{ij}$ of the weighted network's adjacency matrix $A$ represents the traffic volume from node $j$ to $i$, where each node denotes a country or region. The total passengers daily are approximately $\Phi=8.9\times10^6$; taking into account the population $P_i$ of each node $i$, the adjacency matrix is modified to

$$\hat{A}_{ij}=\frac{\Phi}{\sum_{i=1}^{n}P_i}A_{ij}. \tag{3}$$

The magnitude order of entries in matrix $\hat{A}$ is around $10^{-2}$ to $10^{-3}$. The nodal activities $x_i(t)$ are extracted from the daily reports of infected cases in each country or region. Here we consider the nodes whose accumulated H1N1 cases are more than 100 and focus on the early spreading dynamics, that is, within the 45 days since the first case was reported in each node: this captures the system behaviour before government control.

Based on these empirical data, our approach successfully infers a concise effective dynamical equation

$$\frac{\mathrm{d}x_i}{\mathrm{d}t}=ax_i+b\sum_{j=1}^{N}\hat{A}_{ij}\frac{1}{1+\mathrm{e}^{-(x_j-x_i)}}, \tag{4}$$

where $a=0.074$ and $b=7.130$ (Supplementary Section VI and Supplementary Fig. 18). It is interesting that our approach infers a

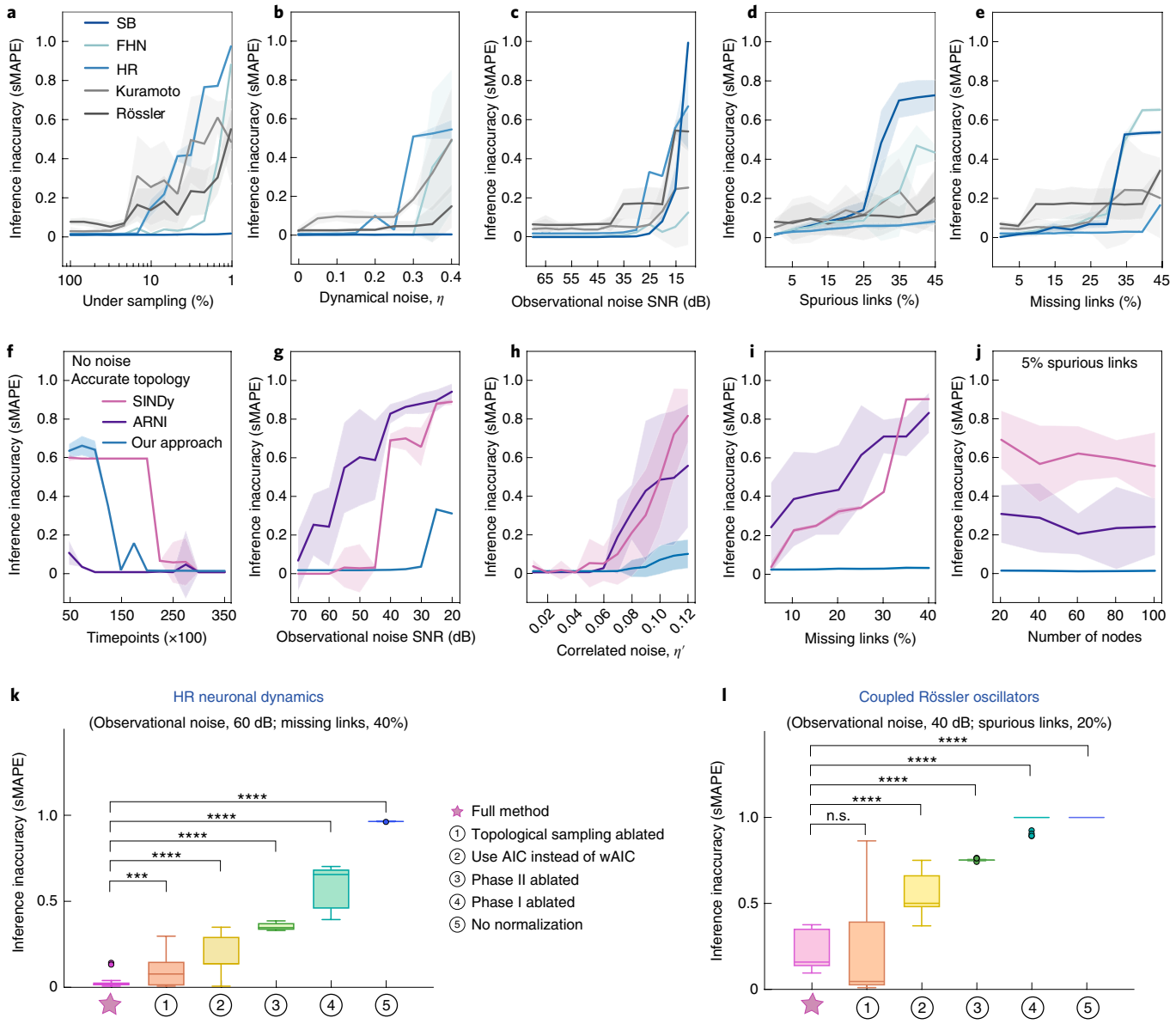


**Fig. 5 | Inference robustness against incompleteness and noises. a–e**, Inference inaccuracies (sMAPE) when the nodes' activities data are low resolution (**a**), have dynamical noises (Gaussian white noise with intensity $\eta$ (**b**)) or observational noises (intensity quantified by signal-to-noise ratio (**c**)), or when the topology data have spurious (**d**) and missing (**e**) links. **f–j**, Comparisons of inference inaccuracies between SINDy, ARNI and our approach for inferring HR neuronal network dynamics, with varying amounts of time points (**f**), observational noise (**g**), correlated dynamical noise (**h**), missing links (**i**) and different network sizes (**j**). Simulation details are shown in Supplementary Table 4. **k**,**l**, Comparison results of ablation studies: HR (**k**) and Rössler (**l**). The box-whisker plots are visualized with the Tukey method (the box represents the IQR and the line in the box indicates the median, with whiskers that extend 1.5 times the IQR from the box edges; the outliers are also shown) and the sample size is 20. Five ablation studies were performed: removing topological sampling (①), using original AIC instead of wAIC (②), removing phase II (③), removing phase I (④) or without normalization to $\Theta_F$ and $\Theta_G$ (⑤). Statistical significance is obtained through multiple Mann–Whitney tests. Three or four asterisks indicate a $p$ value of $<10^{-3}$ or $10^{-4}$, and n.s. means not significant.

sigmoid (nonlinear) form, rather than the linear form of epidemic models, to better capture the interaction dynamics. This might be caused by the fact that people usually consciously travel less if their countries/regions or the destinations have a higher infection risk. Although equation (4) describes the dynamics of all the nodes with the same parameters $a$ and $b$, we also extend it by taking into account dynamical heterogeneity in the nodes, that is, to obtain $a_i$ and $b_i$ from each node $i$'s activity data (Fig. 6b–e and Supplementary Fig. 18).

Because empirical systems lack ground truths, we verify the inferred equation (4) by testing its generalizability to the spread of severe acute respiratory syndrome (SARS) and coronavirus disease 2019 (COVID-19). Based on the daily reported numbers within the first 45 days in each node, we find that equation (4) is also able to capture the early spread of SARS and COVID-19 on the worldwide airline network. Indeed, as shown in Fig. 6f–i and Supplementary Figs. 19 and 20, evolution of the cumulative numbers of SARS cases (for nodes whose eventual infected cases are more than 100) and COVID-19 cases (for nodes whose eventual infected cases are more than 2,000) agree well with the activities generated by equation (4) with heterogeneous parameters $a_i$ and $b_i$.

## Discussion

Many real networks have been mapped so far, but there are still complex systems whose network structure information is totally missing.

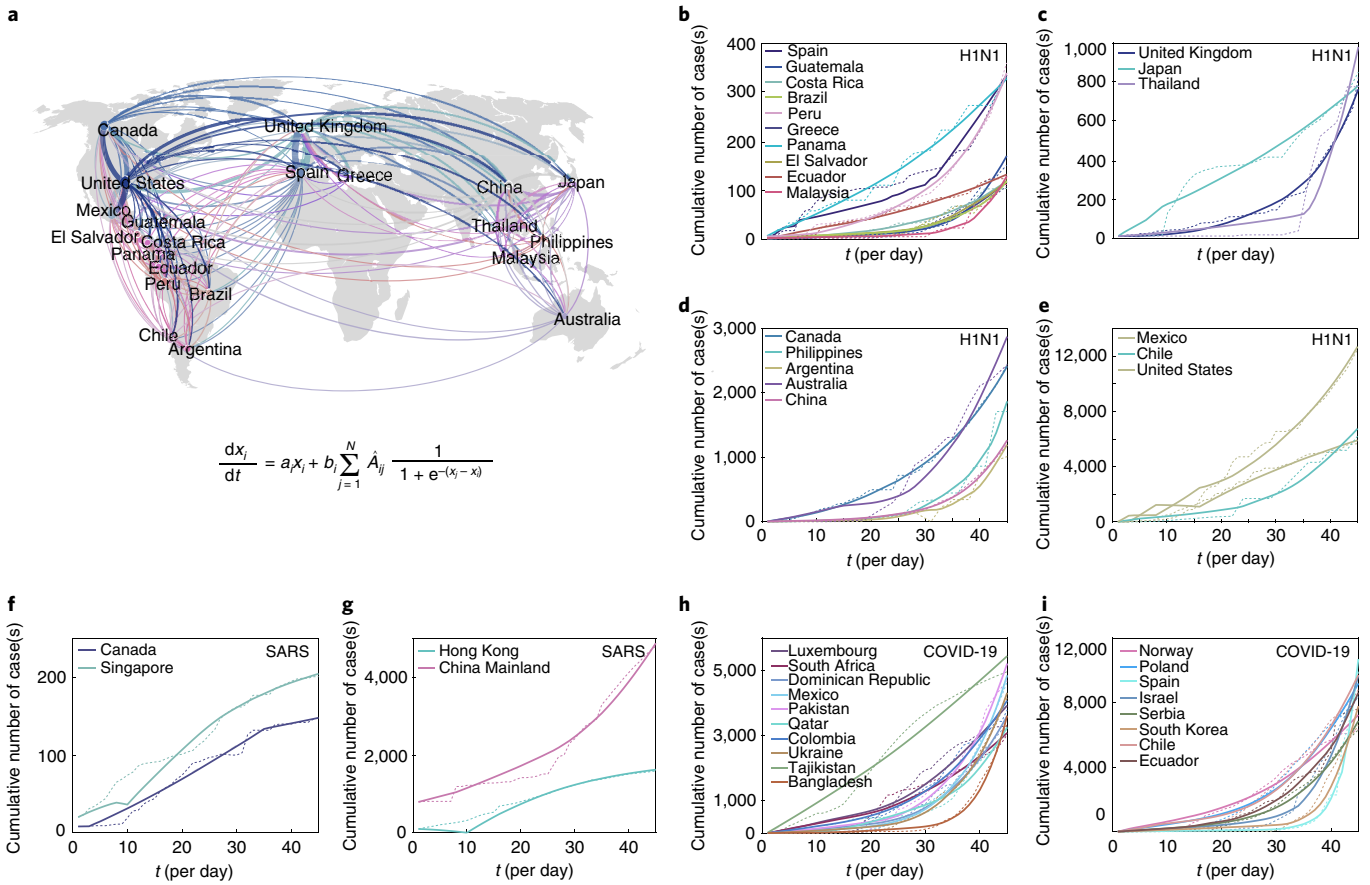


**Fig. 6 | Inference of early spreading dynamics from empirical data. a**, Worldwide airline network (partial) used for the inference. Each node represents a country or region, and the line thickness represents the amount of passenger flow. The form describes the dynamical equation inferred by the two-phase approach. **b**–**e**, Comparisons between the empirical cumulative number of H1N1 cases for different nodes (dashed lines) and the cumulative number generated by the inferred equation (solid lines). For better visualization, the comparisons are displayed in four plots (from **b** to **e**) and all the dates (when the first case is reported for each node) are shifted to the first day ($t=1$). **f**–**i**, Comparisons between the empirical and inferred cumulative numbers in different nodes for SARS (**f** and **g**) and COVID-19 (**h** and **i**).

For the latter, a possible scheme is inferring their topological structure, especially directed or causal networks[30,37–41], from nodes' activities data first and then applying our approach to infer the system dynamics. It is worth noting that inferring the network structure from nodes' activities data is also challenging, especially when the number of nodes is large[42,43], because the number of parameters needing to be estimated is about $n^2$ (where $n$ is the network size). Therefore, how to simultaneously infer both structure and dynamics of large, complex systems is still an outstanding problem.

Our work also raises several questions worthy of future pursuit. First, stochasticity in the dynamics of some real complex systems might be stronger than that considered in this work. Such highly stochastic systems are better described by stochastic differential equations[29,44–46]. Second, our approach does not account for discrete or Boolean dynamics, or systems that contain thresholding terms or exhibit irregular dynamics with instability properties[47]. Third, when the nodal activity is multidimensional, experimental access might be limited to a sub-dimension of the activity vector. The Koopman operator and time-delay embedding techniques are helpful for capturing the dynamical properties of sub-dimension observable systems[48]. Yet, the problem remains unsolved for complex networked systems. Finally, the nodes in a complex system can have higher-order—beyond pairwise—couplings, and such higher-order interactions may impact the dynamics of networked systems[49,50]. Hence, it is an interesting direction to extend the approach to inferring higher-order network dynamics.

## Methods

**Two-phase inference approach.** The left-hand side of equation (1) represents the time-varying derivative of each node's activity, which can be numerically obtained from $\mathbf{x}_i(t)$ through the five-point approximation[51]

$$\dot{x}_t \approx \frac{x_{t-2\delta t} - 8x_{t-\delta t} + 8x_{t+\delta t} - x_{t+2\delta t}}{12\delta t}, \quad (5)$$

where $\delta t$ is the time step. Hence, the specific goal is to infer both exact structure and corresponding coefficients of the self-dynamics function $\mathbf{F}(\mathbf{x}_i(t))$ and the interaction dynamics function $\mathbf{G}(\mathbf{x}_i(t), \mathbf{x}_j(t))$.

Because we lack a priori knowledge of the forms of $\mathbf{F}$ and $\mathbf{G}$, we construct two comprehensive libraries, namely, $L_F$ and $L_G$, for self- and interaction dynamics, respectively, including polynomial, trigonometric, exponential, fractional, rescaling and various activation functions (Supplementary Tables 1 and 2). By introducing the observed time series of nodes' activities to the elementary functions in $L_F$ and $L_G$, we obtain two matrices $\Theta_F(t) = L_F(\mathbf{x}_i(t))$ and $\Theta_G(t) = L_G(\mathbf{x}_i(t), \mathbf{x}_j(t))$ that describe the corresponding behaviours of these elementary functions (Supplementary Fig. 1). To infer the compact forms that best match equation (2), we propose a two-phase approach.

Phase I, global regression: the purpose of this phase is to assess the relevance of each elementary function in $L_F$ and $L_G$ to the true, yet unknown, network dynamics. Given the observations of $\mathbf{x}_i(t)$ for all $i$ at time $t$, we approximate the derivatives $\dot{\mathbf{x}}(t)$ and calculate the matrices $\widetilde{\Theta}_F(t)$ and $\widetilde{\Theta}_G(t)$. These values are highly skewed and can span several orders of magnitude (Supplementary Fig. 3) due to the skewness of node degrees and nonlinearity of system dynamics, which could induce an overestimation of the importance for inherently low-value constituents. To eliminate this severe effect, it is crucial to normalize each column in $\dot{\mathbf{x}}(t)$, $\widetilde{\Theta}_F(t)$ and $\widetilde{\Theta}_G(t)$. Then, the inference problem described by equation (2) is further recast to an optimization formula:

$$\arg\min_{\xi_F,\xi_G}\int_0^T\left(\|\widetilde{\Theta}_F(t)\xi_F+\widetilde{A}\widetilde{\Theta}_G(t)\xi_G-\dot{\mathbf{x}}(t)\|^2\right)\mathrm{d}t+\lambda(\|\xi_F\|+\|\xi_G\|),\quad(6)$$

where $\lambda>0$ is a hyper-parameter that regulates the sparsity of coefficient vectors $\xi_F$ and $\xi_G$. We employ the regression analysis method of the least absolute shrinkage and selection operator to solve equation (6) and perform a fivefold validation to obtain the most appropriate value of $\lambda$ (Supplementary Section IB). The resultant $\xi_F$ and $\xi_G$ capture the relevance of each elementary function in $L_F$ and $L_G$, enabling the identification of leading elementary functions that are most probably the constituents of the true **F** and **G** (Fig. 1c). Consequently, phase I is able to narrow down the model space. However, the dynamical equation inferred by such regression alone lacks generative power. For instance, as shown in Fig. 1d, the trajectory generated by an inferred dynamical equation of phase I deviates from that of the true network dynamics.

Phase II, local fine-tuning: to reconstruct generative and concise expressions for **F** and **G**, we next perform fine-tuning in the reduced model space (Supplementary Section IB). In contrast to phase I, we now use the original values of $\dot{\mathbf{x}}(t)$, $\widetilde{\Theta}_F(t)$ and $\widetilde{\Theta}_G(t)$, that is, without normalization, to further identify the necessary elementary functions and learn their precise coefficients. Since spurious or missing links in the observed network topology have an adverse effect on learning, we perform topological sampling (discussed later) that imitates the feature of observed—usually incomplete—topologies. Another issue is to determine the minimal number of elementary functions required for reconstructing **F** and **G**. To do so, we sequentially remove the elementary functions with the smallest inferred coefficients and calculate, using a weighted version of Akaike's information criterion (wAIC; discussed later), the information inconsistency between the observed nodes' activities and the remaining set of elementary functions. This process stops when removing a certain elementary function consistently increases the value of wAIC. As shown in Fig. 1e, each curve in a plot at the left column represents the information inconsistency versus model complexity for one topological sample. We find that, indeed, the joint operation with wAIC and topological sampling is helpful for inference from noisy and incomplete data (Fig. 5k,l).

The final sets of elementary functions and their coefficients $\hat{\xi}_F$ and $\hat{\xi}_G$ compose the forms $\hat{\mathbf{F}}$ and $\hat{\mathbf{G}}$, leading to the successful inference of network dynamics described by equation (1). Indeed, as demonstrated in Fig. 1f, the trajectory generated by the inferred dynamical equation agrees well with the numerical simulations of the true network dynamics. It is worth noting that the ground truth, that is, the form of the true equation, remains unknown during the whole procedure and is only used to assess the accuracy of the final inferred results; hence, our approach works in an autonomous, unsupervised way.

**wAIC.** The original Akaike's information criterion (AIC)[52] is a frequently used method to balance the fitting and complexity of a model with respect to the observed data, defined as AIC$=n\log\mathrm{MSE}+2p$, where $n$ is the number of observations, MSE is the mean squared error of the regression result of the model and $p$ is the number of variables. By using AIC, one aims to select an optimal model that best fits the observations with the fewest variables from the model candidates. However, we find that the original AIC does not work well in the inference problem we aim to solve in the present work. Hence, we introduce a weighted version of AIC (namely, wAIC) as

$$\mathrm{wAIC}=\begin{cases}w(n\log\mathrm{MSE}+2p), & (n\log\mathrm{MSE}+2p)\geq 0,\\ (n\log\mathrm{MSE}+2p)/w, & (n\log\mathrm{MSE}+2p)<0,\end{cases}\quad(7)$$

which balances the fitting accuracy and model complexity. Here $w$ is the inferred coefficient of a term from phase I. A term with a larger $w$ inferred by phase I is more likely to be able to capture the properties of the underlying unknown dynamics. Thus, multiplying $w$ or $1/w$ with AIC amplifies the impact of removing this term from the equation. The smaller the wAIC, the more consistent is the composition of the elementary functions with the observed data and less important is this removed term.

To be specific, to evaluate the relevance of term $i$, we remove this term from the equation inferred by phase I and calculate the value of $\mathrm{wAIC}_i$ of the new, shorter equation (Supplementary Fig. 2). We repeat this process to obtain the wAIC for each term. Then, we sort these terms based on their wAIC values, and remove terms one by one with wAIC values from small to large. This operation gives a shorter equation at each step, and we calculate the AIC values of these shortened equations. The optimal equation is determined at the turning point where the curve starts to consistently increase (Fig. 1e, purple stars).

**Topological sampling.** We perform topological sampling in phase II as follows. We randomly choose $S$ nodes from all $n$ nodes, and obtain the activities of these $S$ nodes' partial neighbours. Introducing the sampled ego structures and nodes' activities into libraries $L_F$ and $L_G$ allows us to construct the self- and interaction matrices $\widetilde{\Theta}_F$ and $\widetilde{\Theta}_G$, respectively, as well as to further distil the elementary functions and their coefficients. We repeat the process to obtain $K$ sets of samples and average the coefficients of the elementary functions inferred from the $K$ sample sets. In the present work, we set $S=10$ and $K=20$.

**sMAPE.** The inference inaccuracy is quantified by sMAPE[53]:

$$\mathrm{sMAPE}=\frac{1}{m}\sum_{i=1}^{m}\frac{|I_i-R_i|}{(|I_i|+|R_i|)},\quad(8)$$

where $m$ is the cardinal number of the set that contains both inferred and true elementary functions and $I_i$ and $R_i$ are the inferred and true coefficients, respectively. The range of sMAPE is $[0,1]$. The more accurate the inferred equation, the lower is the value of sMAPE. Note that if an inferred elementary function should not exist in the true equation or a true elementary function is not successfully inferred, the value of sMAPE increases. Therefore, sMAPE captures not only the errors of inferred coefficients but also the incorrectness of the inferred equation form.

**NED.** To evaluate the discrepancy between the inferred and true dynamics, we used the metric of normalized Euclidean distance (NED) that represents the distance between the two trajectories generated by the inferred and true dynamical equations. That is,

$$\mathrm{NED}\,(x_i,\hat{x}_i)=\frac{1}{D_{\max}(T-t_0)}\sum_{t=t_0}^{T}\sqrt{(x_i(t)-\hat{x}_i(t))^2+(\dot{x}_i(t)-\dot{\hat{x}}_i(t))^2}.\quad(9)$$

Here $x_i$ is the true trajectory and $\hat{x}_i$ is the trajectory generated by the inferred equation; $t_0$ and $T$ are the beginning and ending times, respectively; and $D_{\max}$ is the longest Euclidean distance between a pair of points of the true trajectory.

## Data availability

Source data are provided with this paper. The empirical network data include *C. elegans* connectome[54–56], the mushroom-body region of *Drosophila*[57], Northern Europe power grid[58], the US power grid[59], Advogato social network[60] retrieved from https://networkrepository.com/ and worldwide airline network data retrieved from OpenFlights (https://openflights.org/data.html). The empirical data of epidemic spreading include daily reported numbers of H1N1 and SARS cases available at Kaggle (https://www.kaggle.com/lnunes/a-brief-comparative-study-of-epidemics/data) and the daily reported numbers of COVID-19 cases[61].

## Code availability

All the source codes are publicly available at the Code Ocean capsule[62].

Received: 22 June 2021; Accepted: 17 February 2022;
Published online: 24 March 2022

## References

1. Grewe, B. F., Langer, D., Kasper, H., Kampa, B. M. & Helmchen, F. High-speed in vivo calcium imaging reveals neuronal network activity with near-millisecond precision. *Nat. Methods* **7**, 399–405 (2010).
2. Stetter, O., Battaglia, D., Soriano, J. & Geisel, T. Model-free reconstruction of excitatory neuronal connectivity from calcium imaging signals. *PLoS Comput. Biol.* **8**, e1002653 (2012).
3. Reuter, J. A., Spacek, D. V. & Snyder, M. P. High-throughput sequencing technologies. *Mol. Cell.* **58**, 586–597 (2015).
4. Levy, S. E. & Myers, R. M. Advancements in next-generation sequencing. *Annu. Rev. Genom. Hum. Genet.* **17**, 95–115 (2016).
5. 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).
6. Brockmann, D. & Helbing, D. The hidden geometry of complex, network-driven contagion phenomena. *Science* **342**, 1337–1342 (2013).
7. Chang, S. et al. Mobility network models of COVID-19 explain inequities and inform reopening. *Nature* **589**, 82–87 (2021).
8. Newman, M., Barabási, A.-L. & Watts, D. J. *The Structure and Dynamics of Networks* (Princeton Univ. Press, 2011).
9. Barzel, B. & Barabási, A.-L. Universality in network dynamics. *Nat. Phys.* **9**, 673–681 (2013).
10. Harush, U. & Barzel, B. Dynamic patterns of information flow in complex networks. *Nat. Commun.* **8**, 2181 (2017).
11. Stankovski, T., Pereira, T., McClintock, P. V. & Stefanovska, A. Coupling functions: universal insights into dynamical interaction mechanisms. *Rev. Mod. Phys.* **89**, 045001 (2017).
12. Breakspear, M. Dynamic models of large-scale brain activity. *Nat. Neurosci.* **20**, 340–352 (2017).
13. Santolini, M. & Barabási, A.-L. Predicting perturbation patterns from the topology of biological networks. *Proc. Natl Acad. Sci. USA* **115**, E6375–E6383 (2018).
14. Buldyrev, S. V., Parshani, R., Paul, G., Stanley, H. E. & Havlin, S. Catastrophic cascade of failures in interdependent networks. *Nature* **464**, 1025–1028 (2010).

15. Yang, Y., Nishikawa, T. & Motter, A. E. Small vulnerable sets determine large network cascades in power grids. *Science* **358**, eaan3184 (2017).
16. Pastor-Satorras, R., Castellano, C., Van Mieghem, P. & Vespignani, A. Epidemic processes in complex networks. *Rev. Mod. Phys.* **87**, 925 (2015).
17. Castellano, C., Fortunato, S. & Loreto, V. Statistical physics of social dynamics. *Rev. Mod. Phys.* **81**, 591–646 (2009).
18. Becker, J., Brackbill, D. & Centola, D. Network dynamics of social influence in the wisdom of crowds. *Proc. Natl Acad. Sci. USA* **114**, E5070–E5076 (2017).
19. Arenas, A., Díaz-Guilera, A., Kurths, J., Moreno, Y. & Zhou, C. Synchronization in complex networks. *Phys. Rep.* **469**, 93–153 (2008).
20. Barzel, B., Liu, Y.-Y. & Barabási, A.-L. Constructing minimal models for complex system dynamics. *Nat. Commun.* **6**, 7186 (2015).
21. Schmidt, M. & Lipson, H. Distilling free-form natural laws from experimental data. *Science* **324**, 81–85 (2009).
22. 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).
23. Brunton, S. L., Proctor, J. L. & Kutz, J. N. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proc. Natl Acad. Sci. USA* **113**, 3932–3937 (2016).
24. Rudy, S. H., Brunton, S. L., Proctor, J. L. & Kutz, J. N. Data-driven discovery of partial differential equations. *Sci. Adv.* **3**, e1602614 (2017).
25. Udrescu, S.-M. & Tegmark, M. AI Feynman: a physics-inspired method for symbolic regression. *Sci. Adv.* **6**, eaay2631 (2020).
26. Raissi, M. & Karniadakis, G. E. Hidden physics models: machine learning of nonlinear partial differential equations. *J. Comput. Phys.* **357**, 125–141 (2018).
27. Iten, R., Metger, T., Wilming, H., Del Rio, L. & Renner, R. Discovering physical concepts with neural networks. *Phys. Rev. Lett.* **124**, 010508 (2020).
28. Frishman, A. & Ronceray, P. Learning force fields from stochastic trajectories. *Phys. Rev. X* **10**, 021009 (2020).
29. Brückner, D. B., Ronceray, P. & Broedersz, C. P. Inferring the dynamics of underdamped stochastic systems. *Phys. Rev. Lett.* **125**, 058103 (2020).
30. Shandilya, S. G. & Timme, M. Inferring network topology from complex dynamics. *New J. Phys.* **13**, 013004 (2011).
31. Newman, M. E. J. Network structure from rich but noisy data. *Nat. Phys.* **14**, 542–545 (2018).
32. Rabinovich, M. I., Varona, P., Selverston, A. I. & Abarbanel, H. D. Dynamical principles in neuroscience. *Rev. Mod. Phys.* **78**, 1213 (2006).
33. Marvel, S. A., Kleinberg, J., Kleinberg, R. D. & Strogatz, S. H. Continuous-time model of structural balance. *Proc. Natl Acad. Sci. USA* **108**, 1771–1776 (2011).
34. Strogatz, S. H. Exploring complex networks. *Nature* **410**, 268–276 (2001).
35. Barahona, M. & Pecora, L. M. Synchronization in small-world systems. *Phys. Rev. Lett.* **89**, 054101 (2002).
36. Mangan, N. M., Kutz, J. N., Brunton, S. L. & Proctor, J. L. Model selection for dynamical systems via sparse regression and information criteria. *Proc. Math. Phys. Eng. Sci.* **473**, 20170009 (2017).
37. Casadiego, J., Nitzan, M., Hallerberg, S. & Timme, M. Model-free inference of direct network interactions from nonlinear collective dynamics. *Nat. Commun.* **8**, 2192 (2017).
38. Runge, J., Nowack, P., Kretschmer, M., Flaxman, S. & Sejdinovic, D. Detecting and quantifying causal associations in large nonlinear time series datasets. *Sci. Adv.* **5**, eaau4996 (2019).
39. Sugihara, G. et al. Detecting causality in complex ecosystems. *Science* **338**, 496–500 (2012).
40. Sun, J., Taylor, D. & Bollt, E. M. Causal network inference by optimal causation entropy. *SIAM J. Appl. Dyn. Syst.* **14**, 73–106 (2015).
41. Kralemann, B., Pikovsky, A. & Rosenblum, M. Reconstructing effective phase connectivity of oscillator networks from observations. *New J. Phys.* **16**, 085013 (2014).
42. Frässle, S. et al. Regression DCM for fMRI. *NeuroImage* **155**, 406–421 (2017).
43. Gilson, M., Moreno-Bote, R., Ponce-Alvarez, A., Ritter, P. & Deco, G. Estimation of directed effective connectivity from fMRI functional connectivity hints at asymmetries of cortical connectome. *PLoS Comput. Biol.* **12**, e1004762 (2016).
44. Deco, G., Rolls, E. T. & Romo, R. Stochastic dynamics as a principle of brain function. *Prog. Neurobiol.* **88**, 1–16 (2009).
45. Genkin, M., Hughes, O. & Engel, T. A. Learning non-stationary Langevin dynamics from stochastic observations of latent trajectories. *Nat. Commun.* **12**, 5986 (2021).
46. Zhao, H. Inferring the dynamics of 'black-box' systems using a learning machine. *Sci. China Phys. Mech. Astron.* **64**, 270511 (2021).
47. Jahnke, S., Memmesheimer, R.-M. & Timme, M. Stable irregular dynamics in complex neural networks. *Phys. Rev. Lett.* **100**, 048102 (2008).
48. Champion, K. P., Brunton, S. L. & Kutz, J. N. Discovery of nonlinear multiscale systems: sampling strategies and embeddings. *SIAM J. Appl. Dyn. Syst.* **18**, 312–333 (2019).
49. Battiston, F. et al. The physics of higher-order interactions in complex systems. *Nat. Phys.* **17**, 1093–1098 (2021).
50. Lambiotte, R., Rosvall, M. & Scholtes, I. From networks to optimal higher-order models of complex systems. *Nat. Phys.* **15**, 313–320 (2019).
51. Sauer, T. Numerical solution of stochastic differential equations in finance. in *Handbook of Computational Finance* 529–550 (Springer, 2012).
52. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Automat. Contr.* **19**, 716–723 (1974).
53. Flores, B. E. A pragmatic view of accuracy measurement in forecasting. *Omega* **14**, 93–98 (1986).
54. White, J. G., Southgate, E., Thomson, J. N. & Brenner, S. The structure of the nervous system of the nematode *Caenorhabditis elegans*. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **314**, 1–340 (1986).
55. Varshney, L. R., Chen, B. L., Paniagua, E., Hall, D. H. & Chklovskii, D. B. Structural properties of the *Caenorhabditis elegans* neuronal network. *PLoS Comput. Biol.* **7**, e1001066 (2011).
56. Yan, G. et al. Network control principles predict neuron function in the *Caenorhabditis elegans* connectome. *Nature* **550**, 519–523 (2017).
57. Scheffer, L. K. et al. A connectome and analysis of the adult *Drosophila* central brain. *eLife* **9**, e57443 (2020).
58. Menck, P. J., Heitzig, J., Kurths, J. & Schellnhuber, H. J. How dead ends undermine power grid stability. *Nat. Commun.* **5**, 3969 (2014).
59. Kunegis, J. KONECT: the Koblenz network collection. In *Proc. 22nd International Conference on World Wide Web* 1343–1350 (ACM, 2013).
60. Rossi, R. & Ahmed, N. The network data repository with interactive graph analytics and visualization. In *Twenty-Ninth AAAI Conference on Artificial Intelligence* (2015).
61. Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect. Dis.* **20**, 533–534 (2020).
62. Gao, T.-T. & Yan, G. A two-phase approach for inferring complex network dynamics. *Code Ocean* https://doi.org/10.24433/CO.4774495.v1 (2022).


## Acknowledgements

T.-T.G. and G.Y. are supported by the National Key Research and Development Program of China (grant no. 2021ZD0204500), National Natural Science Foundation of China (grant nos. 12161141016 and 11875043), Shanghai Municipal Science and Technology Major Project (grant no. 2021SHZDZX0100), Shanghai Municipal Commission of Science and Technology Project (grant nos. 18ZR1442000 and 19511132101) and Fundamental Research Funds for the Central Universities. We are also grateful for the helpful discussion with B. Barzel, J. Moore, X. Ru and T. Li.

## Author contributions

G.Y. conceived the research. G.Y. and T.-T.G. designed the research. T.-T.G. performed the research. T.-T.G. and G.Y. analysed the results. G.Y. and T.-T.G. wrote the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s43588-022-00217-0.

**Correspondence and requests for materials** should be addressed to Gang Yan.

**Peer review information** *Nature Computational Science* thanks Matthieu Gilson and the other, anonymous reviewer(s) for their contribution to the peer review of this work. Handling editor: Jie Pan, in collaboration with the *Nature Computational Science* team.

**Reprints and permissions information** is available at www.nature.com/reprints.

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