Supplementary Materials for A universal data based method for reconstructing complex networks with binary-state dynamics

Jingwen Li, Zhesi Shen, Wen-Xu Wang, Celso Grebogi, and Ying-Cheng Lai

1 Computation details

Parameter values in the binary-state dynamics used for network reconstruction are displayed in Supplementary Table S1. The only requirement for choosing the parameter values is that the switching dynamics should be monotonic. Since all the binary-state dynamics are monotonic, there is no specific restriction for the parameter values. Note that several models have convergent behaviors. If the states of nodes converge into a stable state, there will be no more useful information for network reconstruction. If this occurs, we randomly initialize the states of all nodes after a certain period.

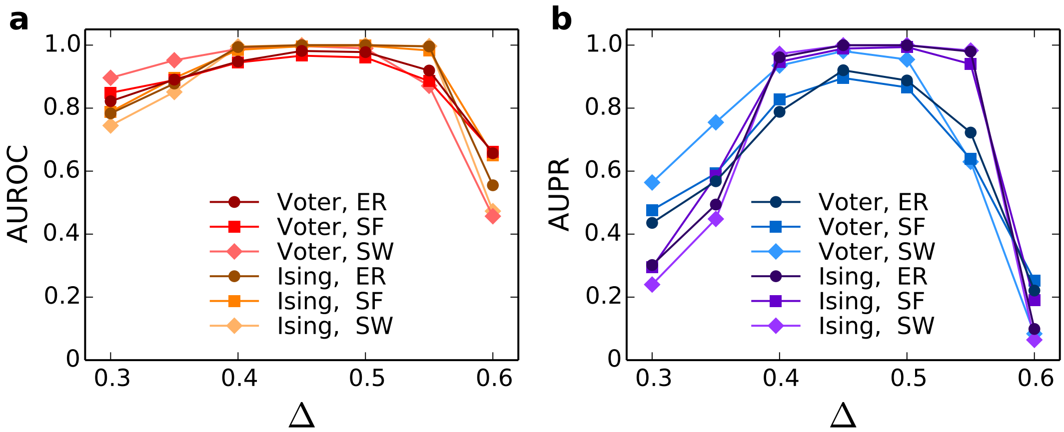
The set of the threshold parameter Δ for realizing the merging process for network reconstruction is independent of network structure and binary-state dynamics. We investigate the dependence of the reconstruction performance on threshold Δ. The results are shown in Supplementary Fig S1. We found that AUROC and AUPR can always reach high values when $0.4 \leqslant \Delta \leqslant 0.55$ in all cases. Thus, we set the threshold Δ to be 0.45 for simplicity.

Regarding the selection of bases, the method is relatively time consuming because it requires calculating the Hamming distance between each pair of strings in different time steps. Hence, to improve computational efficiency, for large-size networks with $N \geqslant 500$, we choose bases randomly instead of using the base-selection method presented in the main text. It reduces accuracy a little in a few cases, but the computational complexity is considerably reduced. Supplementary Figs. S2 (**a**) and (**b**) show the results of reconstruction for Ising and Voter dynamics on ER, WS

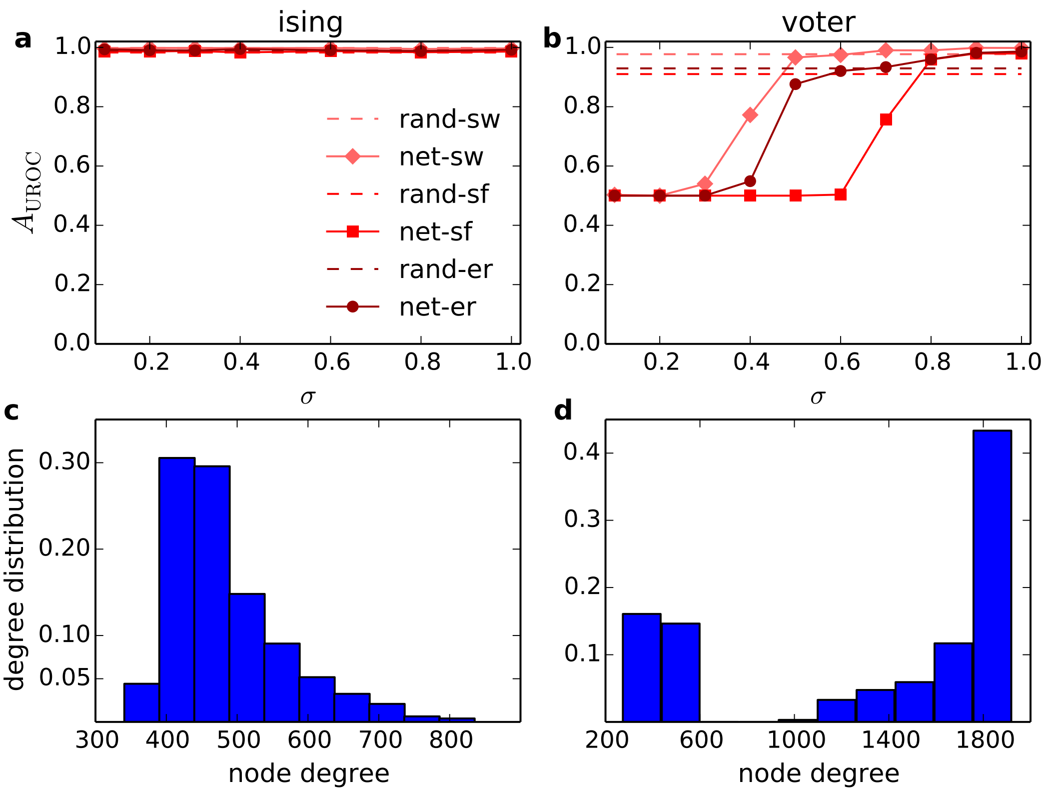
Supplementary Table S1 | Settings in numerical simulations. Parameter values in various binary-state dynamics and the period for initiating node states because of converging to steady state.

Model	Parameters	Convergent	Update period
Voter	—	Yes	100 (5 for N=100)
Kirman	$c_1 = 0.1$, $c_2 = 0.1$, $d = 0.08$	No	—
Ising Gluaber	$\beta = 2$	No	—
SIS	$\lambda = 0.2$, $\mu = 0.5$	No	—
Game	$\alpha = 0.1$, $\beta = 1$, $a = 6$, $b = 5$, $c = 1$, $d = 0$	No	—
Language	$s = 0.5$, $\alpha = 0.7$	No	—
Threshold	$M_k = 2/k$	Yes	5
Majority vote	$Q = 0.3$	Yes	10 (5 for N=100)

and SF networks. We found that for Ising dynamics, the results are almost not affected by the value of σ; but for Voter dynamics, a large value of σ is preferred. The possible reason is that, for dynamics with convergence such as Voter, the time series is dominated by all zeros or all ones. In the similarity networks, the nodes representing all zeros(or ones) time strings densely connect to each other, which leads to a bimodal degree distribution, as shown in Supplementary Fig. S2(**d**). We can see that there are more than 40% nodes in the rightmost bin. Thus, σ should be large enough to exclude these nodes, and then the reconstructing performance will approach high accuracy then, as shown in Fig. S2(**b**). For Ising dynamic, the degree distribution is like a bell shape, and there are no dominant zeros(or ones) time strings, so σ is not a key parameter. We also compare the performance of the networked base selection with the performance of randomly base selection. The networked indeed shows better performance, especially for dynamics with convergence.



Supplementary Figure S1 | Determination of threshold Δ. (**a**) AUROC as a function of threshold parameter Δ for the voter and Ising model on ER, SF and SW networks. (**b**) AUPR as a function of Δ for the two models and three networks. The network size $N = 100$ and $\langle k \rangle = 6$. The length of time series is 1.5×10^4. Other parameters of dynamics are shown in Supplementary Table S1.

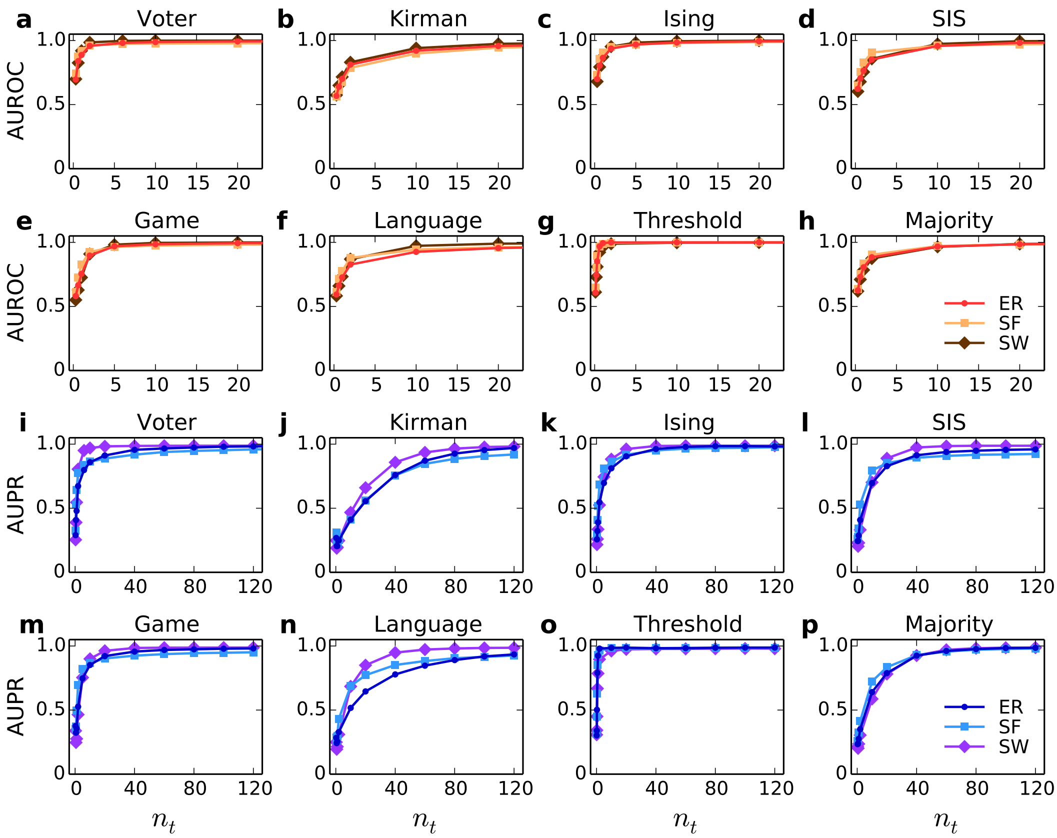


Supplementary Figure S2 | Determination of threshold σ. (**a,b**) AUROC as a function of threshold parameter σ for (**a**) the voter and (**b**) Ising model on ER, SF and SW networks, respectively. The dashed lines are the results of randomly selected bases. (**c**) The degree distribution of the constructed similarity network for Ising dynamic on ER network. (**d**) The degree distribution of the constructed similarity network for Voter dynamic on ER network. The network size $N = 100$ and $\langle k \rangle = 6$. The length of time series is 1.5×10^4. Other parameters of dynamics are shown in Supplementary Table S1.

There is an adjustable parameter λ in the lasso. In general, the parameter is determined by using cross-validation method, such as sklearn.linear_model.LassoCV in python. In terms of the cross-validation method, we obtained the proper value of λ, which is set to be 10^{-4} and 10^{-3} for reconstructing networks with $N \leqslant 500$ and $N = 1000$, respectively, in all reconstructions. All the convex optimizations are implemented in Python(version 2.7) and Sklearn(version 0.14).

2 Dependence of performance on data amount

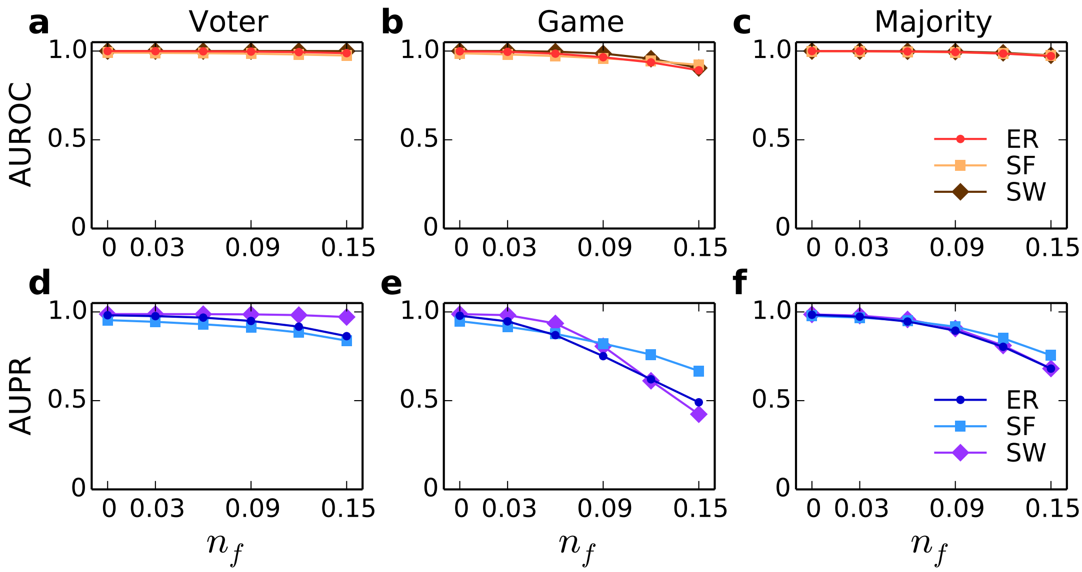
We examine how the length of time series affects reconstruction accuracy. We let n_t denote the ratio of the total length of time series normalized to the network size N. Supplementary Fig. S3 shows the reconstruction performance measured by AUROC and AUPR for various dynamics in combination with different types of networks. We find that AUROC and AUPR rapidly increases as n_t increases. After n_t exceeds a relatively small value, nearly full reconstruction can be achieved, which provides additional evidence for the high efficiency of our method. The results are summarized in Table II in the main text.



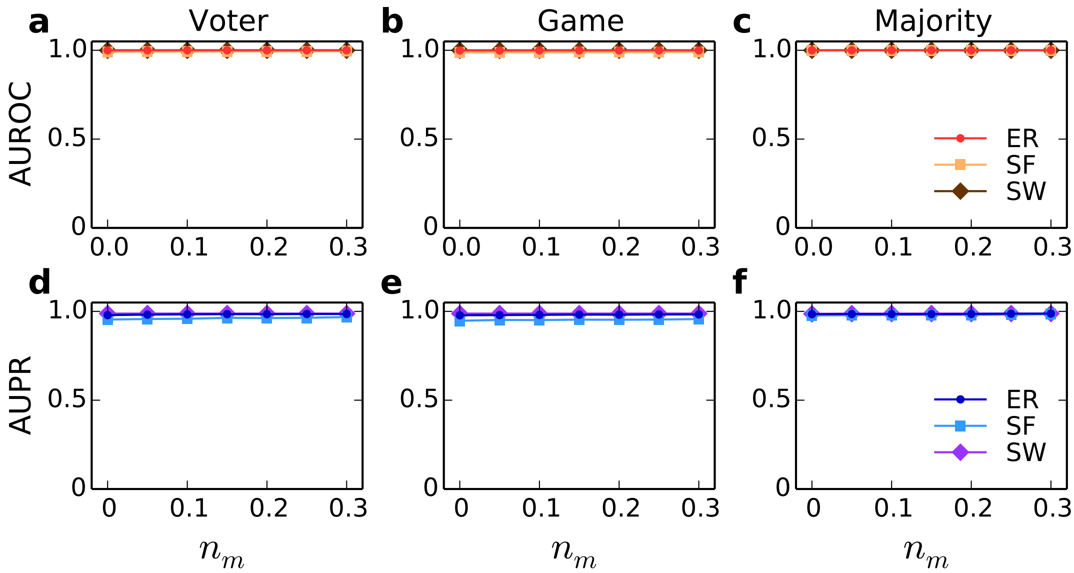
Supplementary Figure S3 | Reconstruction performance with respect to the length of time series. (**a-h**) AUROC and (**i-p**) AUPR as functions of the normalized length of time series n_t for various dynamics on ER, SF and SW networks. The network size $N = 500$ and $\langle k \rangle = 6$. Other parameter values of binary-state dynamics are shown in Supplementary Table S1.

3 Robustness against noise and missing data

Robustness against noise and missing data is important for evaluating the applicability of a method. We consider the scenario of noise-induced wrong records in time series. Specifically, we assume that a fraction n_{f} of binary states are wrong, and flip from 1 to zero or from zero to 1. The presence of unobservable nodes or missing data is quite often in the real situation. We assume that the data of a fraction of nodes, n_{m}, cannot be observed. We investigate the reconstruction accuracy as a function of n_{f} and n_{m}, respectively. As shown in Supplementary Fig. S4 and Supplementary Fig. S5, respectively, we find that high AUROC and AUPR remains in a wide range of n_{f} and n_{m}, providing strong evidence for the robustness of our reconstruction framework against measurement noise and missing data. The results are summarized in Table III in the main text.



Supplementary Figure S4 | Robustness against measurement noise. (**a,b,c**) AUROC and (**d,e,f**) AUPR as functions of the fraction n_f of wrong states in time series for the voter, Ising and majority model on ER, SF and SW networks. Parameters of networks and dynamics are the same as in Supplementary Fig. S3. $n_t = 100$.



Supplementary Figure S5 | Robustness against missing data. (**a,b,c**) AUROC and (**d,e,f**) AUPR as functions of the fraction n_m of unobservable nodes for the voter, Ising and majority model on ER, SF and SW networks. Parameters of networks and dynamics are the same as in Supplementary Fig. S3. $n_t = 100$.