# Supplementary Information for
# Reconstructing Complex Networks With Binary-State Dynamics

## 1 Theoretical validation of data-based linearization

We provide a heuristic analysis for the completely data-based linearization that gives rise to the general relationship

$$\langle s_i(\hat{t}+1)\rangle \approx c_i \cdot \sum_{j=1,j\neq i}^{N} a_{ij}\langle s_j(\hat{t})\rangle + d_i, \tag{1}$$

from general binary-state dynamics characterized by the switching probability

$$P_i^{01}(t) = F\left(m_i(t), k_i\right), \tag{2}$$

where $\langle s_i(\hat{t}+1)\rangle$ and $\langle s_j(\hat{t})\rangle$ can be obtained and calculated exclusively from data, $a_{ij}$ are the elements of adjacent matrix and are to be reconstructed, $c_i$ and $d_i$ are constants of node $i$, $P_i^{01}(t)$ is the switching probability of node $i$ from 0 to 1 in time $t$, and $F(m_i(t), k_i)$ is the dynamic function that depends on the number of active neighbors $m$ of node $i$ and the node's degree $k_i$. We will confirm through a heuristic analysis that the data-based linearization resulting from the merging process presented in the main text is valid for general binary-state dynamics subject to formula (2).

For nodes with only one neighbor, the linear relationship (1) can be rigorously proved. In this scenario, the number of active neighbors is either 0 or 1. Let $P_{\hat{t}}(1)$ denote the proportion of strings with single active neighbors in the set of base $\hat{t}$, and denote the proportion of strings with null active neighbors as $1 - P_{\hat{t}}(1)$. Let the switching probability of null active neighbors and single active neighbors be $f(0)$ and $f(1)$. Then we have

$$\begin{aligned}\langle s_i(\hat{t}+1)\rangle \approx \langle P_i^{01}(t)\rangle &= f(0)\left[1 - P_{\hat{t}}(1)\right] + f(1)P_{\hat{t}}(1) \\ &= \left[f(1) - f(0)\right]P_{\hat{t}}(1) + f(0)\end{aligned} \tag{3}$$

and

$$\sum_{j=1,j\neq i}^{N} a_{ij}\langle s_j(\hat{t})\rangle = P_{\hat{t}}(1). \tag{4}$$

Inserting Eq. (4) into Eq. (3), we have

$$\langle s_i(\hat{t}+1)\rangle \approx [f(1)-f(0)] \sum_{j=1,j\neq i}^{N} a_{ij}\langle s_j(\hat{t})\rangle + f(0), \tag{5}$$

which is a linear form that is subject to Eq. (1), because both $[f(1)-f(0)]$ and $f(0)$ are constants and they are determined by the specific binary-state dynamics.

Figure 1a,b shows two representative examples of reconstructing the local structure of a node with one neighbor for evolutionary game model and threshold model. We see explicit linear relationship for both models. With respect to different number of active neighbors in the original bases, two sets of groups are classified, as shown in Fig. 1c. (what does Fig. 1c means?)

For nodes with more than one neighbor, the linear relationship can be justified and predicted based on binomial distribution and Taylor linear approximation. For an arbitrary node, say, node $i$ with $k$ neighbors, we will substantiate the linear relationship between $\langle s_i(\hat{t}+1)\rangle$ and $\sum_{j=1,j\neq i}^{N} a_{ij}\langle s_j(\hat{t})\rangle$ resulting from the data-based linearization, where

$$\langle s_i(\hat{t}+1)\rangle \approx \langle P_i^{01}(t)\rangle = \sum_{m=0}^{k_i} F(m,k_i)P_{\hat{t}}(m), \tag{6}$$

and

$$\sum_{j=1,j\neq i}^{N} a_{ij}\langle s_j(\hat{t})\rangle = \sum_{m=0}^{k_i} mP_{\hat{t}}(m), \tag{7}$$

where $P_{\hat{t}}(m)$ represents the proportion of strings with $m$ active neighbors among all strings that belong to the set of base $\hat{t}$. The key to validating the linear relationship lies in the distribution that $P_{\hat{t}}(m)$ obeys.

Regarding the effect of the merging process (Fig. 1 in the main text), we hypothesize that $P_{\hat{t}}(m)$ follows is binomial distribution with different binomial coefficient $p_{\hat{t}}$. We denote the proportion of state 0 in data to be $p_0$. If the strings are randomly chosen for each set of a base, $P_{\hat{t}}(m)$ exactly obeys binomial distribution with binomial coefficient $p_0$. However, due to the process of selecting strings that are similar to each set of a base, the distribution will be biased toward the number of active neighbors in the base. Despite the original complex influence of the base and string selections based on Hamming distance, their effects can be simply regarded as selecting a group of strings with similar proportion of state 0 since we actually do not know which the node's neighbors are. This process leads to the binomial coefficient that depends on the base string. Figure 2a,b shows the comparison between the actual distribution of $P_{\hat{t}}(m)$ obtained from numerical simulations and the binomial distributions with different binomial coefficients in game and majority model, where the binomial coefficients approximately range from 0.4 to 0.6 because $p_0 \approx 0.5$ in the data. We see that $P_{\hat{t}}(m)$ can be well approximated by a binomial distribution with different parameter values, which indeed validates our binomial distribution hypothesis.

Based on the binomial distribution hypothesis, we have

$$P_{\hat{t}}(m) = \mathrm{C}_{k_i}^{m} {p_{\hat{t}}}^{m} (1 - p_{\hat{t}})^{k_i - m}. \tag{8}$$

Inserting Eq. (8) into Eq. (6) yields

$$\begin{aligned} \langle s_i(\hat{t}+1) \rangle &\approx \sum_{m=0}^{k_i} F(m, k_i) \mathrm{C}_{k_i}^{m} {p_{\hat{t}}}^{m} (1 - p_{\hat{t}})^{k_i - m} \\ &= \sum_{m=0}^{k_i} \mathrm{C}_{k_i}^{m} \sum_{l=0}^{m} \Big[ (-1)^{m-l} \mathrm{C}_{m}^{l} F(l, k_i) \Big] {p_{\hat{t}}}^{m}. \end{aligned} \tag{9}$$

The fact that $p_{\hat{t}}$ fluctuates around $p_0$ allows us to apply the Taylor series expansion around $p_0$ to Eq. (9), leading to

$$\begin{aligned} \langle s_i(\hat{t}+1) \rangle &\approx \sum_{m=0}^{k_i} \mathrm{C}_{k_i}^{m} \sum_{l=0}^{m} \Big[ (-1)^{m-l} \mathrm{C}_{m}^{l} F(l, k_i) \Big] p_0^{m} \\ &+ \sum_{m=0}^{k_i} \mathrm{C}_{k_i}^{m} \sum_{l=0}^{m} \Big[ (-1)^{m-l} \mathrm{C}_{m}^{l} F(l, k_i) \Big] m p_0^{m-1} (p_{\hat{t}} - p_0) \\ &+ \mathcal{O}(p_{\hat{t}} - p_0). \end{aligned} \tag{10}$$

Omitting the high-order term $\mathcal{O}(p_{\hat{t}} - p_0)$, we have

$$\begin{aligned} \langle s_i(\hat{t}+1) \rangle &\approx \sum_{m=0}^{k_i} \mathrm{C}_{k_i}^{m} \sum_{l=0}^{m} \Big[ (-1)^{m-l} \mathrm{C}_{m}^{l} F(l, k_i) \Big] (1 - m) p_0^{m} \\ &+ \sum_{m=0}^{k_i} \mathrm{C}_{k_i}^{m} \sum_{l=0}^{m} \Big[ (-1)^{m-l} \mathrm{C}_{m}^{l} F(l, k_i) \Big] m p_0^{m-1} p_{\hat{t}}. \end{aligned} \tag{11}$$

On the other hand, substitute Eq. (8) into Eq. (7) yields

$$\begin{aligned} \sum_{j=1, j \neq i}^{N} a_{ij} \langle s_j(\hat{t}) \rangle &= \sum_{m=0}^{k_i} m \mathrm{C}_{k_i}^{m} {p_{\hat{t}}}^{m} (1 - p_{\hat{t}})^{k_i - m} \\ &= k_i p_{\hat{t}}. \end{aligned} \tag{12}$$

Combining Eq. (11) and Eq. (12), we have

$$\begin{aligned} \langle s_i(\hat{t}+1) \rangle &\approx \sum_{m=0}^{k_i} \mathrm{C}_{k_i}^{m} \sum_{l=0}^{m} \Big[ (-1)^{m-l} \mathrm{C}_{m}^{l} F(l, k_i) \Big] (1 - m) p_0^{m} \\ &+ \left\{ \frac{1}{k_i} \sum_{m=0}^{k_i} \mathrm{C}_{k_i}^{m} \sum_{l=0}^{m} \Big[ (-1)^{m-l} \mathrm{C}_{m}^{l} F(l, k_i) \Big] m p_0^{m-1} \right\} \sum_{j=1, j \neq i}^{N} a_{ij} \langle s_j(\hat{t}) \rangle. \end{aligned} \tag{13}$$

Note that all variables in the first term on the right hand side of Eq. (13) are only determined by the binary-state dynamics and the node degree of $i$. Hence, the first term corresponding to $d_i$ is a constant with respect to node state $s_i$. In analogy, all variables in the coefficient of the second term are determined by the the binary-state dynamics and the node degree of $i$ as well, indicating the coefficient is a constant corresponding to $c_i$ in Eq. (1). Taken together, we theoretically justified that Eq. (13) is approximately a linear equation in the form of Eq. (1).

Figure 2c,d shows the relationship between $\langle s_i(\hat{t}+1)\rangle$ and $\sum_{j=1,j\neq i}^{N} a_{ij}\langle s_j(\hat{t})\rangle$(namely $\langle m\rangle$) of each set of bases and the linear relationship calculated by using Eq. 13 for game model and majority model with nonlinear and piecewise switching dynamics. We see that the theoretical predictions are in good agreement with the results from the merging process for linearization, which strongly validates the data-based linearization for general binary-state dynamics.

It is noteworthy that the key to the success of the data-based linearization lies in selecting similar strings subject to a base and the average over each set of bases. The selection similar strings accounts for the binomial distribution of active neighbors in a set, and different bases induces different binomial coefficients. Then the average of the binomial distributions leads to the relatively small range of $\langle m\rangle$ compared to the original range in the switching function, allowing us to use Taylor linear approximation. Moreover, high-order terms in the Taylor series expansion contribute little to the binomial distribution, which justifies the low-order approximation. Based on the linear relationship, the reconstruction of local structure can be realized by employing the Lasso without requiring the linear coefficients and intercept. In other words, the data-based linearization is general valid for arbitrary binary-state dynamics without any knowledge of the switching function.

## 2 AUROC and AUPR

To quantify the performance of our reconstruction method, we introduce two standard measurement indices, the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPR). True positive rate (TPR), false positive rate (FPR), Precision and Recall that are used to calculate AUROC and AUPR are defined as follows:

$$\mathrm{TPR}(l) = \frac{\mathrm{TP}(l)}{P}, \tag{14}$$

where $l$ is the cutoff in the edge list, $\mathrm{TP}(l)$ is the number of true positives in the top $l$ predictions in the edge list, and $P$ is the number of positives in the gold standard.

$$\mathrm{FPR}(l) = \frac{\mathrm{FP}(l)}{Q}, \tag{15}$$

where $\mathrm{FP}(l)$ is the number of false positive in the top $l$ predictions in the edge list, and Q is the number of negatives in the gold standard.

$$\mathrm{Precision}(l) = \frac{\mathrm{TP}(l)}{\mathrm{TP}(l) + \mathrm{FP}(l)} = \frac{\mathrm{TP}(l)}{l}, \tag{16}$$

$$\mathrm{Recall}(l) = \frac{\mathrm{TP}(l)}{P}, \tag{17}$$

where $\mathrm{Recall}(l)$, which is called sensitivity, is equivalent to $\mathrm{TPR}(l)$.

## 3 Computation details

Parameter values in the binary-state dynamics used for network reconstruction are displayed in Supplementary Table 1. 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 $\Delta$ 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 $\Delta$. The results are shown in Supplementary Fig 3. 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 $\Delta$ 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 more than $500$, we choose bases randomly instead of using the base-selection method presented in the main text, which will lose some reconstruction accuracy and increase requirement of data amount, but considerably reduce computational complexity. The reconstruction performance is displayed in Table II in the main text.

There is an adjustable parameter $\lambda$ 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 $\lambda$, which is set to be $10^{-4}$ and $10^{-3}$ for reconstructing networks with $N \leqslant 500$ and $N = 1000$, respectively, in all reconstructions.

## 4 Data sets of empirical networks

The details of several empirical networks in nature and society used in the main text for examining our reconstruction method are shown in Supplementary Table 2.

# 5 Dependence of performance on amount of data

We first examine how the number of base strings affects reconstruction accuracy. we denote $n_{\hat{t}}$ to be the number of bases divided by the network size $N$. Supllementary Fig. 4 shows AUROC and AUPR as functions of $n_{\hat{t}}$ for Voter (linear), Game (nonlinear) and Majority (piecewise) model, respectively. We see that due to the advantage of the lasso in reconstructing sparse signals, nearly perfect reconstruction is achieved from using a small amount of $n_{\hat{t}}$, and in a wide range of $n_{\hat{t}}$, such high accuracy is guaranteed, suggesting both high efficiency and robustness of our reconstruction method.

The length of time series is also an important for evaluating reconstruction efficiency as well, because all the required bases and subordinate strings are selected from limited time series. We denote $n_t$ to be the ratio of the total length of time series normalized to the network size $N$. Supplementary Fig. 5 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 reconstruction method.

# 6 Influence of network properties to reconstruction performance

We explore how network properties affect the reconstruction accuracy. In additional to investigations on model and real networks in paper, we explore the effect of mean degree $\langle k \rangle$ on the reconstruction accuracy, as shown in Supplementary Fig. 6. The reconstruction accuracy decreases as $\langle k \rangle$ increases. The main reason for this result is that the low-order approximation in the data-based linearization is better for smaller node degree, as discussed in Supplementary Sec. I. Moreover, with the increase of $\langle k \rangle$, the vector $\mathbf{X}_i$ to be reconstructed will become denser. Note that it usually requires larger amounts of data to reconstruct a denser signal by using the Lasso according to the compressed sensing theory. Thus, in general a network with larger $\langle k \rangle$ will be more difficult to be reconstructed.

We explore how network size affects data requirement. Supplementary Fig. 7 shows the minimum normalized length of time series $n_t^{min}$ to acquire at least $0.95$ AUROC and AUPR simultaneously as functions of network size $N$. We see that $n_t^{min}$ decreases as $N$ increases, which is because of network sparsity as well. In general, for the same average node degree $\langle k \rangle$, a network with larger size will be sparser, leading to a sparser vector $\mathbf{X}_i$. According to the compressed sensing theory, less amount of data is required for reconstructing a sparser $\mathbf{X}_i$, accounting for the decrease of $n_t^{min}$ with the increase of $N$. These results indicate that our reconstruction method is scalable and of practical importance for dealing with large real networked systems.

## 7 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 $n_{\rm f}$ of binary states are wrong, 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, say, $n_{\rm m}$, cannot be observed. We investigate the reconstruction accuracy as a function of $n_{\rm f}$ and $n_{\rm m}$. As shown in Supplementary Fig. 8 and Supplementary Fig. 9, respectively, we find that high AUROC and AUPR remains in a wide range of $n_{\rm f}$ and $n_{\rm m}$, providing strong evidence for the robustness of our reconstruction framework against measurement noise and missing data.

# 8 References

1.

2. M. Girvan and M. E. Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.

3. D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson. The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. *Behavioral Ecology and Sociobiology*, 54(4):396–405, 2003.

4. R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon. Superfamilies of evolved and designed networks. *Science*, 303(5663):1538–1542, 2004.

5. M. A. Van Duijn, E. P. Zeggelink, M. Huisman, F. N. Stokman, and F. W. Wasseur. Evolution of sociology freshmen into a friendship network. *Journal of Mathematical Sociology*, 27(2-3):153–191, 2003.

6. W. W. Zachary. An information flow model for conflict and fission in small groups. *Journal of anthropological research*, pages 452–473, 1977.

# Supplementary tables

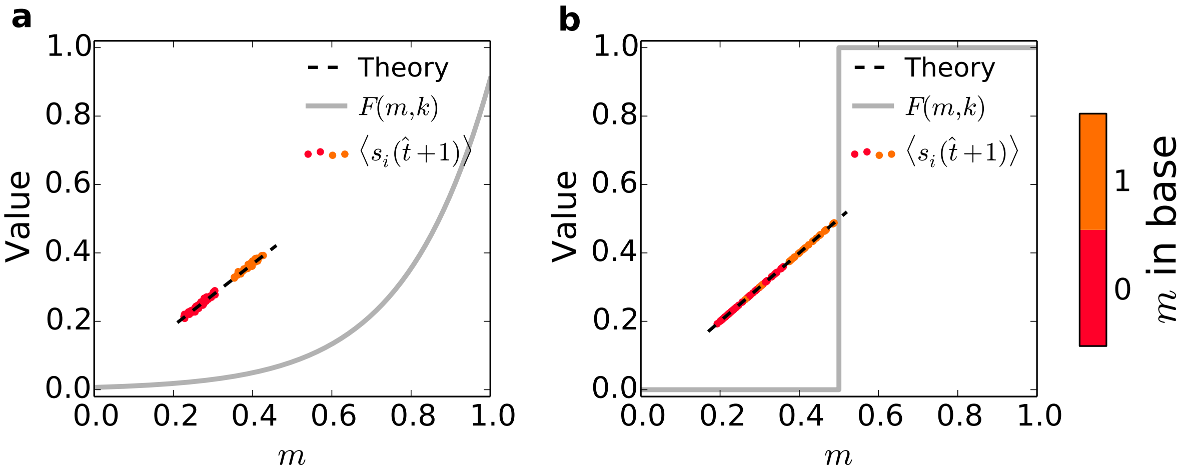
**Supplementary Table 1** | Parameter values in various binary-state dynamics and the period for initiating node states because of converging to steady state.

| Model | Parameters | Convergent | Update period |
|---|---|---|---|
| Voter | — | Yes | 100 |
| 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 | $\alpha = 0.1$, $\beta = 1$, $a = 5$, $b = 5$ | No | — |
| Language | $s = 0.5$, $\alpha = 0.7$ | No | — |
| Threshold | $M_k = 2/k$ | Yes | 5 |
| Majority vote | $Q = 0.3$ | Yes | 10 |

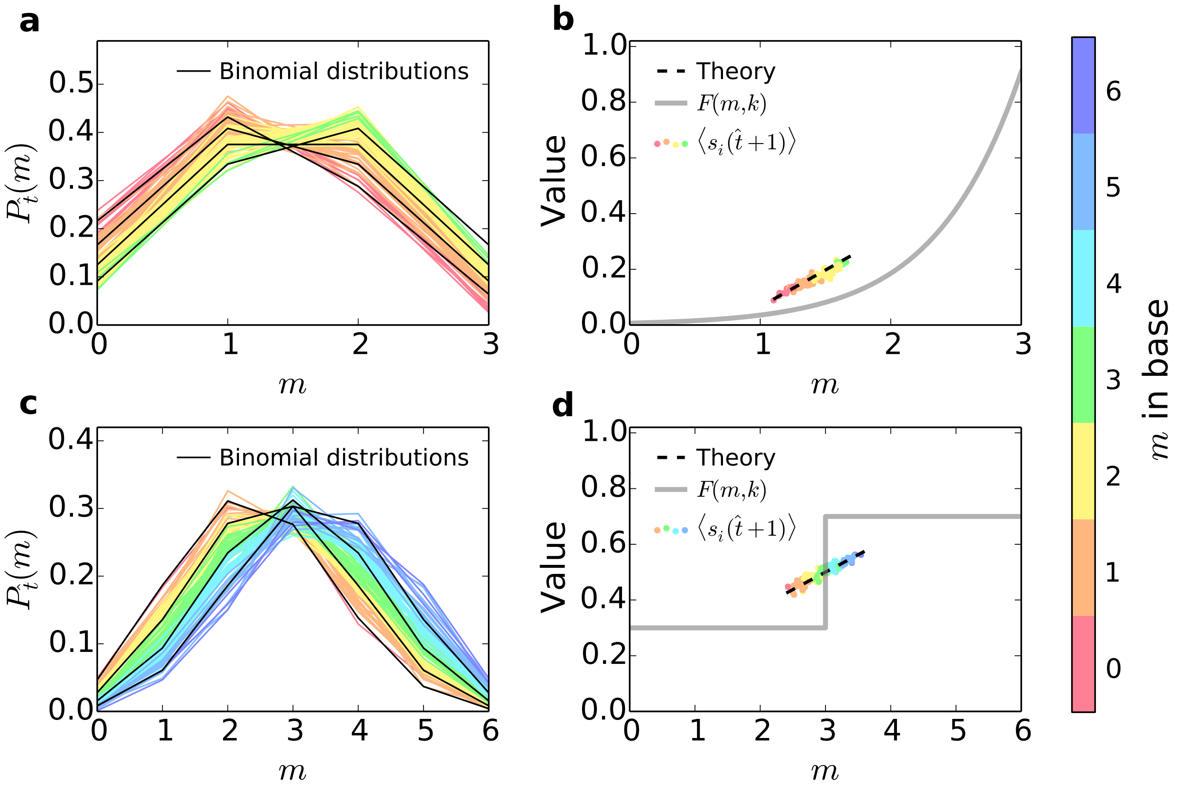
**Supplementary Table 2 | Feature and description of the empirical networks.** $N$ and $L$ denote the numbers of nodes and links of the empirical network studied in the paper.

| Name | $N$ | $L$ | Description |
|---|---|---|---|
| Dolphins [3] | 62 | 159 | Social network of dolphins |
| Football [2] | 115 | 613 | The network of American football games, Fall 2000. |
| Karate [6] | 34 | 78 | Social network of friendships of a karate club |
| Leader [4] | 32 | 96 | College students in a course about leadership |
| Polbooks [1] | 105 | 441 | A network of books about US politics |
| Prison [4, 5] | 67 | 182 | Social networks of positive sentiment (prison inmates) |
| Santa Fe [2] | 118 | 200 | Scientific collaboration network of the Santa Fe Institute |

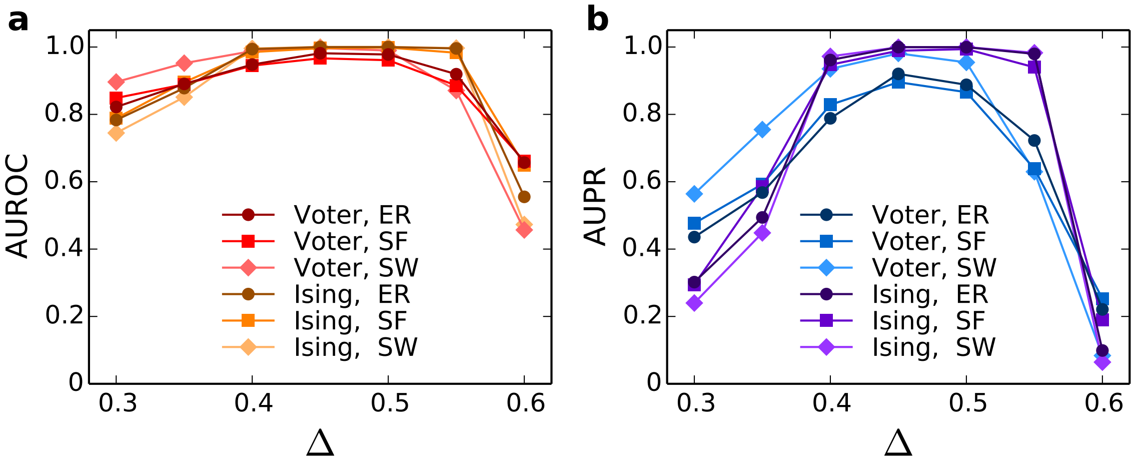
# Supplementary figures



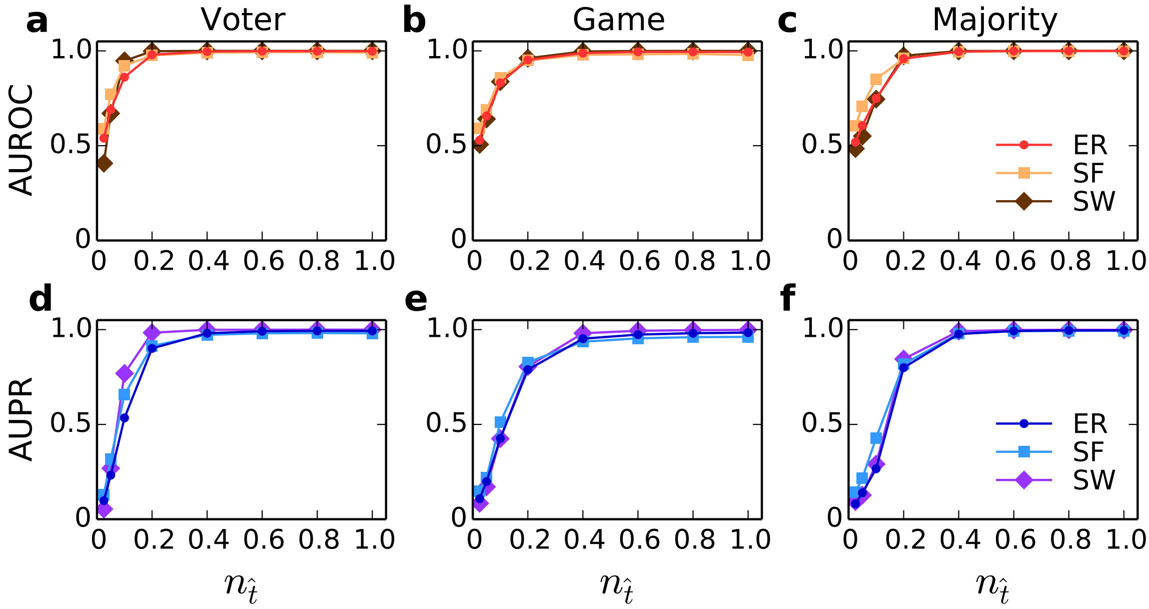
**Supplementary Figure 1 | linearization of switching function with $k_i = 1$.** Linearization of switching function for nodes with a single neighbor for (a) game model and (b) threshold model. The grey solid curves are the original switching functions, data points are the results of data-based linearization (E-q, (1)), and the dashed lines are theoretical predictions from Eq. (5). The color of data points represents two sets of subordinate strings whose base string has no active neighbors ($m = 0$) or has a single active neighbors ($m = 1$). For both nonlinear and piecewise switching functions, the theoretical predictions are in exact agreement with data-based linearization, because for $k_i = 1$ the linearization is rigorous without any approximation.



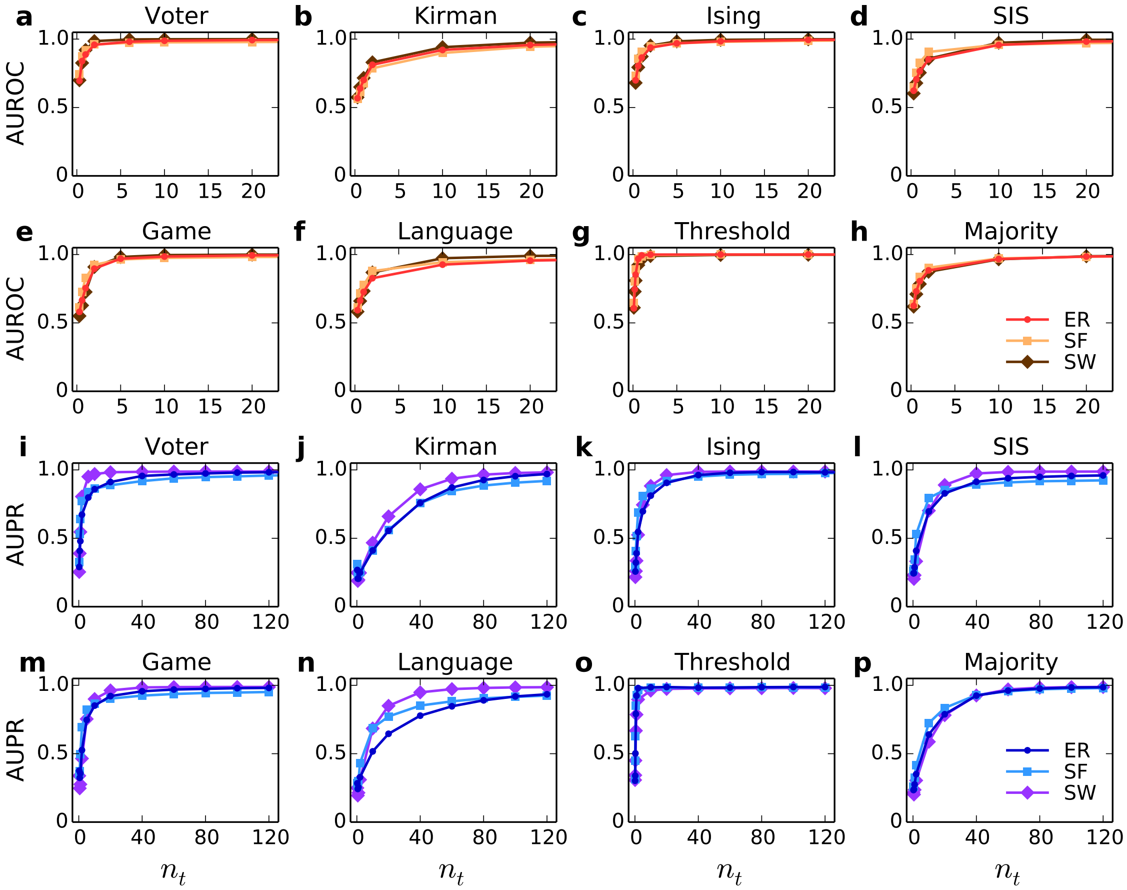
**Supplementary Figure 2 | Theoretical analysis of the data-based linearization.** (**a, b**) The distribution of active neighbors $m$ in subordinate strings subject to each base string and binomial distributions for reconstructing node $i$ with $k_i = 3$ for game model (a) and $k_i = 6$ for majority model (b), respectively. Each color of curves represents a set of subordinate strings whose base string has $m$ active neighbors. The distribution can be well described by binomial distributions under different binomial coefficients, as exemplified by black curves. There is a good agreement between the distribution of active neighbors in subordinate strings and binomial distributions. (**c, d**) The original switching function and the linearized function with theoretical prediction based on binomial distribution for game model (c) and majority model (d), respectively. The color of data points represents different sets of subordinate strings whose base string has different number of active neighbors $m$ (the same meaning as in (a, b)). The grey curves are the original switching function in the binary-state dynamics. The black dashed line are the theoretical prediction of the linear relationship through Eq. 13 based on binomial distribution and Taylor linear approximation. The theoretical predictions are in good agreement with numerical results.



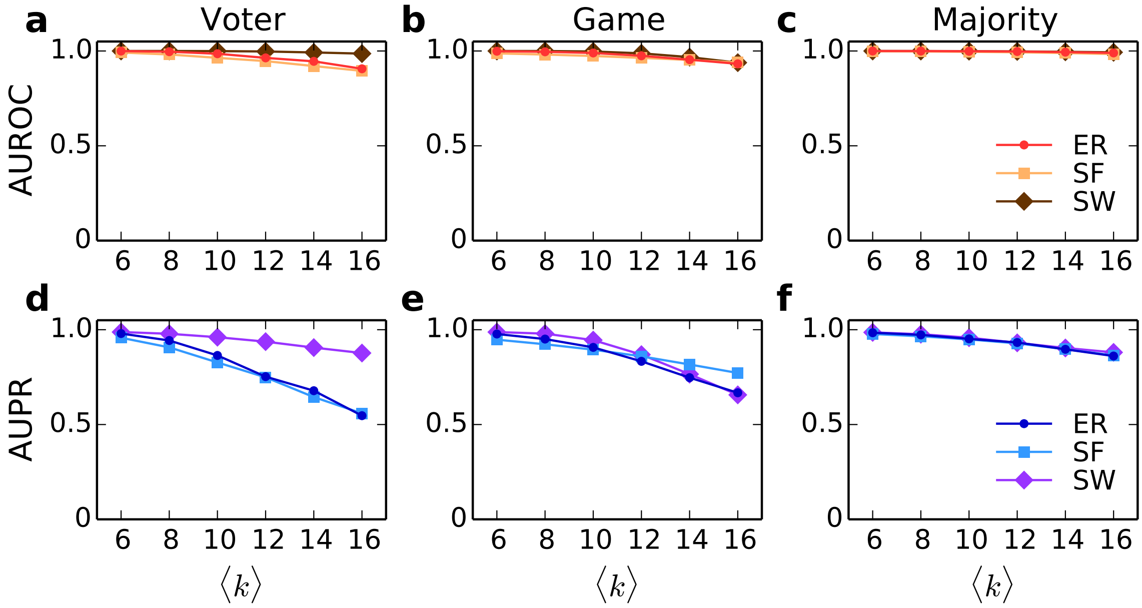
**Supplementary Figure 3 | Determination of threshold $\Delta$.** (a) AUROC as a function of threshold parameter $\Delta$ for Voter and Ising model on ER, SF and SW networks. (b) AUROC as a function of $\Delta$ for the two models and three networks. The network size $N = 100$ and $\langle k \rangle = 6$. The length of time series is $1.5 \times 10^4$. Other parameters of dynamics are shown in Supplementary Table 1.



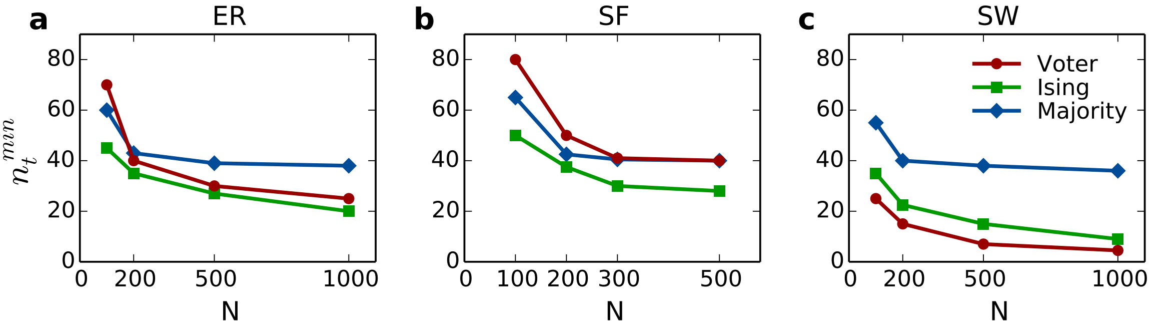
**Supplementary Figure 4 | Reconstruction performance with respect to the number of base strings.** (**a,b,c**) AUROC and (**d,e,f**) AUPR as functions of the normalized number of base strings $n_{\hat{t}}$ for Voter, Game and Majority model on ER, SF and SW networks. The network size $N = 100$ and $\langle k \rangle = 6$. The length of time series is $1.5 \times 10^4$. Other parameter values of binary-state dynamics are shown in Supplementary Table 1.



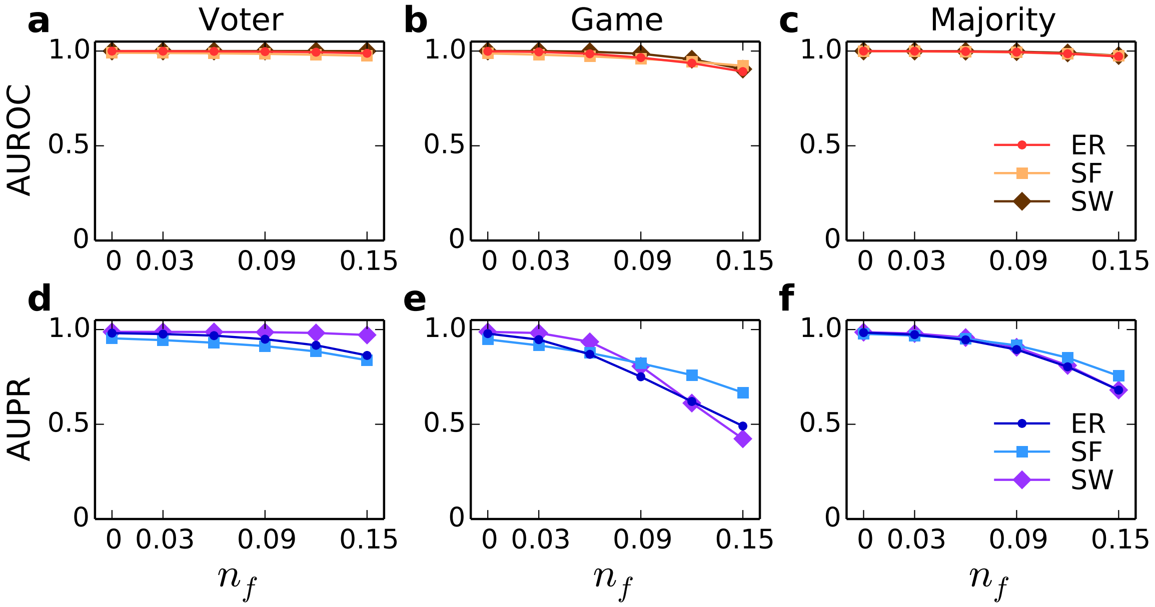
**Supplementary Figure 5 | Reconstruction performance with respect to the length of time series.** (**a-h**) AUROC and (**i-p**) AUPR as functions of the normalized length of time series $n_t$ for various dynamics on ER, SF and SW networks. The network size $N = 500$ and $\langle k \rangle = 6$. Other parameter values of binary-state dynamics are shown in Supplementary Table 1.



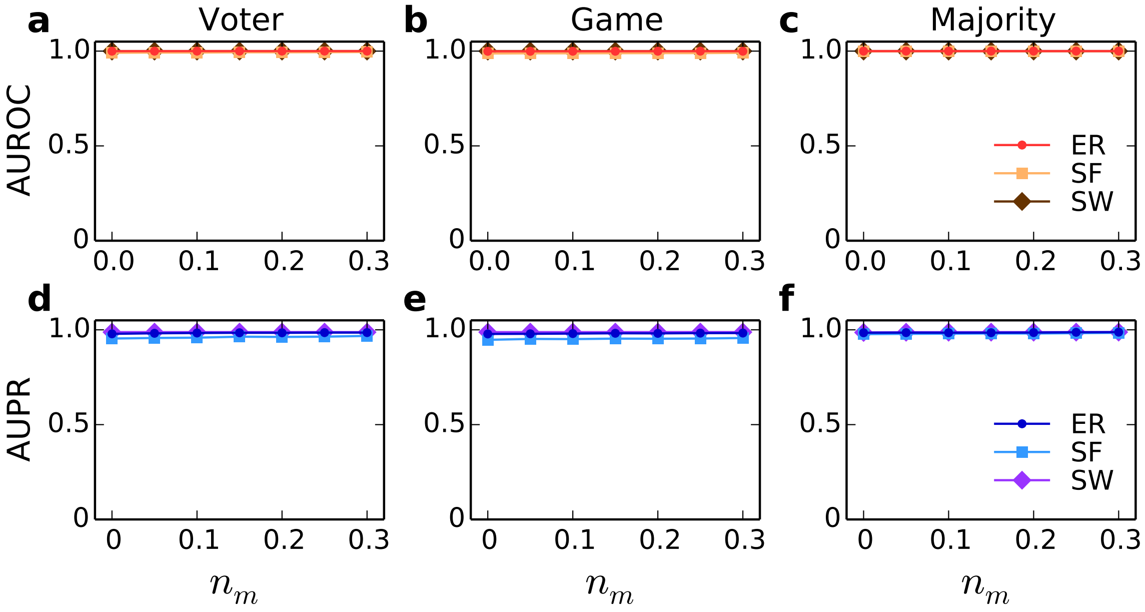
**Supplementary Figure 6 | Reconstruction performance affected by average node degree. (a,b,c)** AUROC and (**d,e,f**) AUPR as functions of the average node degree $\langle k \rangle$ for Voter, Game and Majority on ER, SF and SW networks. The network size $N = 500$ and normalized length of time series $n_t = 100$. Other parameter values of binary-state dynamics are shown in Supplementary Table 1.



**Supplementary Figure 7 | Reconstruction performance affected by network size.** The minimum normalized length $n_t^{min}$ to acquire at least $0.95$ AUROC and AUPR simultaneously as a function of network size $N$ for Voter, Ising and Majority on (a) ER, (b) SF and (c) SW networks. The mean degree of networks is $6$. Other parameter values of binary-state dynamics are shown in Supplementary Table 1.



**Supplementary Figure 8 | 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 Voter, Ising and Majority on ER, SF and SW networks. Parameters of networks and dynamics are the same as in Fig. 5. $n_t = 100$.



**Supplementary Figure 9 | Robustness against missing data.** (**a,b,c**) AUROC and (**d,e,f**) AUPR as functions of the fraction $n_m$ of unobservable nodes for Voter, Ising and Majority on ER, SF and SW networks. Parameters of networks and dynamics are the same as in Fig. 5. $n_t = 100$.