

ARTICLE

<https://doi.org/10.1038/s41467-022-30706-9>

OPEN

Full reconstruction of simplicial complexes from binary contagion and Ising data

Huan Wang¹, Chuang Ma², Han-Shuang Chen³, Ying-Cheng Lai⁴ & Hai-Feng Zhang¹✉

Previous efforts on data-based reconstruction focused on complex networks with pairwise or two-body interactions. There is a growing interest in networks with higher-order or many-body interactions, raising the need to reconstruct such networks based on observational data. We develop a general framework combining statistical inference and expectation maximization to fully reconstruct 2-simplicial complexes with two- and three-body interactions based on binary time-series data from two types of discrete-state dynamics. We further articulate a two-step scheme to improve the reconstruction accuracy while significantly reducing the computational load. Through synthetic and real-world 2-simplicial complexes, we validate the framework by showing that all the connections can be faithfully identified and the full topology of the 2-simplicial complexes can be inferred. The effects of noisy data or stochastic disturbance are studied, demonstrating the robustness of the proposed framework.

¹The Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, School of Mathematical Science, Anhui University, Hefei 230601, China. ²School of Internet, Anhui University, Hefei 230601, China. ³School of Physics and Material Science, Anhui University, Hefei 230601, China. ⁴School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ 85287, USA. ✉email: haifengzhang1978@gmail.com

In network science and engineering, a subfield of research is to find the network topology and nodal dynamical equations from data¹. This is important because networks are ubiquitous in the real world but the details of their connection topology and the intrinsic dynamical systems governing the properties and physical observables of the network are often unknown. The details are desired not only for understanding but also for protecting, disabling, or controlling the network dynamical behaviors (depending on the specific applications), and a viable way is to solve the inverse problem of determining the network details through observational data if they are available. As for any inverse problems in mathematics and physical sciences, the network inverse problem is challenging. Previous works in this area focused on “conventional” networks with pairwise interactions only^{1–16}. Existing methods include those which are based on drive-response^{3,5}, adaptive synchronization^{2,11}, noise correlation^{6,15}, compressive sensing^{7,9,17}, maximum likelihood estimation^{13,14,16}, and Granger causality^{4,8}. The data can be from continuous- or discrete-time dynamical processes. For example, the drive-response and adaptive synchronization methods use data from continuous-time nonlinear coupled systems^{2,3,5,11}, while the maximum likelihood estimation method is suitable for data from discrete-time dynamics^{13,14,16}. In this paper, motivated by the fact that higher-order networks have become a state-of-the-art subfield of research in network science^{18–24}, we develop a reconstruction framework for finding from time-series data network topology with higher-order interactions.

While pairwise or node-to-node interactions are the familiar type in networks, it has been recognized that higher-order interactions are also ubiquitous and important. For example, in a social network, the collective recommendation of multiple friends can often be more persuasive than the recommendation of a single friend to convince the individual to buy a new product. In a rumor spreading process, a piece of false news is likely to be accepted by an individual if it is shared or promoted simultaneously by many people^{25–27}. A similar situation occurs in neuronal networks, where a firing event is often the result of excitatory and inhibitory interactions among many neurons. In all these cases, the interaction arises simultaneously among a group of nodes in the network, and to describe the network by the conventional pairwise interactions is no longer adequate²⁸: higher-order interactions beyond the pairwise relationship must be taken into account. Mathematically, higher-order interactions can be described as hypergraphs or simplicial complexes²⁹, i.e., networks containing higher-order simplexes. In particular, a k -simplex describes the simultaneous interaction among $(k+1)$ nodes, where a zero-simplex specifies an isolated node (i.e., without any interaction), a 1-simplex represents the conventional pairwise interaction, a 2-simplex underlies the simultaneous interaction among three nodes, and so on.

The past three years have witnessed a growing interest in higher-order networks³⁰. For example, random walks on hypergraphs were studied, where a walker chooses the next destination depending on the number and the size of the shared hyperedges³¹. A family of random walks on hypergraphs with a parameter controlling the bias of the dynamics towards hyperedges of small or large size was constructed and the impacts of walk strategy and walk time on community detection were elucidated³². The stability conditions of the general dynamical processes on hypergraphs were found¹⁸, and a social contagion model on hypergraphs was constructed which presents dynamical phenomena such as first- and second-order transitions, bistability and hysteresis³³. A simplicial model of social contagion was proposed and it was demonstrated that the reinforcement mechanisms in 2-simplex can lead to a discontinuous phase transition³⁴. The impacts of the heterogeneity of simplicial complexes on the SIS (susceptible-infected-susceptible) spreading

model with collective and individual contagion were analyzed³⁵, and a pair approximation theory to study the SIS dynamics in simplicial complexes was developed, which was argued to be more accurate than the Markov-chain and mean field methods³⁶. A social communication model including idea integration and information transmission in simplicial complexes was proposed and the critical condition leading to the outbreak of information was identified³⁷. A simplicial activity driven model was proposed and the impact of both simplicial and temporally evolving interactions were analyzed³⁸. In terms of network reconstruction, a statistical method to detect higher-order interactions from network data of pairwise links has recently been developed²¹.

In this paper, we develop a framework to reconstruct complex networks with higher-order interactions from data. To be concrete, we focus on networks with 2-simplexes and assume that the dynamical processes on the network are social contagion and simplicial Ising dynamics that generate binary time-series data. Our method is of the statistical inference type pivoted on maximum likelihood estimation, with the aim to fully reconstruct both pairwise interactions (links) and 2-simplexes at the same time, thereby distinguishing our work from the recent method based on link data²¹. In particular, the central task is to estimate the probabilities of each node connecting to the reconstruction or target node (pairwise interaction) and of any two nodes forming a three-body 2-simplex with the target node. We articulate a two-step process to greatly enhance the computational efficiency and an effective truncation process to determine the final reconstructed structure of the simplicial complex. Using three synthetic and four real-world simplicial complexes, we demonstrate the accuracy of our reconstruction method and establish its robustness with respect to variations in the average degree of the network and stochastic fluctuations. Our work represents an initial effort in reconstructing complex networks with higher-order interactions based on observed time-series data.

Results

Simplicial complexes. A k -simplex σ is formed by a filled clique of a set of $k+1$ nodes $[v_0, \dots, v_k]$, which defines a $(k+1)$ -body interaction³⁹. A 1-simplex is two nodes connected by an edge, a 2-simplex is three nodes connected pairwise by edges and with an additional single face, i.e., a triangle, and a 3-simplex is four vertices connected pairwise by edges and joined by four faces, which are filled in to form a solid tetrahedron, and so on. A simplicial complex $\mathcal{K}$ composed of a set of nodes $\mathcal{V}$ is a collection of simplexes, with the additional requirement^{39,40} that if a simplex is in $\mathcal{K}$ ($\sigma \in \mathcal{K}$), then any simplex q composed of subsets of simplex σ should also be included in $\mathcal{K}$. For example, a 2-simplicial complex $\mathcal{K}$ is a collection of 0-, 1- and 2-simplexes.

Social contagion dynamics. Peer influence and reinforcement mechanisms are ubiquitous in the dynamical process of social contagion⁴¹, from which higher-order interactions in the network are originated. A social contagion model taking reinforcement into account on 2-simplicial complexes was proposed³⁴, which exploits the SIS type of spreading dynamics with binary-state dynamical variables. In particular, let S_i^t be the state of node i at time t . Each node has two possible states: susceptible ($S_i^t = 0$) or infected ($S_i^t = 1$). At the initial time, a fraction ρ_0 of nodes is infected. A susceptible node i can get infection from an infected neighbor j through their pairwise interaction (i, j) with probability β_1 . Node i can also be infected through a 2-simplex (i, j, k) , where both j and k have already been infected, with the probability β_2 , and this event can be understood as a synergistic reinforcement effect. For convenience, we set $\beta_1 = \alpha/k_1$ and $\beta_2 = \omega/k_2$, where α and ω are two non-zero positive constants, k_1 and k_2 are the

average degrees of the two-body and three-body interactions in a 2-simplicial complex, respectively. In general, we have $\alpha < \omega$ so as to ensure that the role of 2-simplex is embodied in the spreading dynamics. Each infected node recovers to the susceptible state with the probability μ . In our work, the values of β_1 and β_2 are selected near their respective thresholds to facilitate efficient and accurate reconstruction of 2-simplicial complexes. The effects of varying the values of β_1 and β_2 on the reconstruction accuracy are also studied (Sec. I of Supplementary Information (SI)).

Simplicial Ising dynamics. The Ising model arises in many fields due to its fundamental role in phase transitions in statistical physics. It has also been applied to many social systems^{42,43}. While the Ising dynamics on networks have been extensively studied^{44–46}, previous studies were exclusively conducted for networks with pairwise interactions only. To our knowledge, Ising dynamics on networks with higher-order interactions have not been studied.

To address the synergistic reinforcement effect of 2-simplex, we define a simplicial Ising dynamics on 2-simplicial complexes. Each node has two possible states: spin-down ($S_i^t = -1$) or spin-up ($S_i^t = +1$). At the initial time, the state of each node i is randomly assigned as $+1$ or -1 with equal probability. Defining the Hamiltonian as:

$$H(t) = -J_1 \sum_{(i,j)} S_i^t S_j^t - J_2 \sum_{(i,j,k)} S_i^t S_j^t S_k^t, \quad (1)$$

where J_1 and J_2 are the strengths of two-body and three-body interactions, and (i, j) and (i, j, k) denote the two-body and the three-body connections in the 2-simplicial complex, respectively. The first term in the Hamiltonian characterizes the interaction between the edges (i.e., two-body connections) and the second term contains three-body interactions from the 2-simplex. At each time step, the spin-flipping probability of each node i is given by $(1 + e^{\delta \Delta E_i^t})^{-1}$, where δ is the inverse temperature. The quantity

$$\Delta E_i^t = 2J_1 \sum_{(i,j) \in \partial_i} S_i^t S_j^t + 2J_2 \sum_{(i,j,k) \in \nabla_i} S_i^t S_j^t S_k^t$$

represents the change in the energy caused by a flipping of node i at time t , where ∂_i and ∇_i are the 1-simplex set and the 2-simplex set containing node i , respectively.

Statistical inference framework. For SIS and Ising processes taking place on a 2-simplicial complex of size N , the available time-series data representing the states of nodes at different time steps can be stored in a data matrix S , where each row is a time string representing all nodes' states at that time step and each column is a node's state at different time steps. We reconstruct 2-simplicial complexes from the data matrix S with our statistical inference framework. This task consists of three steps: (1) establishing the likelihood function based on the available data matrix S ; (2) obtaining the connection probabilities of two- and three-body interactions by maximizing the likelihood function according to the idea of the expectation maximization (EM) method, and (3) executing an improved two-step reconstruction strategy to significantly increase the computational efficiency. The details of the framework are described in Methods.

Quantification of reconstruction performance. We use F1 score to quantify the reconstruction accuracy⁴⁷, a global performance indicator defined as

$$F1 = \frac{2P * R}{P + R}, \quad (2)$$

where $P = TP/(TP + FP)$ and $R = TP/(TP + FN)$, with TP, FP, TN, FN being the numbers of true positive, false positive, true negative and false negative classes, respectively. A larger value of F1

corresponds to a higher accuracy and $F1 = 1$ indicates that the original network structure has been fully reconstructed with zero error.

Reconstructing synthetic and real-world simplicial complexes.

For readability, the results from the social contagion dynamics are presented in the main text, while those from the simplicial Ising dynamics are presented in Sec. III of SI.

Figure 1 presents results of reconstructing three synthetic 2-simplicial complexes (see Sec. “Construction of synthetic and real-world 2-simplicial complexes” in Methods on how these networks are constructed), where squares, diamonds and circles denote the performance of reconstructing the two-body connections while triangles with different orientations demonstrate the performance of reconstructing three-body connections. Several features can be seen from Fig. 1. First, the reconstruction accuracy increases with the length T of the time series and can reach the unity value for $T \gtrsim 8000$. Second, the average degrees k_1 and k_2 of two-body and three-body simplexes, respectively, have different effects on the reconstruction accuracy. In particular, as shown in Fig. 1a–c, a small value of k_1 tends to increase the reconstruction accuracy of both types of simplexes. This can be understood by noting that a small value of k_1 means that there are fewer two-body connections that need to be reconstructed, thereby enhancing the accuracy of the two-body connections for the same length of the time series. At the same time, fewer two-body connections reduce the complexity in reconstructing three-body connections and thereby improving the reconstruction accuracy. Regarding the effects of k_2 , Fig. 1d–f reveal that its value affects only the reconstruction accuracy of three-body connections and has little effect on the accuracy of reconstructing two-body connections that have no dependence on the three-body connections in a 2-simplicial complex. Third, the reconstruction accuracy of three-body interactions is lower than that of two-body interactions owing to the complicated structure of former and its dependence on the latter.

Figure 2 shows the results of reconstructing four real-world 2-simplicial complexes: Hypertext2009⁴⁸, Thiers12⁴⁹, InVS15⁵⁰, and LyonSchool^{51,52} (see Sec. “Construction of synthetic and real-world 2-simplicial complexes” in Methods for the details of these real-world networks). The basic parameters of these 2-simplicial complexes constructed from the datasets are listed in Table 1. It can be seen from Fig. 2 that, as for the real-world networks, the reconstruction accuracies for both the two-body and three-body interactions increase with the length of the time series. Remarkably, these network structures are quite irregular, complicating the reconstruction. Nonetheless, for $T = 20,000$, the F1 score can exceed 80%.

An issue of practical significance is the robustness of our reconstruction framework against random perturbations. To address this issue, we randomly flip a fraction f of infected states and the same number of susceptible states in the data matrix S (see Sec. “Details of the statistical inference framework” in Methods) and investigate the effect of f on the reconstruction accuracy as characterized by F1. The results are shown in Fig. 3 for three synthetic 2-simplicial complexes and three real-world 2-simplicial complexes. It can be seen that increasing the fraction f of flipping leads to a reduction of F1. In particular, the value of F1 for the two-body connections can be as high as 50% even when 30% of the infected states have been flipped ($f = 0.3$), attesting to the robustness of our framework in reconstructing pairwise links against stochastic fluctuations in the data.

Discussion

To find the network structure from observational data has been an active research field for more than fifteen years¹. In previous

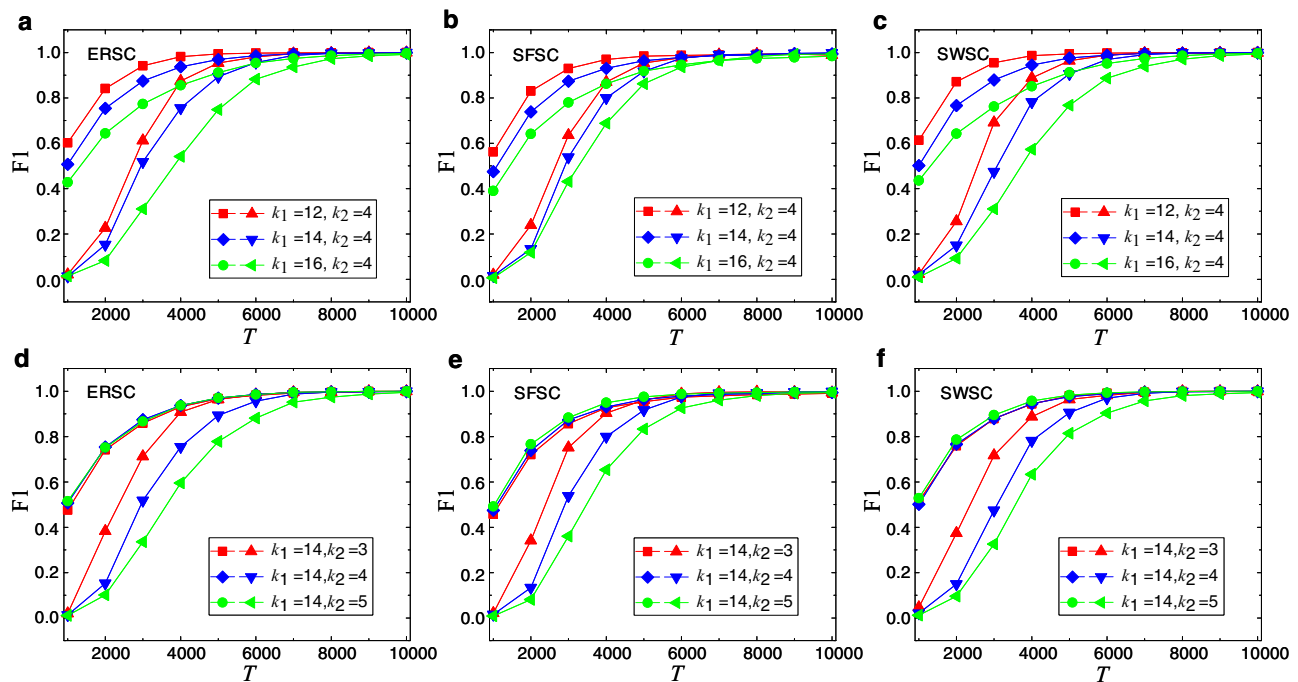


Fig. 1 Reconstruction performance for synthetic 2-simplicial complexes. Shown is F1 score as a function of the length T of the observational binary time series for three synthetic 2-simplicial complexes: (a, d) random simplicial complex (ERSC), (b, e) scale-free simplicial complex (SFSC), and (c, f) small-world simplicial complex (SWSC). In each panel, squares, diamonds and circles denote the performance of reconstructing the two-body connections while triangles with different orientations demonstrate the performance of reconstructing three-body connections, and different values of the average degree are distinguished by colors. All simplicial complexes have the same size $N = 200$. Other parameter values are $\alpha = 0.8$, $\omega = 2.4$, $\rho_0 = 0.2$, and $\mu = 1$. The results are averaged over five realizations.

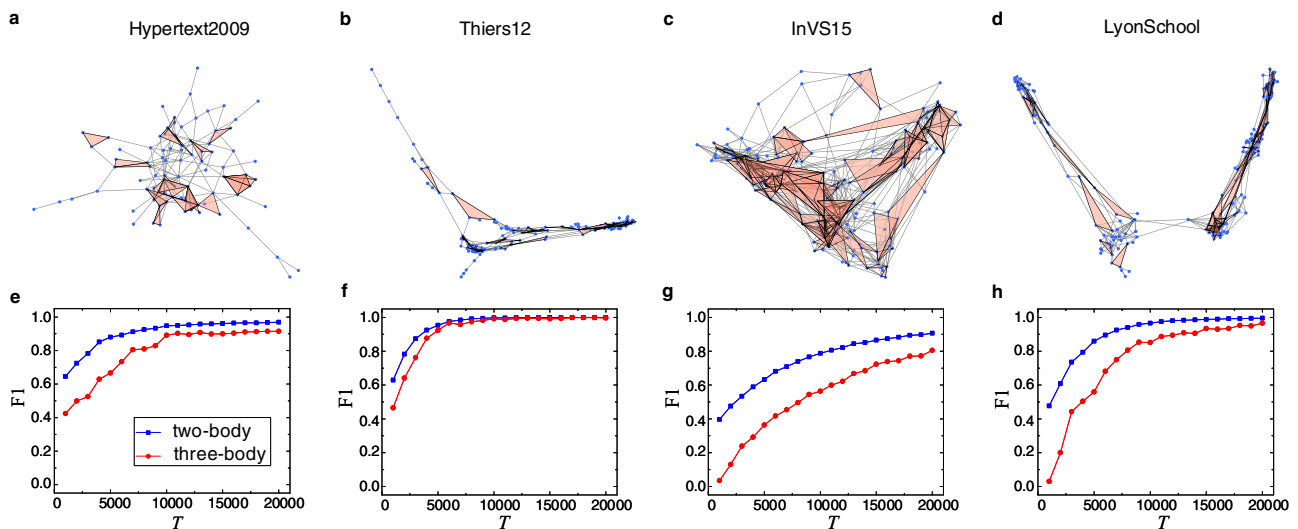


Fig. 2 Visualization and reconstruction performance for real-world 2-simplicial complexes. Visualization results for (a) Hypertext2009, (b) Thiers12, (c) InVS15, and (d) LyonSchool. The corresponding reconstruction performances are shown in panels (e–h) in terms of F1 score versus the length T of the available time series. The blue squares and red circles demonstrate the reconstruction performance of two-body and three-body connections, respectively. Parameter values are $\alpha = 0.3$, $\omega = 1$, $\rho_0 = 0.2$, and $\mu = 1$. Each data point is the result of averaging over five realizations.

studies, the term “network structure” is largely referred to as the collection of pairwise connections as characterized by the adjacency matrix of the network. Since the goal is to figure out whether there is a link between any two nodes, the existing methods focused on measures that are suitable for ascertaining the “two-body” interactions, such as those based on pairwise correlation or synchronization. From the beginning of modern network science and engineering slightly over two decades ago,

networks with only pairwise connections have represented the standard setting of study. Likewise, the inverse problem of data-based discovery of the network structure has been exclusively carried out in this setting. To our knowledge, in the current literature, the problem of finding higher-order connections in complex networks from time-series data has not been addressed.

Higher-order interactions are nonetheless ubiquitous in complex networks and its importance has been gradually recognized

with an accumulating interest, eventually generating an explosive growth of research recently^{18–24}. The structure of networks with higher-order connections, also known as simplicial complexes, are represented by tensors of high orders. For example, three-body interactions or 2-simplices in a network can be described by a tensor of rank 3. Structurally, simplicial complexes are significantly more sophisticated than the conventional networks with pairwise links only, and richer dynamics can be anticipated in the former, which have begun to be studied. From the point of view of inverse problem, to reconstruct simplicial complexes from time-series data represents a great challenge.

We have taken the first step to address this inverse problem. Focusing on complex networks with 2-simplices, we have developed a statistical inference framework by which all two-body and three-body interactions in the network can be found simultaneously from binary time-series data only, i.e., no prior

knowledge about the network to be reconstructed is required. The backbone of our reconstruction framework is maximum likelihood estimation that yields the probabilities of all the possible pairwise and three-body connections and a criterion to associate the probabilities with the actual interactions. To significantly increase the computational efficiency, we have proposed and tested a two-step process and a truncation process to determine the true structure of the simplicial complexes. The reconstruction framework has withstood tests on synthetic and real-world simplicial complexes with respect to accuracy and robustness against random fluctuations.

Many open problems remain. For example, our reconstruction framework is formulated in terms of binary time-series data from social contagion dynamics and simplicial Ising dynamics (Sec. III of SI). How to reconstruct higher-order networks from data generated by different dynamical processes needs to be studied. Also, our statistical inference method is developed for 2-simplicial complexes that are perhaps the “simplest” network structure beyond the conventional networks with pairwise interactions. To reconstruct networks with higher-order interactions such as 3-simplicial complexes and hypergraphs is worth pursuing. It is also necessary to develop methods to improve the reconstruction accuracy with shorter time series. We hope our work will stimulate further research in this emerging subfield of data-based reconstruction of complex networks with higher-order interactions.

Methods

Details of the statistical inference framework. We describe the details of our statistical inference framework through an illustrative example, as shown in Fig. 4, where a 2-simplicial complex with $N = 30$ nodes and its data matrix are illustrated

Table 1 Basic parameters of the four 2-simplicial complexes constructed from real-world datasets.					
Data set	Context	N	k_1	k_2	ζ
Hypertext2009	Conference	85	4.52	1.16	20
Thiers12	High school	156	4.56	1.21	20
InVS15	Workplace	211	7.52	2.19	20
LyonSchool	Primary school	222	5.42	2.18	50

N is the number of nodes, k_1 and k_2 are the average degrees of two-body and three-body connections, respectively, ζ is a threshold to filter out certain connections with low interaction frequency.

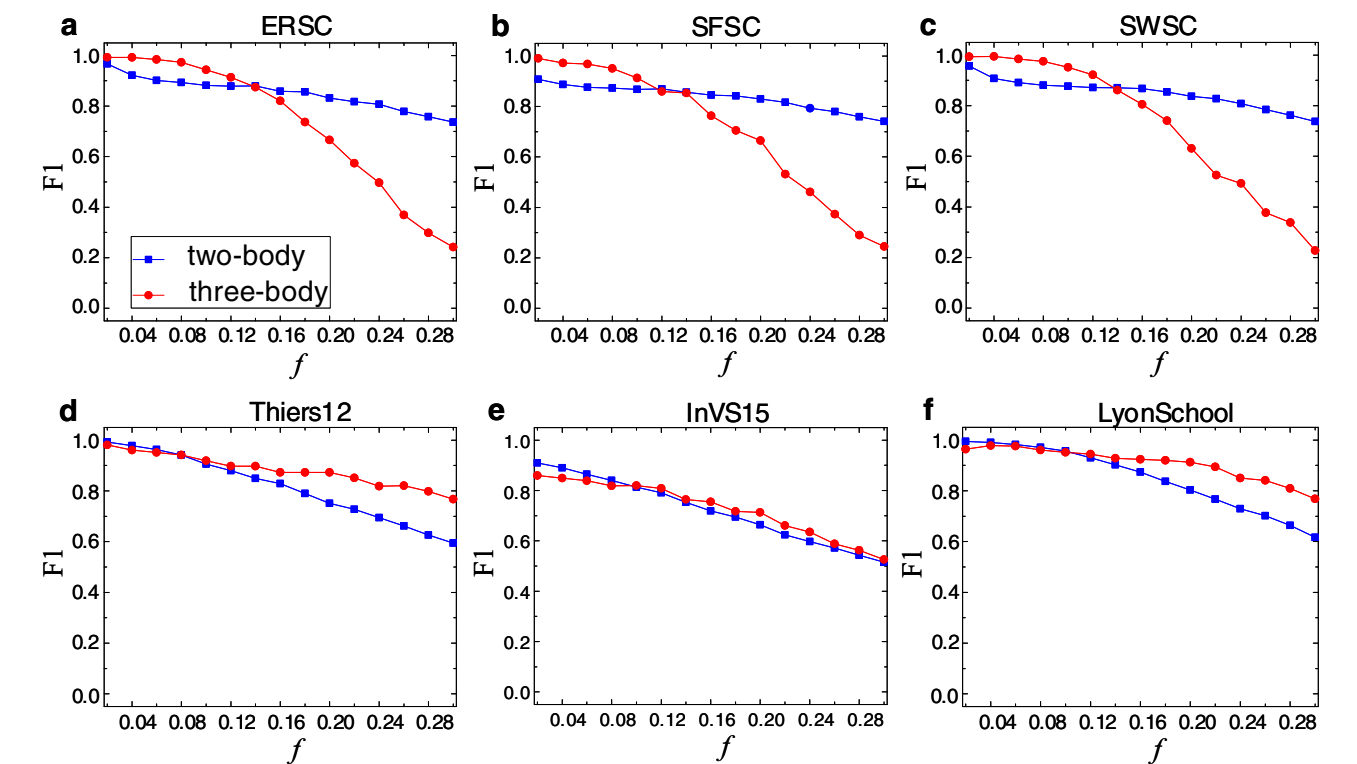
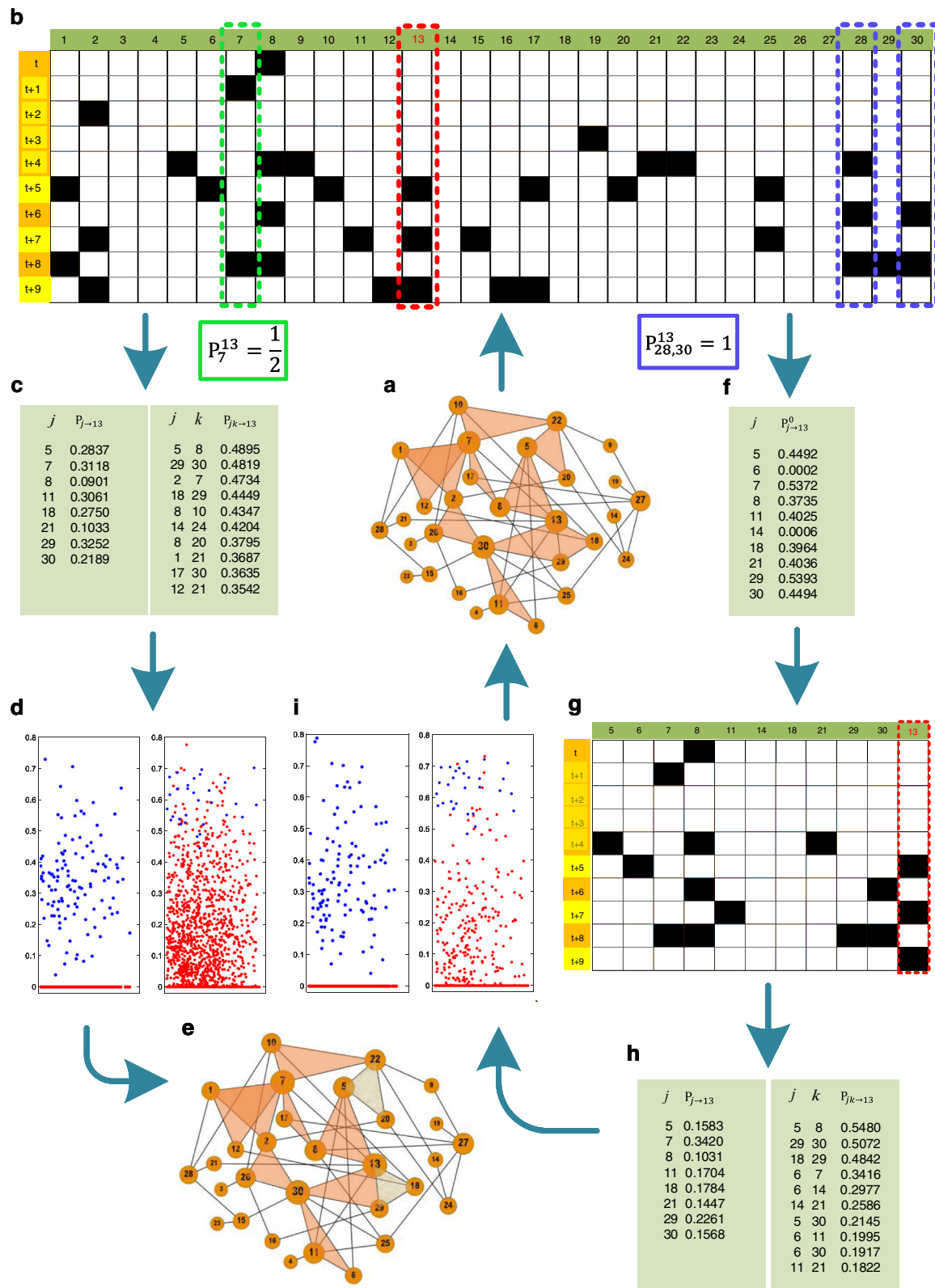


Fig. 3 Effect of random flipping ratio f on reconstruction performance for synthetic and real-world simplicial complexes. Shown is the F1 score for (a) random simplicial complex (ERSC), (b) scale-free simplicial complex (SFSC), (c) small-world simplicial complex (SWSC), (d) Thiers12, (e) InVS15, and (f) LyonSchool. The blue squares and red circles demonstrate the reconstruction performance of two-body and three-body connections, respectively. The parameter values for the synthetic simplicial complexes are $N = 200$, $k_1 = 12$, $k_2 = 4$, $T = 10,000$, $\alpha = 0.8$, $\omega = 2.4$, $\rho_0 = 0.2$, and $\mu = 1$. For the real-world simplicial complexes, the parameter values are $T = 20,000$, $\alpha = 0.3$, $\omega = 1$, $\rho_0 = 0.2$, and $\mu = 1$. Each data point is the result of averaging over ten realizations.



in Fig. 4a, b, respectively. For such a network hosting SIS dynamics, the probability of a susceptible node i (i.e., $S_i^t = 0$) being infected (i.e., $S_i^{t+1} = 1$) is determined only by the infected neighbors and the infected 2-simplexes in which two other nodes in the 2-simplex are both infected, at time t . The transition probability from the infected state to the susceptible state does not depend on the states of the neighbors, so it is only necessary to consider the transition probability from the susceptible state to the infected state for constructing the network. We stress that the details of the dynamical process, such as the infection probabilities β_1 and β_2 as

well as the recovery probability μ , are assumed to be unknown but only the binary time series of the nodal states are available. Figure 4 presents an illustrative example to describe the details of our method.

Establishing the likelihood function. Let $j \rightarrow i$ denote the event that node j has a direct impact on the state of node i . For example, node j can directly spread the virus or send a piece of information to node i , which means that node j is one of immediate neighbors of node i . Nodes i and j thus form a 1-simplex, a property

Fig. 4 Schematic illustration of reconstructing a 2-simplicial complex based on binary social contagion. **a** A 2-simplicial complex of size $N = 30$, where the black links represent 1-simplices and the orange shadows represent 2-simplices. **b** Data matrix S that stores all nodes' states at different time steps, where each row is a time string representing all nodes' states at that time step and each column is a node's state at different time steps, where the black and blank squares denote the 1 and 0 states, respectively. Take node 13 as an example (the red frame), the values of $P_{j \rightarrow 13}^{13} (\forall j \neq 13)$ and $P_{jk \rightarrow 13}^{13} (\forall j \neq k \neq 13)$ can be calculated from data matrix S : $P_{j \rightarrow 13}^{13} = 1/2$ (the green frame) and $P_{28,30 \rightarrow 13}^{13} = 1$ (the purple frame). **c** The values of $P_{j \rightarrow 13}$ and $P_{jk \rightarrow 13}$ are obtained through the EM algorithm, where only non-zero values of $P_{j \rightarrow 13}$ and the top 10 values of $P_{jk \rightarrow 13}$ are shown. **d** The values of $P_{j \rightarrow i}$ (each column in the left subgraph) and $P_{jk \rightarrow i}$ (each column in the right subgraph) for each node i based on the method described in Secs. 4.1.1 and 4.1.2, where the blue and red dots denote the actual and nonexistent two-body or three-body connections, respectively. **e** The 2-simplicial complex is inferred based on the probabilities in **(d)**, in which the two-body connections are exactly predicted, but two 2-simplices (5, 20, 22) and (13, 18, 29) in **(a)** cannot be predicted (marked by the light-yellow shadows). **f** The values of $P_{j \rightarrow 13}^0$ for node 13 obtained by iterating Eqs. 22–25, and only non-zero values are shown. **g** Compressed data matrix that records only the columns in S giving $P_{j \rightarrow 13}^0 > 0$. **h** The values of $P_{j \rightarrow 13}$ and $P_{jk \rightarrow 13}$ for node 13 based on the compressed data in **(g)**. **i** The values of $P_{j \rightarrow i}$ and $P_{jk \rightarrow i}$ for each node i . Finally, the full 2-simplicial complex in **(a)** can be exactly reconstructed by determining whether $P_{j \rightarrow i} > 0$ or $P_{jk \rightarrow i} > \Delta_i$.

that is independent of time t . Similarly, let $jk \rightarrow i$ denote the event that the synergistic reinforcement effect coming from nodes j and k has a direct impact on the state of node i , which is also independent of time. In the following, we determine the probabilities of node i and node j being connected and of three nodes i, j, k forming a three-body connection (i, j, k) .

The conditional probability of $S_i^{t+1} = 1$ and $j \rightarrow i$ given $S_j^t = 1$ and $S_i^t = 0$ can be written as

$$\begin{aligned} & P(S_i^{t+1} = 1, j \rightarrow i | S_i^t = 0, S_j^t = 1) \\ &= P(j \rightarrow i | S_i^t = 0, S_j^t = 1, S_i^{t+1} = 1) * P(S_i^{t+1} = 1 | S_i^t = 0, S_j^t = 1) \quad (3) \\ &= P_{j \rightarrow i} P_j^i, \end{aligned}$$

where $P_{j \rightarrow i} \triangleq P(j \rightarrow i | S_i^t = 0, S_j^t = 1, S_i^{t+1} = 1)$ is the probability of node i being infected by node j under the conditions $S_i^t = 0, S_j^t = 1$ and $S_i^{t+1} = 1, P_{j \rightarrow i} > 0$ indicates that node j is a neighbor of node i , and $P_j^i \triangleq P(S_i^{t+1} = 1 | S_i^t = 0, S_j^t = 1)$ is the probability of $S_i^{t+1} = 1$ under the conditions $S_i^t = 0$ and $S_j^t = 1$, which can be estimated from the data matrix S . Take the matrix in Fig. 4b as an example and suppose we wish to estimate the value of P_{13}^{13} , where nodes 13 (i.e., node 13 is the target node) and 7 are highlighted by red and green frames, respectively. It is necessary to extract each pair including the time string with $S_{13}^t = 0$ and its next time strings at $t + 1$. It can be seen that seven pairs of time strings can be extracted: $(t, t + 1), (t + 1, t + 2), (t + 2, t + 3), (t + 3, t + 4), (t + 4, t + 5), (t + 6, t + 7)$, and $(t + 8, t + 9)$. It can also be seen that two time moments: $t + 1$ and $t + 8$, satisfy the conditions that node 13 is in the susceptible state and node 7 is in the infected state. The only time at which node 13 can be infected at the next time step is $t + 8$. As a result, we have $P_{13}^{13} = 1/2$.

Similarly, the conditional probability of $S_i^{t+1} = 1$ and $jk \rightarrow i$ given $S_j^t S_k^t = 1$ and $S_i^t = 0$ can be written as

$$\begin{aligned} & P(S_i^{t+1} = 1, jk \rightarrow i | S_i^t = 0, S_j^t S_k^t = 1) \\ &= P(jk \rightarrow i | S_i^t = 0, S_j^t S_k^t = 1, S_i^{t+1} = 1) * P(S_i^{t+1} = 1 | S_i^t = 0, S_j^t S_k^t = 1) \quad (4) \\ &= P_{jk \rightarrow i} P_{jk}^i, \end{aligned}$$

where $P_{jk \rightarrow i} \triangleq P(jk \rightarrow i | S_i^t = 0, S_j^t S_k^t = 1, S_i^{t+1} = 1)$ is the probability of node i being infected through the synergistic interaction from nodes j and k , under the conditions $S_i^t = 0, S_j^t S_k^t = 1$, and $S_i^{t+1} = 1$, and $P_{jk \rightarrow i} > 0$ indicates that the three nodes i, j, k form a 2-simplex. The probability $P_{jk}^i \triangleq P(S_i^{t+1} = 1 | S_i^t = 0, S_j^t S_k^t = 1)$ can be estimated from the data matrix S in a similar way. Again, take the three nodes 13, 28, and 30 in Fig. 4b as an example. It can be seen that the time instants at which $S_{13}^t = 0, S_{28}^t = 1$ and $S_{30}^t = 1$ are fulfilled are $t + 6$ and $t + 8$. Because $S_{13}^{t+7} = 1$ and $S_{13}^{t+9} = 1$, we have $P_{28,30 \rightarrow 13}^{13} = 1$.

According to Eqs. 3, 4, the expected number of susceptible node i being infected at $t_m + 1$ is given by

$$\begin{aligned} E_i^{t_m+1} &= \sum_{j(j \neq i)} P(S_i^{t_m+1} = 1, j \rightarrow i | S_i^{t_m} = 0, S_j^{t_m} = 1) \Psi_j^{t_m} \\ &\quad + \sum_{j,k(j \neq k \neq i)} P(S_i^{t_m+1} = 1, jk \rightarrow i | S_i^{t_m} = 0, S_j^{t_m} S_k^{t_m} = 1) \Psi_{jk}^{t_m} + \varepsilon_i \\ &= \sum_{j(j \neq i)} P_{j \rightarrow i} P_j^i \Psi_j^{t_m} + \sum_{j,k(j \neq k \neq i)} P_{jk \rightarrow i} P_{jk}^i \Psi_{jk}^{t_m} + \varepsilon_i, \quad (5) \end{aligned}$$

where $\Psi_j^{t_m}$ represents the events that nodes j are infected at time t_m , similarly, $\Psi_{jk}^{t_m}$ is the events that both nodes j and k are infected at time t_m , and their values are zero or one. For example, if $\Psi_j^{t_m} = 1$, it means that node j is infected at time t_m ;

otherwise, $\Psi_j^{t_m} = 0$ when it is not infected at time t_m . The quantity ε_i represents the noise due to the errors from the collected data.

In general, the probability of a given number of events occurring in a fixed interval of time is characterized by the Poisson distribution, so we use it to capture the random nature of the times that node i is infected. An advantage of the Poisson distribution is that it makes a mathematical analysis and computations with the EM algorithm feasible^{53–56}. Specifically, the likelihood function can be described as

$$P\left(\left\{\Psi_i^{t_m+1}\right\}_{m=1,\dots,M} \middle| \Theta, \left\{\Psi_j^{t_m}\right\}_{m=1,\dots,M; j=1,\dots,N}\right) = \prod_{m(\Psi_i^{t_m}=0)} \frac{e^{-E_i^{t_m+1}} (E_i^{t_m+1})^{\Psi_i^{t_m+1}}}{\Psi_i^{t_m+1}!}, \quad (6)$$

where Θ denotes the set of variables $P_{j \rightarrow i}, P_{jk \rightarrow i}$ and ε_i . We have $\Psi_i^{t_m+1}! \equiv 1$ since $\Psi_i^{t_m+1}$ is either zero or one.

Maximizing the likelihood function based on EM algorithm. We next use the EM method to maximize the likelihood function⁵⁷ for determining the parameter Θ in Eq. 6. Taking the logarithm form of Eq. 6, we get

$$L(\Theta) = \sum_{m(\Psi_i^{t_m}=0)} \left(\Psi_i^{t_m+1} \log E_i^{t_m+1} - E_i^{t_m+1} \right) = \sum_{m(\Psi_i^{t_m}=0)} \left[\begin{aligned} & \Psi_i^{t_m+1} \log \left(\sum_{j(j \neq i)} P_{j \rightarrow i} P_j^i \Psi_j^{t_m} \right) \\ & + \sum_{j,k(j \neq k \neq i)} P_{jk \rightarrow i} P_{jk}^i \Psi_{jk}^{t_m} + \varepsilon_i \\ & - \left(\sum_{j(j \neq i)} P_{j \rightarrow i} P_j^i \Psi_j^{t_m} \right) \\ & + \sum_{j,k(j \neq k \neq i)} P_{jk \rightarrow i} P_{jk}^i \Psi_{jk}^{t_m} + \varepsilon_i \end{aligned} \right]. \quad (7)$$

Applying the Jensen's inequality to the logarithmic term on the right side of Eq. 7 yields

$$\begin{aligned} & \log \left(\sum_{j(j \neq i)} P_{j \rightarrow i} P_j^i \Psi_j^{t_m} + \sum_{j,k(j \neq k \neq i)} P_{jk \rightarrow i} P_{jk}^i \Psi_{jk}^{t_m} + \varepsilon_i \right) \\ &= \log \left(\sum_{j(j \neq i)} \rho_j^{t_m} \frac{P_{j \rightarrow i} P_j^i \Psi_j^{t_m}}{\rho_j^{t_m}} + \sum_{j,k(j \neq k \neq i)} \rho_{jk}^{t_m} \frac{P_{jk \rightarrow i} P_{jk}^i \Psi_{jk}^{t_m}}{\rho_{jk}^{t_m}} + \rho_{\varepsilon_i}^{t_m} \frac{\varepsilon_i}{\rho_{\varepsilon_i}^{t_m}} \right) \\ &\geq \sum_{j(j \neq i)} \rho_j^{t_m} \log \frac{P_{j \rightarrow i} P_j^i \Psi_j^{t_m}}{\rho_j^{t_m}} + \sum_{j,k(j \neq k \neq i)} \rho_{jk}^{t_m} \log \frac{P_{jk \rightarrow i} P_{jk}^i \Psi_{jk}^{t_m}}{\rho_{jk}^{t_m}} + \rho_{\varepsilon_i}^{t_m} \log \frac{\varepsilon_i}{\rho_{\varepsilon_i}^{t_m}} \quad (8) \\ &= \sum_{j(j \neq i)} \rho_j^{t_m} \log (P_{j \rightarrow i} P_j^i \Psi_j^{t_m}) + \sum_{j,k(j \neq k \neq i)} \rho_{jk}^{t_m} \log (P_{jk \rightarrow i} P_{jk}^i \Psi_{jk}^{t_m}) \\ &\quad + \rho_{\varepsilon_i}^{t_m} \log \varepsilon_i - \sum_{j(j \neq i)} \rho_j^{t_m} \log \rho_j^{t_m} - \sum_{j,k(j \neq k \neq i)} \rho_{jk}^{t_m} \log \rho_{jk}^{t_m} - \rho_{\varepsilon_i}^{t_m} \log \rho_{\varepsilon_i}^{t_m}, \end{aligned}$$

where the equality holds if

$$\rho_j^{t_m} = \frac{P_{j \rightarrow i} P_j^i \Psi_j^{t_m}}{\varepsilon_i + \left(\sum_{j'(j' \neq i)} P_{j' \rightarrow i} P_{j'}^i \Psi_{j'}^{t_m} + \sum_{j',k'(j' \neq k' \neq i)} P_{j'k' \rightarrow i} P_{j'k'}^i \Psi_{j'k'}^{t_m} \right)}, \quad (9)$$

$$\rho_{jk}^{t_m} = \frac{P_{jk \rightarrow i} P_{jk}^i \Psi_{jk}^{t_m}}{\varepsilon_i + \left(\sum_{j'(j' \neq i)} P_{j' \rightarrow i} P_{j'}^i \Psi_{j'}^{t_m} + \sum_{j',k'(j' \neq k' \neq i)} P_{j'k' \rightarrow i} P_{j'k'}^i \Psi_{j'k'}^{t_m} \right)}, \quad (10)$$

and

$$\rho_{\varepsilon_i}^{t_m} = \frac{\varepsilon_i}{\varepsilon_i + \left(\sum_{j(j \neq i)} P_{j \rightarrow i} \Psi_j^{t_m} + \sum_{j, k (j \neq k \neq i)} P_{j \rightarrow i} P_{j \rightarrow k} \Psi_j^{t_m} \right)}. \quad (11)$$

The maximization problem of Eq. 7 can then be transformed into maximizing the following equation:

$$\begin{aligned} \hat{L}(\Theta) = & \sum_{m(\Psi_i^{t_m}=0)} \sum_{j(j \neq i)} \left(\Psi_i^{t_m+1} \rho_j^{t_m} \log(P_{j \rightarrow i} \Psi_j^{t_m}) - \Psi_i^{t_m+1} \rho_j^{t_m} \log \rho_j^{t_m} - P_{j \rightarrow i} \Psi_j^{t_m} \right) \\ & + \sum_{m(\Psi_i^{t_m}=0)} \sum_{j, k (j \neq k \neq i)} \left(\Psi_i^{t_m+1} \rho_{jk}^{t_m} \log(P_{jk \rightarrow i} \Psi_{jk}^{t_m}) - \Psi_i^{t_m+1} \rho_{jk}^{t_m} \log \rho_{jk}^{t_m} - P_{jk \rightarrow i} \Psi_{jk}^{t_m} \right) \\ & + \sum_{m(\Psi_i^{t_m}=0)} \left(\Psi_i^{t_m+1} \rho_{\varepsilon_i}^{t_m} \log \varepsilon_i - \Psi_i^{t_m+1} \rho_{\varepsilon_i}^{t_m} \log \rho_{\varepsilon_i}^{t_m} - \varepsilon_i \right). \end{aligned} \quad (12)$$

Calculating the partial derivative of $\hat{L}(\Theta)$ with respect to $P_{j \rightarrow i}$, $P_{jk \rightarrow i}$ and ε_i and setting them to zero, we get

$$\frac{\partial \hat{L}(\Theta)}{\partial P_{j \rightarrow i}} = \sum_{m(\Psi_i^{t_m}=0)} \left(\frac{\Psi_i^{t_m+1} \rho_j^{t_m}}{P_{j \rightarrow i}} - \Psi_j^{t_m} \right) = 0, \quad (13)$$

$$\frac{\partial \hat{L}(\Theta)}{\partial P_{jk \rightarrow i}} = \sum_{m(\Psi_i^{t_m}=0)} \left(\frac{\Psi_i^{t_m+1} \rho_{jk}^{t_m}}{P_{jk \rightarrow i}} - \Psi_{jk}^{t_m} \right) = 0, \quad (14)$$

$$\frac{\partial \hat{L}(\Theta)}{\partial \varepsilon_i} = \sum_{m(\Psi_i^{t_m}=0)} \left(\frac{\Psi_i^{t_m+1} \rho_{\varepsilon_i}^{t_m}}{\varepsilon_i} - 1 \right) = 0, \quad (15)$$

which give

$$P_{j \rightarrow i} = \frac{\sum_{m(\Psi_i^{t_m}=0)} \left(\Psi_i^{t_m+1} \rho_j^{t_m} \right)}{\sum_{m(\Psi_i^{t_m}=0)} \left(\Psi_j^{t_m} \right)}, \quad (16)$$

$$P_{jk \rightarrow i} = \frac{\sum_{m(\Psi_i^{t_m}=0)} \left(\Psi_i^{t_m+1} \rho_{jk}^{t_m} \right)}{\sum_{m(\Psi_i^{t_m}=0)} \left(\Psi_{jk}^{t_m} \right)}, \quad (17)$$

$$\varepsilon_i = \frac{\sum_{m(\Psi_i^{t_m}=0)} \left(\Psi_i^{t_m+1} \rho_{\varepsilon_i}^{t_m} \right)}{\sum_{m(\Psi_i^{t_m}=0)} (1)}. \quad (18)$$

The six equations Eqs. 9–11 and Eqs. 16–18 can be used to solve $P_{j \rightarrow i}$, $P_{jk \rightarrow i}$ and ε_i . In particular, by initializing all values of $P_{j \rightarrow i}$, $P_{jk \rightarrow i}$, ε_i ($\forall j \neq k \neq i$) to be one and then calculating the values of $\rho_j^{t_m}$, $\rho_{jk}^{t_m}$ and $\rho_{\varepsilon_i}^{t_m}$ in Eqs. 9–11, we substitute them into Eqs. 16–18 to find the values of $P_{j \rightarrow i}$, $P_{jk \rightarrow i}$ and ε_i . We repeat this process until convergence is achieved. Since a single iterative process does not ensure global optimization, we carry out the above iteration process a number of times and choose the proper values that give the maximum of the quantity in Eq. 12.

As an example, as shown in Fig. 4c, the values of $P_{j \rightarrow 13}$ and $P_{jk \rightarrow 13}$ are given according to this iteration process, where $P_{j \rightarrow 13} > 0$ and the top 10 values of $P_{jk \rightarrow 13}$ are demonstrated. Similarly, all the values of $P_{j \rightarrow i}$ and $P_{jk \rightarrow i}$ can be calculated for each node i . As presented in Fig. 4d, each column above the abscissa corresponds to the predicted 1-simplex probabilities (the left subgraph of Fig. 4d) and 2-simplex probabilities (the right subgraph of Fig. 4d) of a node, and the blue and red dots denote the actual and nonexistent two-body or three-body connections, respectively.

An improved two-step reconstruction strategy. For a 2-simplicial complex structure with N nodes, when predicting the 2-simplexes of a node i , we randomly choose two nodes (e.g., j and k) and calculate the probability $P_{jk \rightarrow i}$, which requires calculating $(0.0ptN - 12)$ values. To reduce the computational load and increase the reconstruction accuracy, we articulate an improved two-step strategy. The particularity of simplicial complexes stipulates that the other two nodes forming a 2-simplex with node i must be the neighbors of node i , so it is not necessary to calculate the probability $P_{jk \rightarrow i}$ if node j or node k is not a neighbor of node i . The reconstruction process can then be divided into two steps. At the first step, the “approximate” neighborhood of each node is predicted and their corresponding columns in the data matrix S are extracted, leading to a compressed data matrix. At the second step, based on the compressed data matrix, the values of $P_{j \rightarrow i}$ and $P_{jk \rightarrow i}$ for each node i are predicted by iterating Eqs. 9–11 and 16–18. Our two-step method was not designed for the general challenging task of consistently inferring all the subfaces for arbitrarily higher-order simplices. In fact, our method requires the closure condition of simplicial complexes: it is necessary to know in advance that the network under reconstruction is a 2-simplicial complex. Given this premise, the two-step strategy infers first the two-body and then the three-body interactions (i.e., 2-simplex) from the inferred two-body interactions. While the two-step method is efficient to reconstruct 2-simplicial complexes, at the present it

cannot be used to reconstruct the hypergraphs because its second step is to find the triangles from the neighbors (i.e., edges).

For the first step, the predicted neighbors are not accurate because the three-body interactions have been ignored. In fact, the main purpose of this step is to determine an approximate range of neighbors to reduce the time for calculating $P_{jk \rightarrow i}$ ($\forall j \neq k \neq i$). Without taking into account three-body interactions, the expected number of susceptible nodes being infected at $t_m + 1$ can simply be expressed as

$$\begin{aligned} \tilde{E}_i^{t_m+1} = & \sum_{j(j \neq i)} P(S_i^{t_m+1} = 1, j \rightarrow i | S_i^{t_m} = 0, S_j^{t_m} = 1) * \Psi_j^{t_m} + \varepsilon_i \\ = & \sum_{j(j \neq i)} P_{j \rightarrow i}^0 \Psi_j^{t_m} + \varepsilon_i, \end{aligned} \quad (19)$$

where the notation $P_{j \rightarrow i}^0$ is used to emphasize that node j is only an “approximate” neighbor of node i . Assuming that the number Ψ_i of times of node i being infected in each time period obeys the Poisson distribution, we obtain the likelihood function as

$$\begin{aligned} P\left(\left\{\Psi_i^{t_m+1}\right\}_{m=1, \dots, M} \mid \left\{\tilde{E}_i^{t_m+1}\right\}_{m=1, \dots, M}; \left\{\Psi_j^{t_m}\right\}_{m=1, \dots, M}; j=1, \dots, N\right) \\ = \prod_{m(\Psi_i^{t_m}=0)} \frac{e^{-\tilde{E}_i^{t_m+1}} \left(\tilde{E}_i^{t_m+1}\right)^{\Psi_i^{t_m+1}}}{\Psi_i^{t_m+1}!}, \end{aligned} \quad (20)$$

where $\tilde{\Theta}$ denotes the set of variables $P_{j \rightarrow i}^0$ and ε_i . Taking the logarithm of Eq. 20, we have

$$L(\tilde{\Theta}) = \sum_{m(\Psi_i^{t_m}=0)} \left(\Psi_i^{t_m+1} \log \tilde{E}_i^{t_m+1} - \tilde{E}_i^{t_m+1} \right) = \sum_{m(\Psi_i^{t_m}=0)} \left[\Psi_i^{t_m+1} \log \left(\sum_{j(j \neq i)} P_{j \rightarrow i}^0 \Psi_j^{t_m} + \varepsilon_i \right) - \left(\sum_{j(j \neq i)} P_{j \rightarrow i}^0 \Psi_j^{t_m} + \varepsilon_i \right) \right]. \quad (21)$$

Using the EM method to maximize the likelihood function, we obtain the final parameters $\tilde{\Theta}$ as

$$P_{j \rightarrow i}^0 = \frac{\sum_{m(\Psi_i^{t_m}=0)} \left(\Psi_i^{t_m+1} \rho_j^{t_m} \right)}{\sum_{m(\Psi_i^{t_m}=0)} \left(\Psi_j^{t_m} \right)}, \quad (22)$$

$$\varepsilon_i = \frac{\sum_{m(\Psi_i^{t_m}=0)} \left(\Psi_i^{t_m+1} \rho_{\varepsilon_i}^{t_m} \right)}{\sum_{m(\Psi_i^{t_m}=0)} (1)}, \quad (23)$$

where

$$\rho_j^{t_m} = \frac{P_{j \rightarrow i}^0 \Psi_j^{t_m}}{\sum_{j(j \neq i)} P_{j \rightarrow i}^0 \Psi_j^{t_m} + \varepsilon_i}, \quad (24)$$

$$\rho_{\varepsilon_i}^{t_m} = \frac{\varepsilon_i}{\sum_{j(j \neq i)} P_{j \rightarrow i}^0 \Psi_j^{t_m} + \varepsilon_i}. \quad (25)$$

With the initial conditions for $P_{j \rightarrow i}^0$ and ε_i , the values of $P_{j \rightarrow i}$ and ε_i can be obtained by iterating Eqs. 22–25 until convergence is achieved. It is worth noting that $P_{j \rightarrow i}^0$ is a probability and we need to determine the “approximate” neighbors of the node under reconstruction. Theoretically, the “approximate” neighbors can be determined by testing whether $P_{j \rightarrow i}^0$ is non-zero. However, practically this is not feasible due to noise or deviations from the assumptions. For example, as shown in Fig. 4f, nodes 6 and 14 are not neighbors of node 13 even though $P_{6 \rightarrow 13}^0 = 0.0002$ and $P_{14 \rightarrow 13}^0 = 0.0006$. To overcome this difficulty, we articulate a truncation method for determining the neighbors of node i , as follows.

First, note that the time complexity of the second step can be significantly reduced when fewer neighbors are predicted, but too few predicted neighbors can lead to missing neighbors. On the contrary, too many neighbors would increase the time complexity. A solution is to use a reasonable truncation to determine the “approximate” neighbors of each node. To this end, we re-rank the probability $P_{j \rightarrow i}^0$ ($\forall j \neq i$) in a descending order and place a threshold Δ_i in the maximum gap defined as¹⁴:

$$\Delta_i = \arg \max_l \left[\frac{P_l'}{P_{l+1}'} (P_l' - P_{l+1}') \right]. \quad (26)$$

Next, we use Eq. 26 again to find a new threshold $\bar{\Delta}_i$ which is smaller than Δ_i . Finally, node j is regarded as an “approximate” neighbor of node i if $P_{j \rightarrow i}^0 > \bar{\Delta}_i$. The truncation method can ensure the detection of all real neighbors and 2-simplexes.

Once the “approximate” neighbors of node i have been obtained, the time series of these neighbors can be extracted (Fig. 4f, g). The neighbors of node i and its

2-simplexes can be quickly re-predicted based on the second step, i.e., by iterating Eqs. 9–11 and Eqs. 16–18 based on the compressed data matrix. For example, the prediction results for node 13 are shown in Fig. 4h and the values of $P_{j \rightarrow i} (\forall j \neq i)$ and $P_{jk \rightarrow i} (\forall j \neq k \neq i)$ for each node are presented in Fig. 4i. The actual two- and three-body connections of each node can then be determined based on the results in Fig. 4i. Because the identification of two-body connections has been refined in the second step, we simply assume that node j is a neighbor of node i if $P_{j \rightarrow i} > 0$. Following previous work^{14,58}, we assume that nodes i and j are connected when $P_{j \rightarrow i} > 0$ or $P_{i \rightarrow j} > 0$.

The case of three-body interactions is more complicated and the solution is sensitive to noise or errors. In fact, using the condition $P_{jk \rightarrow i} > 0$ as a criterion to detect (i, j, k) as a 2-simplex can lead to many false positives. Our solution is to re-rank $P_{jk \rightarrow i} (\forall j \neq k \neq i)$ in a descending order and obtain a new threshold $\hat{\Delta}_i$ by using Eq. 26 again. As a result, an actual 2-simplex (i, j, k) is formed when $P_{jk \rightarrow i} \geq \hat{\Delta}_i$. To remove the conflicts in the prediction, we assume that there exists a 2-simplex (i, j, k) when two of three conditions hold at least, e.g., $P_{jk \rightarrow i} \geq \hat{\Delta}_i$, $P_{ik \rightarrow j} \geq \hat{\Delta}_j$, and $P_{ij \rightarrow k} < \hat{\Delta}_k$, but a three-body cannot form when $P_{jk \rightarrow i} \geq \hat{\Delta}_i$, $P_{ik \rightarrow j} < \hat{\Delta}_j$, and $P_{ij \rightarrow k} < \hat{\Delta}_k$. Implementing the two-step strategy, we can reconstruct the whole 2-simplicial complexes. As shown in Fig. 4a, the 2-simplicial complex has been accurately reconstructed. Overall, the two-step strategy not only greatly reduces the computational time but also significantly improves the reconstruction accuracy (more details in Sec. II in SI).

Construction of synthetic and real-world 2-simplicial complexes

Synthetic 2-simplicial complexes. Here we describe the main steps of constructing synthetic 2-simplicial complexes of size N , average degrees of two-body and three-body interactions k_1 and k_2 , respectively.

Random simplicial complex (ERSC): First, a random graph is generated by connecting any two nodes with the probability p_1 . We then add 2-simplexes between any three nodes with the probability p_2 , where the formulas of p_1 and p_2 are³⁴:

$$p_1 = \frac{k_1 - 2k_2}{(N - 1) - 2k_2}, \quad (27)$$

$$p_2 = \frac{2k_2}{(N - 1)(N - 2)}. \quad (28)$$

A random 2-simplicial complex with the specified average degrees can then be constructed using the probabilities p_1 and p_2 .

Scale-free simplicial complex (SFSC): First, a scale-free network is generated, in which each new node connects m edges to the old nodes with degree preference⁵⁹. We then add 2-simplexes between any three nodes according to probability p_2 in Eq. 28. The average degree of 1-simplexes can be calculated as

$$k_1 = 2m + 2k_2 \left(1 - \frac{2m}{N}\right). \quad (29)$$

Small-world simplicial complex (SWSC): First, a small-world network⁶⁰ is generated from a regular lattice (all the nodes have the same degree $2m$) with rewiring probability p . We then add 2-simplexes between any three nodes according to probability p_2 in Eq. 28. The average degree of 1-simplexes is given by Eq. 29.

2-simplicial complexes from real-world data. In each real-world data set, the face-to-face interactions have been measured with a temporal resolution of 20 s. First, we generate a weighted network according to the data, where a weight represents the number of interactions between a pair of nodes in the whole time window. Second, we remove any link whose weight is less than a given threshold ζ and set the weights of retained links to one to generate an unweighted network. Finally, we cut the data into multiple segments with a temporal window of 5 min and record all the 2-simplexes. In particular, if three nodes communicate with each other in a short time, they are regarded as constituting a three-body connection. We record the frequencies of the 2-simplexes in each segment. According to the total frequency in all segments, we retain the first 50% of the 2-simplexes with the highest frequencies and count them as the actual 2-simplexes. The visualization of four real-world 2-simplicial complexes is shown in Fig. 2a–d.

Data availability

The SocioPatterns datasets were downloaded from <http://www.sociopatterns.org/datasets/>⁶¹. The source data generated in this study have been deposited on GitHub at: https://github.com/HuanWang2022/reconstruct_simplicial_complex.

Code availability

The code and datasets are available at: https://github.com/HuanWang2022/reconstruct_simplicial_complex⁶².

Received: 23 August 2021; Accepted: 13 May 2022;

Published online: 01 June 2022

References

- Wang, W.-X., Lai, Y.-C. & Grebogi, C. Data based identification and prediction of nonlinear and complex dynamical systems. *Phys. Rep.* **644**, 1–76 (2016).
- Yu, D., Righero, M. & Kocarev, L. Estimating topology of networks. *Phys. Rev. Lett.* **97**, 188701 (2006).
- Timme, M. Revealing network connectivity from response dynamics. *Phys. Rev. Lett.* **98**, 224101 (2007).
- Marinazzo, D., Pellicoro, M. & Stramaglia, S. Kernel method for nonlinear granger causality. *Phys. Rev. Lett.* **100**, 144103 (2008).
- Yu, D. & Parltitz, U. Inferring local dynamics and connectivity of spatially extended systems with long-range links based on steady-state stabilization. *Phys. Rev. E* **82**, 026108 (2010).
- 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).
- 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).
- Wu, X., Wang, W. & Zheng, W. X. Inferring topologies of complex networks with hidden variables. *Phys. Rev. E* **86**, 046106 (2012).
- Shen, Z., Wang, W.-X., Fan, Y., Di, Z. & Lai, Y.-C. Reconstructing propagation networks with natural diversity and identifying hidden sources. *Nat. Commun.* **5**, 1–10 (2014).
- Timme, M. & Casadiego, J. Revealing networks from dynamics: an introduction. *J. Phys. A Math. Theo.* **47**, 343001 (2014).
- Wu, X., Zhao, X., Lü, J., Tang, L. & Lu, J.-A. Identifying topologies of complex dynamical networks with stochastic perturbations. *IEEE Trans. Control Net. Sys.* **3**, 379–389 (2015).
- Casadiego, J., Nitzan, M., Hallerberg, S. & Timme, M. Model-free inference of direct network interactions from nonlinear collective dynamics. *Nat. Commun.* **8**, 1–10 (2017).
- Xiang, B.-B., Ma, C., Chen, H.-S. & Zhang, H.-F. Reconstructing signed networks via Ising dynamics. *Chaos* **28**, 123117 (2018).
- Ma, C., Chen, H.-S., Lai, Y.-C. & Zhang, H.-F. Statistical inference approach to structural reconstruction of complex networks from binary time series. *Phys. Rev. E* **97**, 022301 (2018).
- Zhang, Z., Chen, Y., Mi, Y. & Hu, G. Reconstruction of dynamic networks with time-delayed interactions in the presence of fast-varying noises. *Phys. Rev. E* **99**, 042311 (2019).
- Ma, C., Chen, H.-S., Li, X., Lai, Y.-C. & Zhang, H.-F. Data based reconstruction of duplex networks. *SIAM J. Appl. Dyn. Sys.* **19**, 124–150 (2020).
- Candes, E. J. & Tao, T. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Trans. Inf. Theor.* **52**, 5406–5425 (2006).
- de Arruda, G. F., Tizzani, M. & Moreno, Y. Phase transitions and stability of dynamical processes on hypergraphs. *Commun. Phys.* **4**, 24 (2021).
- Skardal, P. S. & Arenas, A. Higher order interactions in complex networks of phase oscillators promote abrupt synchronization switching. *Commun. Phys.* **3**, 218 (2021).
- Burgio, G., Arenas, A., Gómez, S. & Matamalas, J. T. Network clique cover approximation to analyze complex contagions through group interactions. *Commun. Phys.* **4**, 111 (2021).
- Young, J.-G., Petri, G. & Peixoto, T. P. Hypergraph reconstruction from network data. *Commun. Phys.* **4**, 135 (2021).
- Eriksson, A., Edler, D., Rojas, A., Domenico, M. D. & Rosvall, M. How choosing random-walk model and network representation matters for flow-based community detection in hypergraphs. *Commun. Phys.* **4**, 133 (2021).
- Kovalenko, K. et al. Growing scale-free simplices. *Commun. Phys.* **4**, 43 (2021).
- Ghorbanchian, R., Restrepo, J. G., Torres, J. J. & Bianconi, G. Higher-order simplicial synchronization of coupled topological signals. *Commun. Phys.* **4**, 120 (2021).
- Watts, D. J. & Dodds, P. S. Influentials, networks, and public opinion formation. *J. Consum. Res.* **34**, 441–458 (2007).
- Chierichetti, F., Lattanzi, S. & Panconesi, A. Rumor spreading in social networks. *Theo. Comp. Sci.* **412**, 2602–2610 (2011).
- Lehmann, S. & Ahn, Y.-Y. *Complex Spreading Phenomena in Social Systems* (Springer, 2018).
- Zhang, Z.-K. et al. Dynamics of information diffusion and its applications on complex networks. *Phys. Rep.* **651**, 1–34 (2016).
- Battiston, F. et al. Networks beyond pairwise interactions: structure and dynamics. *Phys. Rep.* **874**, 1–92 (2020).

30. Battiston, F. et al. The physics of higher-order interactions in complex systems. *Nat. Phys.* **17**, 1093–1098 (2021).
31. Carletti, T., Battiston, F., Cencetti, G. & Fanelli, D. Random walks on hypergraphs. *Phys. Rev. E* **101**, 022308 (2020).
32. Carletti, T., Fanelli, D. & Lambiotte, R. Random walks and community detection in hypergraphs. *J. Phys. Complex.* **2**, 015011 (2021).
33. de Arruda, G. F., Petri, G. & Moreno, Y. Social contagion models on hypergraphs. *Phys. Rev. Research* **2**, 023032 (2020).
34. Iacopini, I., Petri, G., Barrat, A. & Latora, V. Simplicial models of social contagion. *Nat. Commun.* **10**, 1–9 (2019).
35. Landry, N. W. & Restrepo, J. G. The effect of heterogeneity on hypergraph contagion models. *Chaos* **30**, 103117 (2020).
36. Matamalas, J. T., Gómez, S. & Arenas, A. Abrupt phase transition of epidemic spreading in simplicial complexes. *Phys. Rev. Res.* **2**, 012049 (2020).
37. Wang, D., Zhao, Y., Leng, H. & Small, M. A social communication model based on simplicial complexes. *Phys. Lett. A* **384**, 126895 (2020).
38. Petri, G. & Barrat, A. Simplicial activity driven model. *Phys. Rev. Lett.* **121**, 228301 (2018).
39. Hatcher, A. *Algebraic Topology* (Cambridge University Press, 2005).
40. Zhao, Y. & Maletić, S. *Simplicial Complexes in Complex Systems: In Search for Alternatives* (World Scientific, 2021).
41. Dakin, R. & Ryder, T. B. Dynamic network partnerships and social contagion drive cooperation. *Proc. R. Soc. B* **285**, 20181973 (2018).
42. Cibra, B. A. An introduction to the Ising model. *Am. Ma. Mon.* **94**, 937–959 (1987).
43. Stauffer, D. Social applications of two-dimensional Ising models. *Am. J. Phys.* **76**, 470–473 (2008).
44. Aleksiejuk, A., Holyst, J. A. & Stauffer, D. Ferromagnetic phase transition in Barabási–Albert networks. *Physica A* **310**, 260–266 (2002).
45. Bianconi, G. Mean field solution of the Ising model on a barabási–albert network. *Phys. Letts. A* **303**, 166–168 (2002).
46. Biswas, S. & Sen, P. Effect of the nature of randomness on quenching dynamics of the Ising model on complex networks. *Phys. Rev. E* **84**, 066107 (2011).
47. Powers, D. M. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *J. Mach. Learn. Tech.* **2**, 37–63 (2011).
48. Isella, L. et al. What’s in a crowd? Analysis of face-to-face behavioral networks. *J. Theo. Biol.* **271**, 166–180 (2011).
49. Fournet, J. & Barrat, A. Contact patterns among high school students. *PloS One* **9**, e107878 (2014).
50. Génois, M. & Barrat, A. Can co-location be used as a proxy for face-to-face contacts? *EPJ Data Sci.* **7**, 11 (2018).
51. Gemmetto, V., Barrat, A. & Cattuto, C. Mitigation of infectious disease at school: targeted class closure vs school closure. *BMC Infect. Dis.* **14**, 1–10 (2014).
52. Stehlé, J. et al. High-resolution measurements of face-to-face contact patterns in a primary school. *PloS ONE* **6**, e23176 (2011).
53. Karrer, B. & Newman, M. E. Stochastic blockmodels and community structure in networks. *Phys. Rev. E* **83**, 016107 (2011).
54. Newman, M. E. Equivalence between modularity optimization and maximum likelihood methods for community detection. *Phys. Rev. E* **94**, 052315 (2016).
55. Newman, M. E. & Reinert, G. Estimating the number of communities in a network. *Phys. Rev. Lett.* **117**, 078301 (2016).
56. De Bacco, C., Power, E. A., Larremore, D. B. & Moore, C. Community detection, link prediction, and layer interdependence in multilayer networks. *Phys. Rev. E* **95**, 042317 (2017).
57. Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B Method.* **39**, 1–22 (1977).
58. Ma, L., Han, X., Shen, Z., Wang, W.-X. & Di, Z. Efficient reconstruction of heterogeneous networks from time series via compressed sensing. *PloS ONE* **10**, e0142837 (2015).
59. Barabási, A.-L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509–512 (1999).
60. Watts, D. J. & Strogatz, S. H. Collective dynamics of small-world networks. *Nature* **393**, 440–442 (1998).
61. SocioPatterns Collaboration. <http://www.sociopatterns.org/>. Accessed Dec 2020.
62. Wang, H., Ma, C., Chen, H.-S., Lai, Y.-C. & Zhang, H.-F. Full reconstruction of simplicial complexes from binary contagion and Ising data. *HuanWang2022/reconstruct_simplicial_complex*, <https://doi.org/10.5281/zenodo.6501187> (2022).

Acknowledgements

H.W., C.M., H.-S.C. and H.-F.Z. acknowledge supports from the National Natural Science Foundation of China (61973001, 12005001, 11875069), the Natural Science Foundation of Anhui Province (2008085QF299), and the University Synergy Innovation Program of Anhui Province (GXXT-2021-032). Y.-C.L. acknowledges support from the Office of Naval Research under Grant No. N00014-21-1-2323.

Author contributions

H.W. and C.M. have contributed equally to this work. H.W., C.M., H.-S.C., Y.-C.L., and H.-F.Z. designed the research project, the models and methods. H.W. and C.M. performed the numerical analysis. H.-F.Z. and Y.-C.L. wrote the paper.

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/s41467-022-30706-9>.

Correspondence and requests for materials should be addressed to Hai-Feng Zhang.

Peer review information *Nature Communication* thanks Giovanni Petri, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at <http://www.nature.com/reprints>

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



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

© The Author(s) 2022

Supplementary Information for
Full reconstruction of simplicial complexes from binary contagion and Ising data

Huan Wang, Chuang Ma, Han-Shuang Chen, Ying-Cheng Lai, and Hai-Feng Zhang

Corresponding author: Hai-Feng Zhang (haifengzhang1978@gmail.com)

CONTENTS

I. Reconstruction accuracy under different epidemic conditions	2
II. Comparison of one-step method and two-step method	3
III. Reconstructing 2-simplicial complexes from simplicial Ising dynamics	3
A. Binary data from simplicial Ising dynamics	3
B. Statistical inference framework for simplicial Ising dynamics	4
C. Representative reconstruction results for simplicial Ising dynamics	7
IV. Reconstruction performance in terms of the number of events	8
V. O-information measure	8
VI. Effect of time-series length on reconstruction performance	9
VII. Supplementary figures	10
VIII. Supplementary references	22

I. RECONSTRUCTION ACCURACY UNDER DIFFERENT EPIDEMIC CONDITIONS

In general, the reconstruction accuracy depends on the epidemic conditions. In the main text, the parameter values for generating the data are selected near the epidemic threshold, i.e., near the triangle-driven and edge-driven transitions. To assess the effects of choosing parameter values in other regimes on the reconstruction results, we focus on two quantities: the rescaled edge infectivity α and the rescaled triangular infectivity ω , and study their impact on the reconstruction accuracy of two-body and three-body connections. In particular, the social contagion probabilities are $\beta_1 = \alpha/k_1$ and $\beta_2 = \omega/k_2$, where k_1 and k_2 are the average degrees of two-body and three-body connections in a 2-simplicial complex, respectively.

Different values of β_1 and β_2 can have significantly different impacts on the reconstruction results. For large values of β_1 or β_2 , many nodes are infected, making it difficult to judge which infected nodes have spread the infection to the node under reconstruction. On the contrary, for small values of β_1 and β_2 , it is difficult for a susceptible node to be infected by other nodes, leading to a lack of the useful data for the reconstruction task. It is then necessary to select the values of β_1 and β_2 properly to achieve acceptable reconstruction accuracy. We have systematically investigated the different parameter settings (e.g., in terms of the rescaled infectivities α and ω) and their effects on the reconstruction accuracy.

To be concrete, we study a scale-free simplicial complex (SFSC) with $N = 200$, $k_1 = 14$ and $k_2 = 4$. Figure S1 shows the average fraction ρ^* of infected nodes in the stationary state versus α for different values of ω . It can be seen that, as ω increases, the nature of the phase transition in the underlying social contagion dynamics changes from continuous to discontinuous. Further, different initial values of the density ρ_0 of the infected nodes can affect the steady-state infection density ρ^* associated with the healthy and endemic states in the bistable region.

We then choose three different values of α (0.5, 0.8 and 1.5), which are below, near and above the edge epidemic threshold α_c , respectively, and investigate the effect of different values of ω on the reconstruction accuracy. Concretely, we choose $\omega = 0.8, 2.4$ and 4.0 , which are below, near and above the triangular epidemic threshold ω_c , respectively. Figure S2(a) reveals that, for $\alpha = 0.5$, the reconstruction accuracies of two-body and three-body connections are not high, because it is difficult for social contagion in simplicial complexes to propagate when the rescaled edge infectivity α is small. Increasing the value of ω can lead to an improvement in the reconstruction accuracy. For example, as shown in Fig. S2(b), for $\alpha = 0.8$ and reasonably long time series, the reconstruction accuracies of two-body and three-body connections for $\omega = 2.4$ are the highest, as a small value of ω means an insignificant synergistic reinforcement effect from the three-body connections but a large value of ω will weaken the interactions from the two-body connections, especially when the value of α is near the edge threshold: both effects lead to difficulties in reconstructing two-body or three-body connections. Figure S2(c) shows, for $\alpha = 1.5$, the reconstruction accuracies of two-body and three-body connections for $\omega = 2.4$ are generally higher than those in the other two cases, because a small value of ω (e.g., $\omega = 0.8$) is not able to generate a strong synergistic reinforcement effect from the three-body connections while a large value of ω (e.g., $\omega = 4$) will cause most nodes to be infected. Taken together, the highest possible reconstruction accuracies are achieved when the values of α and ω are near their respective epidemic thresholds.

II. COMPARISON OF ONE-STEP METHOD AND TWO-STEP METHOD

In order to compare the accuracy and time complexity of one-step and two-step methods, we have used scale-free simplicial complex (SFSC) with different sizes and average degrees. The results are shown in Figs. S3 and S4. In particular, Fig. S3 shows that the accuracy of the two-step method is higher than that of the one-step method for both two-body and three-body reconstruction. The superiority of two-step method is more evident when the network size is larger. Figure S4 demonstrates the running time of the two methods, which are implemented in MATLAB2016a and run on a Linux machine with 2.60-GHz Intel processor, 28 CPU cores, and 192-GB RAM. It can be seen that the required computational time of the two-step method is more than one order of magnitude lower than that required of the one-step method.

III. RECONSTRUCTING 2-SIMPLICIAL COMPLEXES FROM SIMPLICIAL ISING DYNAMICS

We introduce a simplicial Ising model and demonstrate that our statistical inference framework for reconstructing 2-simplicial complexes based on the binary Ising data. The Ising model is fundamental in statistical physics and complex systems, which can be used to describe a variety of phenomena such as phase transitions, coarsening dynamics, and opinion formation.

A. Binary data from simplicial Ising dynamics

Let S_i^t be the state of node i at time t . Each node has two possible states: spin-down ($S_i^t = -1$) or spin-up ($S_i^t = +1$). For a 2-simplicial complex hosting simplicial Ising dynamics, the Hamiltonian is given by

$$H(t) = -J_1 \sum_{(i,j)} S_i^t S_j^t - J_2 \sum_{(i,j,k)} S_i^t S_j^t S_k^t, \quad (\text{S3.1})$$

where J_1 and J_2 are the strengths of two-body and three-body interactions, and (i, j) and (i, j, k) denote the two-body and the three-body connections in the 2-simplicial complex, respectively. The first term in the Hamiltonian characterizes the interaction between the edges (i.e., two-body connections) and the second term contains three-body interactions from the 2-simplex. At each time step, the spin-flipping probability of each node i is given by

$$f_i(t+1) = \frac{1}{1 + e^{\delta \Delta E_i^t}}, \quad (\text{S3.2})$$

in which δ is the inverse temperature. The quantity

$$\Delta E_i^t = 2J_1 \sum_{(i,j) \in \partial_i} S_i^t S_j^t + 2J_2 \sum_{(i,j,k) \in \nabla_i} S_i^t S_j^t S_k^t$$

represents the change in the energy caused by a flipping of node i at time t , where ∂_i and ∇_i are the 1-simplex set and the 2-simplex set containing node i , respectively. At the initial time, the state of each node i is randomly assigned as $+1$ or -1 with equal probability. Binary time series are generated according to Eq. (S3.2) and the data are stored in the data matrix S .

B. Statistical inference framework for simplicial Ising dynamics

Similar to the case of the social contagion dynamics in the main text, we use the two-step reconstruction strategy to reconstruct 2-simplicial complex for simplicial Ising dynamics. At the first step, the “approximate” neighbors of each node are predicted and their corresponding columns in the data matrix S are extracted. At the second step, based on the compressed data matrix, the final neighbors and the 2-simplex of node i are determined. The main processes of step 1 and step 2 are as follows.

Let C and $\bar{C}$ represent the two distinct states of nodes, through which the transition from “+1” to “-1” and that from “-1” to “+1” can be described in a common setting. For the first step, according to the Bayesian formula, the conditional probability of $S_i^{t+1} = \bar{C}$ and $j \rightarrow i$, i.e., the event that node j has a direct impact on the state of node i given $S_i^t = C$ and $S_j^t = \bar{C}$, can be written as

$$\begin{aligned} & P(S_i^{t+1} = \bar{C}, j \rightarrow i | S_i^t = C, S_j^t = \bar{C}) \\ &= P(j \rightarrow i | S_i^t = C, S_j^t = \bar{C}, S_i^{t+1} = \bar{C}) P(S_i^{t+1} = \bar{C} | S_i^t = C, S_j^t = \bar{C}) = P_{j \rightarrow i}^0 P_j^i, \end{aligned} \quad (\text{S3.3})$$

where

$$P_{j \rightarrow i}^0 = P(j \rightarrow i | S_i^t = C, S_j^t = \bar{C}, S_i^{t+1} = \bar{C})$$

denotes the probability that node i changes from C state to $\bar{C}$ state due to the excitation of node j , under the conditions $S_i^t = C$, $S_j^t = \bar{C}$ and $S_i^{t+1} = \bar{C}$. Theoretically, $P_{j \rightarrow i}^0 > 0$ indicates that node j is an “approximate” neighbor of node i ; otherwise, there is no edge connecting nodes i and j . The quantity

$$P_j^i = P(S_i^{t+1} = \bar{C} | S_i^t = C, S_j^t = \bar{C})$$

is the probability of $S_i^{t+1} = \bar{C}$ under the conditions $S_i^t = C$ and $S_j^t = \bar{C}$, which can be estimated from the data matrix S .

From Eq. (S3.3), the expected number of node i in $\bar{C}$ state at $t_m + 1$ can be expressed as

$$\tilde{E}_i^{t_m+1} = \sum_{j(j \neq i)} P(S_i^{t_m+1} = \bar{C}, j \rightarrow i | S_i^{t_m} = C, S_j^{t_m} = \bar{C}) \Psi_j^{t_m} + \varepsilon_i = \sum_{j(j \neq i)} P_{j \rightarrow i}^0 P_j^i \Psi_j^{t_m} + \varepsilon_i, \quad (\text{S3.4})$$

where $\Psi_j^{t_m}$ represents the expected times of node j in $\bar{C}$ state at time t_m , and their values are zero or one. The quantity ε_i represents noise due to the errors from the collected data. Assuming Ψ_i of node i obeys Poisson distribution, we get the likelihood function as

$$P\left(\{\Psi_i^{t_m+1}\}_{m=1, \dots, M} \mid \tilde{\Theta}, \{\Psi_j^{t_m}\}_{m=1, \dots, M; j=1, \dots, N}\right) = \prod_{m(\Psi_i^{t_m}=C)} \frac{e^{-\tilde{E}_i^{t_m+1}} (\tilde{E}_i^{t_m+1})^{\Psi_i^{t_m+1}}}{\Psi_i^{t_m+1}!}, \quad (\text{S3.5})$$

where $\tilde{\Theta}$ denotes the set of variables $P_{j \rightarrow i}^0$ and ε_i . We have $\Psi_i^{t_m+1}! \equiv 1$ since $\Psi_i^{t_m+1}$ is either zero

or one. Taking the logarithm of Eq. (S3.5), we have

$$L(\tilde{\Theta}) = \sum_{m(\Psi_i^{t_m}=C)} \left(\Psi_i^{t_m+1} \log \tilde{E}_i^{t_m+1} - \tilde{E}_i^{t_m+1} \right) = \sum_{m(\Psi_i^{t_m}=C)} \left[\Psi_i^{t_m+1} \log \left(\sum_{j(j \neq i)} P_{j \rightarrow i}^0 P_j^i \Psi_j^{t_m} + \varepsilon_i \right) - \left(\sum_{j(j \neq i)} P_{j \rightarrow i}^0 P_j^i \Psi_j^{t_m} + \varepsilon_i \right) \right]. \quad (\text{S3.6})$$

Using the expectation maximization (EM) method [1] to maximize the likelihood function for determining the parameter $\tilde{\Theta}$ in Eq. (S3.5), we have

$$P_{j \rightarrow i}^0 = \frac{\sum_{m(\Psi_i^{t_m}=C)} (\Psi_i^{t_m+1} \rho_j^{t_m})}{\sum_{m(\Psi_i^{t_m}=C)} (P_j^i \Psi_j^{t_m})}, \quad (\text{S3.7})$$

$$\varepsilon_i = \frac{\sum_{m(\Psi_i^{t_m}=C)} (\Psi_i^{t_m+1} \rho_{\varepsilon_i}^{t_m})}{\sum_{m(\Psi_i^{t_m}=C)} (1)}, \quad (\text{S3.8})$$

where

$$\rho_j^{t_m} = \frac{P_{j \rightarrow i}^0 P_j^i \Psi_j^{t_m}}{\sum_{j'(j' \neq i)} P_{j' \rightarrow i}^0 P_{j'}^i \Psi_{j'}^{t_m} + \varepsilon_i}, \quad (\text{S3.9})$$

$$\rho_{\varepsilon_i}^{t_m} = \frac{\varepsilon_i}{\sum_{j'(j' \neq i)} P_{j' \rightarrow i}^0 P_{j'}^i \Psi_{j'}^{t_m} + \varepsilon_i}. \quad (\text{S3.10})$$

Equations (S3.7)-(S3.10) are the key formulas for finding the unknown quantities $P_{j \rightarrow i}^0$ and ε_i . By initializing $P_{j \rightarrow i}^0$ and ε_i and repeating the above four equations until convergence is achieved and using the same truncation method as in social contagion dynamics, we can determine the “approximate” neighbors of node i .

At the second step, similar to Eq. (S3.3) in the first step, the conditional probability of $S_i^{t+1} = \bar{C}$ and $j \rightarrow i$ given $S_i^t = C$ and $S_j^t = \bar{C}$ can be written as

$$\begin{aligned} & P(S_i^{t+1} = \bar{C}, j \rightarrow i | S_i^t = C, S_j^t = \bar{C}) \\ &= P(j \rightarrow i | S_i^t = C, S_j^t = \bar{C}, S_i^{t+1} = \bar{C}) P(S_i^{t+1} = \bar{C} | S_i^t = C, S_j^t = \bar{C}) = P_{j \rightarrow i} P_j^i, \end{aligned} \quad (\text{S3.11})$$

where

$$P_{j \rightarrow i} = P(j \rightarrow i | S_i^t = C, S_j^t = \bar{C}, S_i^{t+1} = \bar{C})$$

denotes the probability that node i changes from C state to $\bar{C}$ state due to the excitation of node j , under the conditions $S_i^t = C$, $S_j^t = \bar{C}$ and $S_i^{t+1} = \bar{C}$. Theoretically, $P_{j \rightarrow i} > 0$ indicates that node j is a neighbor of node i ; otherwise, there is no edge connecting nodes i and j . The quantity

$$P_j^i = P(S_i^{t+1} = \bar{C} | S_i^t = C, S_j^t = \bar{C})$$

is the probability of $S_i^{t+1} = \bar{C}$ under the conditions $S_i^t = C$ and $S_j^t = \bar{C}$, which can be estimated from the compressed data matrix.

Similarly, the conditional probability of $S_i^{t+1} = \bar{C}$ and $jk \rightarrow i$ (i.e., the event that the synergistic reinforcement effect coming from nodes j and k has a direct impact on the state of node i) given the conditions $S_j^t S_k^t = \bar{C}$ and $S_i^t = C$, can be written as

$$\begin{aligned} & P(S_i^{t+1} = \bar{C}, jk \rightarrow i | S_i^t = C, S_j^t S_k^t = \bar{C}) \\ &= P(jk \rightarrow i | S_i^t = C, S_j^t S_k^t = \bar{C}, S_i^{t+1} = \bar{C}) P(S_i^{t+1} = \bar{C} | S_i^t = C, S_j^t S_k^t = \bar{C}) \quad (\text{S3.12}) \\ &= P_{jk \rightarrow i} P_{jk}^i, \end{aligned}$$

where

$$P_{jk \rightarrow i} = P(jk \rightarrow i | S_i^t = C, S_j^t S_k^t = \bar{C}, S_i^{t+1} = \bar{C})$$

is the probability that node i changes from C state to $\bar{C}$ state because of the synergistic excitation from nodes j and k , under the conditions $S_i^t = C$, $S_j^t S_k^t = \bar{C}$ and $S_i^{t+1} = \bar{C}$. The quantity $P_{jk \rightarrow i}$ being positive indicates that nodes i , j and k form a 2-simplex, otherwise, they do not form a 2-simplex. The probability

$$P_{jk}^i = (S_i^{t+1} = \bar{C} | S_i^t = C, S_j^t S_k^t = \bar{C})$$

can be estimated from the compressed data matrix in a similar way.

According to Eqs. (S3.11) and (S3.12), the expected number of node i in $\bar{C}$ state at $t_m + 1$ is given by

$$\begin{aligned} E_i^{t_m+1} &= \sum_{j(j \neq i)} P(S_i^{t_m+1} = \bar{C}, j \rightarrow i | S_i^{t_m} = C, S_j^{t_m} = \bar{C}) \Psi_j^{t_m} \\ &+ \sum_{j,k(j \neq k \neq i)} P(S_i^{t_m+1} = \bar{C}, jk \rightarrow i | S_i^{t_m} = C, S_j^{t_m} S_k^{t_m} = \bar{C}) \Psi_{jk}^{t_m} + \varepsilon_i \quad (\text{S3.13}) \\ &= \sum_{j(j \neq i)} P_{j \rightarrow i} P_j^i \Psi_j^{t_m} + \sum_{j,k(j \neq k \neq i)} P_{jk \rightarrow i} P_{jk}^i \Psi_{jk}^{t_m} + \varepsilon_i, \end{aligned}$$

where $\Psi_j^{t_m}$ represents the expected times of node j in $\bar{C}$ state at time t_m . Similarly, $\Psi_{jk}^{t_m}$ is the expected times of both nodes j and k being in $\bar{C}$ state at time t_m , and their values are zero or one. Assuming that the number Ψ_i in each time period obeys Poisson distribution, we obtain the likelihood function as

$$P(\{\Psi_i^{t_m+1}\}_{m=1, \dots, M} | \Theta, \{\Psi_j^{t_m}\}_{m=1, \dots, M; j=1, \dots, N}) = \prod_{m(\Psi_i^{t_m}=C)} \frac{e^{-E_i^{t_m+1}} (E_i^{t_m+1})^{\Psi_i^{t_m+1}}}{\Psi_i^{t_m+1}!}, \quad (\text{S3.14})$$

where Θ denotes the set of variables $P_{j \rightarrow i}$, $P_{jk \rightarrow i}$ and ε_i .

Using the EM method to maximize the likelihood function for determining the parameter Θ in

Eq. (S3.14), we obtain

$$P_{j \rightarrow i} = \frac{\sum_{m(\Psi_i^{t_m}=C)} (\Psi_i^{t_m+1} \rho_j^{t_m})}{\sum_{m(\Psi_i^{t_m}=C)} (P_j^i \Psi_j^{t_m})}, \quad (\text{S3.15})$$

$$P_{jk \rightarrow i} = \frac{\sum_{m(\Psi_i^{t_m}=C)} (\Psi_i^{t_m+1} \rho_{jk}^{t_m})}{\sum_{m(\Psi_i^{t_m}=C)} (P_{jk}^i \Psi_{jk}^{t_m})}, \quad (\text{S3.16})$$

$$\varepsilon_i = \frac{\sum_{m(\Psi_i^{t_m}=C)} (\Psi_i^{t_m+1} \rho_{\varepsilon_i}^{t_m})}{\sum_{m(\Psi_i^{t_m}=C)} (1)}, \quad (\text{S3.17})$$

where

$$\rho_j^{t_m} = \frac{P_{j \rightarrow i} P_j^i \Psi_j^{t_m}}{\left(\sum_{j' (j' \neq i)} P_{j' \rightarrow i} P_{j'}^i \Psi_{j'}^{t_m} + \sum_{j', k' (j' \neq k' \neq i)} P_{j' k' \rightarrow i} P_{j' k'}^i \Psi_{j' k'}^{t_m} + \varepsilon_i \right)}, \quad (\text{S3.18})$$

$$\rho_{jk}^{t_m} = \frac{P_{jk \rightarrow i} P_{jk}^i \Psi_{jk}^{t_m}}{\left(\sum_{j' (j' \neq i)} P_{j' \rightarrow i} P_{j'}^i \Psi_{j'}^{t_m} + \sum_{j', k' (j' \neq k' \neq i)} P_{j' k' \rightarrow i} P_{j' k'}^i \Psi_{j' k'}^{t_m} + \varepsilon_i \right)}, \quad (\text{S3.19})$$

$$\rho_{\varepsilon_i}^{t_m} = \frac{\varepsilon_i}{\left(\sum_{j' (j' \neq i)} P_{j' \rightarrow i} P_{j'}^i \Psi_{j'}^{t_m} + \sum_{j', k' (j' \neq k' \neq i)} P_{j' k' \rightarrow i} P_{j' k'}^i \Psi_{j' k'}^{t_m} + \varepsilon_i \right)}. \quad (\text{S3.20})$$

Taken together, the six equations Eqs. (S3.15)-(S3.20) can be used to solve $P_{j \rightarrow i}$, $P_{jk \rightarrow i}$, and ε_i by initializing $P_{j \rightarrow i}$, $P_{jk \rightarrow i}$, ε_i and repeating the above six equations until convergence is achieved.

We remark that, according to the flipping probability given by Eq. (S3.2), each node always has a large or small flipping probability regardless of the neighbors state at each time step, so noise is always present in the time-series data. As a result, the prediction of two-body connections at the second step is truncated according to Eq. (26) in the main text, instead of simply cutting from zero as in the case of social contagion dynamics. For predicting the three-body connections, the same truncation method as in the social contagion dynamics can be used.

C. Representative reconstruction results for simplicial Ising dynamics

Figures S5 and S6 present results on random simplicial complex (ERSC), scale-free simplicial complex (SFSC), and small-world simplicial complex (SWSC) for $N = 100$ and $N = 200$,

respectively. Reconstruction results from four real-world 2-simplicial complexes are shown in Fig. S7.

Similar to the case of reconstructing 2-simplicial complexes from the social contagion model, Figs. S5-S7 reveal that increasing the time-series length T can improve the reconstruction accuracy of two-body and three-body connections. In addition, a small value of k_1 tends to increase the reconstruction accuracies of both types of simplexes. However, different values of k_2 have little effect on the reconstruction accuracies. For the same amount of data, larger networks generally lead to lower accuracies. Finally, the reconstruction accuracy of two-body connections is generally higher than that of the three-body connections.

IV. RECONSTRUCTION PERFORMANCE IN TERMS OF THE NUMBER OF EVENTS

We present F1 score as a function of the number of events for scale-free simplicial complex (SFSC) (Fig. S8) and four real-world 2-simplicial complexes (Fig. S9), where the number of events is calculated on each node. For example, 100 events means that each node has 100 switches between susceptible and infected states. It can be seen that the reconstruction accuracy increases with the number of events. In addition, the reconstruction accuracies of two-body and three-body connections decrease with k_1 (i.e., average degree of 1-simplex), but the value of k_2 (i.e., average degree of 2-simplex) affects only the accuracy of three-body connections and has little effect on the accuracy of reconstructing two-body connections. While a larger simplicial complex requires more events to reach certain F1 score, with sufficient data a high reconstruction accuracy can still be achieved. The conclusion is that the results based on the number of events are essentially the same as those in terms of the simulation time T .

V. O-INFORMATION MEASURE

In the recent work [2], an important information metric named O-information (Ω) was introduced, which can be used to characterize synergy- and redundancy-dominated systems and quantify higher-order interdependencies. It can be seen from Lemma 1 in the paper that O-information can only capture the interactions that go beyond the pairwise relationships, so it is not suitable for describing the interaction of two variables, i.e., it cannot be used to reconstruct the pairwise relationships.

More Specifically, for a system of three discrete variables, the O-information is defined as

$$\Omega(X^n) = H(X^n) + \sum_{j=1}^n [H(X_j) - H(X_{-j}^n)],$$

where

$$\begin{aligned} H(X^n) &= - \sum_{X^n} P_{X^n}(X^n) \log P_{X^n}(X^n), \\ X^n &= (X_1, \dots, X_n), \\ X_{-j}^n &= (X_1, \dots, X_{j-1}, X_{j+1}, \dots, X_n), \end{aligned}$$

for $n = 3$. According to this metric, the system is redundancy dominated if $\Omega(X^3) > 0$; otherwise ($\Omega(X^3) < 0$), it is synergy dominated. To compare our method with this O-information based

method, we have calculated the O-information values between any three variables. If it is negative, there is an interaction among the three variables (three-body connection), otherwise the connection does not exist. The results are shown in Fig. S10, where the O-information values of any three points are displayed with the blue and red dots denoting the existent and nonexistent three-body connections, respectively. It can be seen that the O-information values associated with most of the existent three-body interactions are indeed negative, but the O-information values of many nonexistent three-body interactions are also negative, making it impossible to distinguish the two cases and to ascertain the existent three-body interactions.

Overall, the O-information method proposed in [PRE 100, 032305 (2019)] relies on strong correlations among the time series of the dynamical variables for predicting the synergy structure. In our case, the dynamical time-series data are obtained by alternating updated iterations, that is, the state transition of each node is determined by the states of neighbors at the previous time, leading to only weak correlations among different nodal pairs at any given time. As a result, the O-information method is not suitable for predicting higher-order structures from dynamical time-series data.

VI. EFFECT OF TIME-SERIES LENGTH ON RECONSTRUCTION PERFORMANCE

The reconstruction performance on synthetic and real-world 2-simplicial complexes versus the length T of the time series has been studied for different noise level (the flip ratio) f : $f = 0$ (without noise), 0.1, 0.2, and 0.3, with results shown in Figs. S11 and S12, respectively. These results indicate that, for short time series, the reconstruction performance is sensitive to noise. For example, when the value of f changes from 0 to 0.1, there is a sizable reduction in the F1 score. Regardless of the noise level, increasing the length of time series can always improve the reconstruction performance for both two-body and three-body interactions.

VII. SUPPLEMENTARY FIGURES

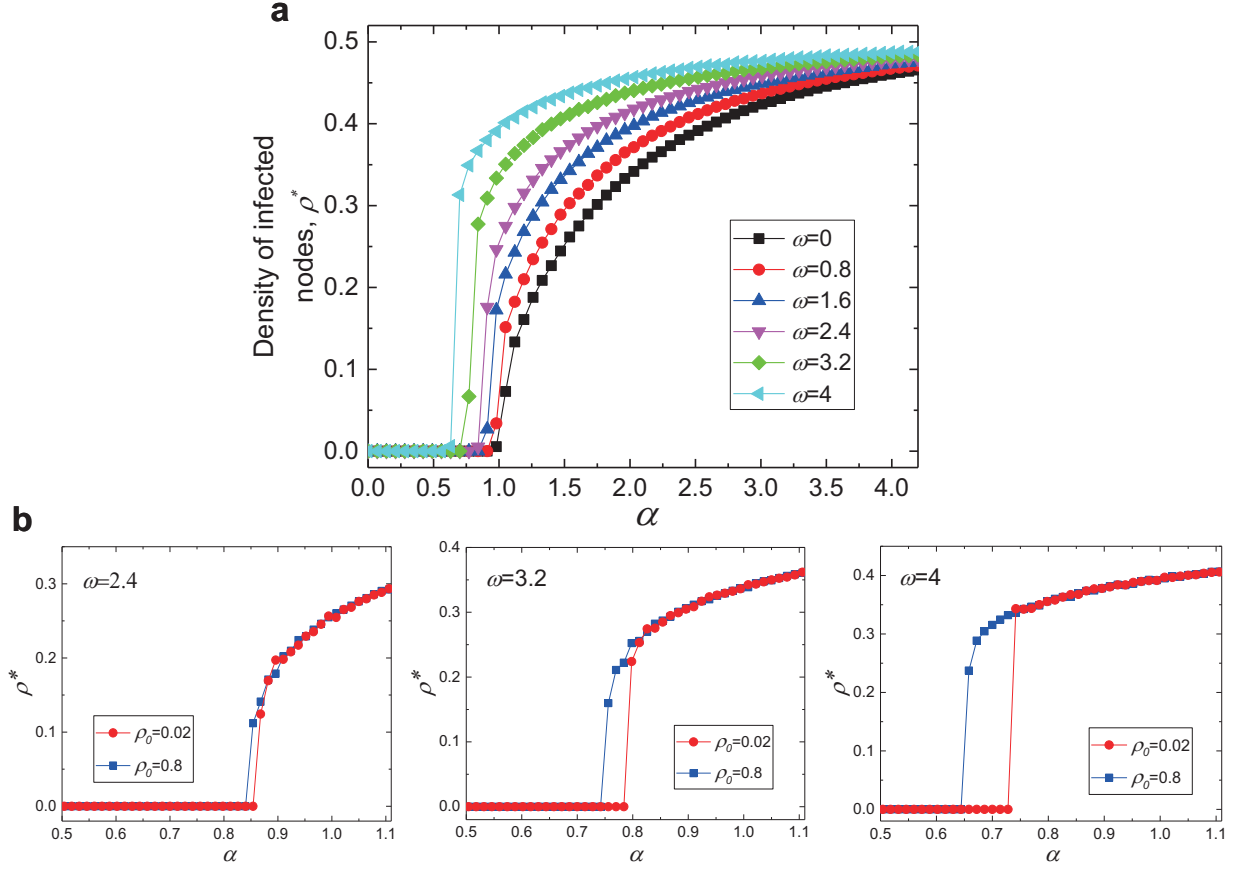


FIG. S1. Average fraction ρ^* of infected nodes in the stationary state as a function of the rescaled edge infectivity α for SFSC. **(a)** The resulting ρ^* -vs- α curves for different values of the rescaled triangular infectivity ω for $\rho_0 = 0.2$ and $\mu = 1$. As ω increases, the nature of the phase transition in the underlying social contagion dynamics changes from continuous to discontinuous. **(b)** The effect of the initial density of the infected nodes on the ρ^* -vs- α curve for three values ω (three subpanels). In each subpanel, the ρ^* -vs- α curves for two values of ρ_0 are shown: $\rho_0 = 0.02$ (red circles) and $\rho_0 = 0.8$ (blue squares). Different initial values of the density ρ_0 of the infected nodes can affect the steady-state infection density ρ^* associated with the healthy and endemic states in the bistable region. For all the curves in **(a)** and **(b)**, each data point is the result of averaging over 50 statistical realizations.

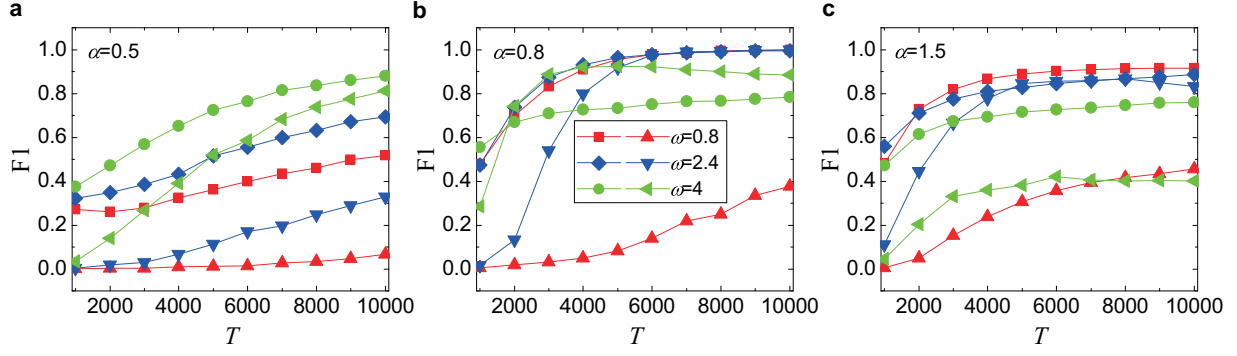


FIG. S2. Reconstruction performance for different values of α and ω . Shown is F1 score as a function of the time series length T for SFSC for (a) $\alpha = 0.5$, (b) $\alpha = 0.8$, and (c) $\alpha = 1.5$. In each panel, squares, diamonds and circles denote the accuracy of reconstructing two-body connections, while triangles with different orientations denote the accuracy of reconstructing three-body connections. The results from different values of ω are distinguished by colors. Other parameter values are $\rho_0 = 0.2$ and $\mu = 1$, and five realizations are used to generate the results.

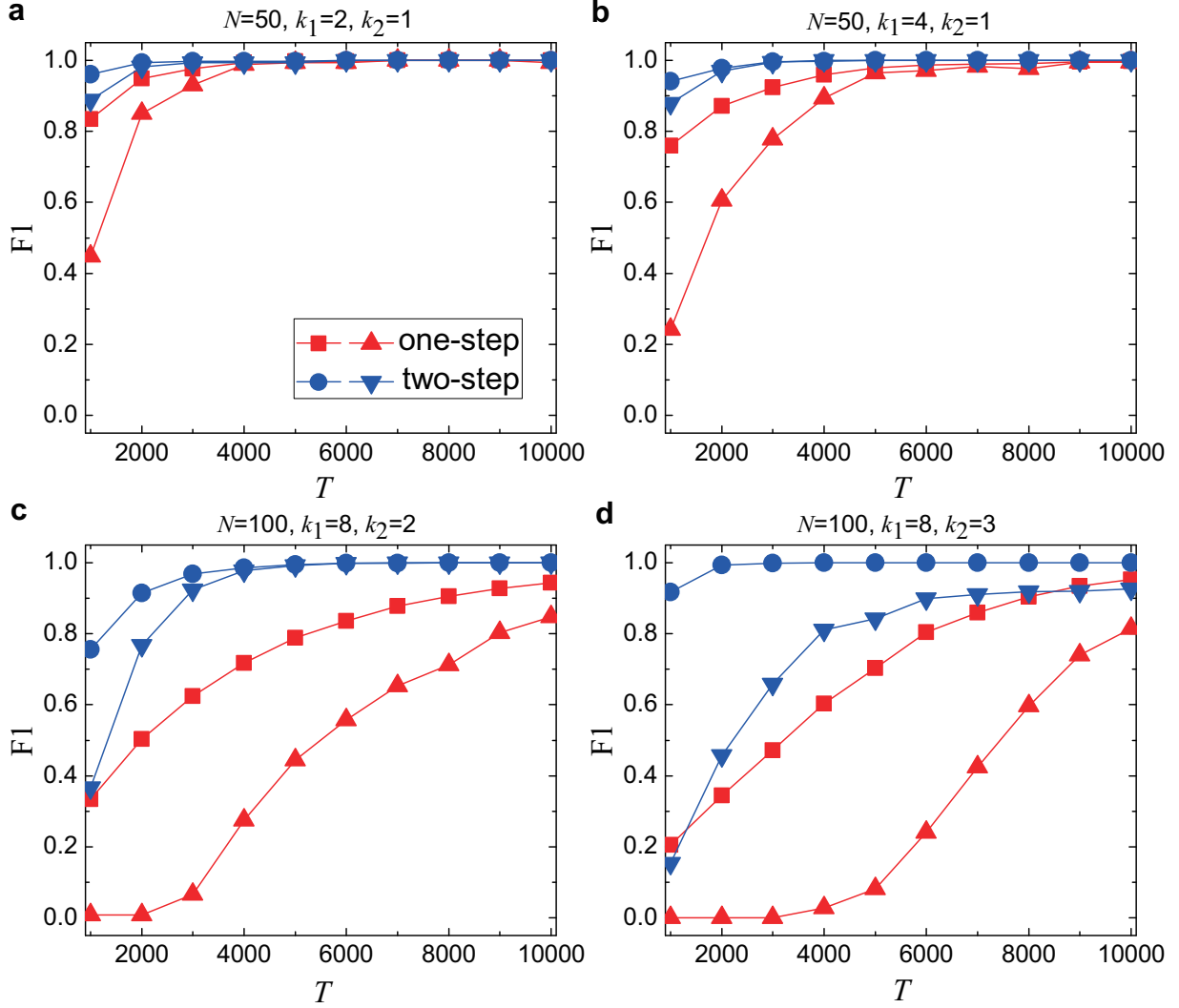


FIG. S3. Comparison of the reconstruction performance between one-step and two-step methods in SFSC with different sizes and degrees. Shown is F1 score as a function of the time-series length T . Squares and circles denote the performance of reconstructing two-body connections while triangles with different orientations are for reconstructing three-body connections. The results from the two methods are distinguished by colors. The parameter values in each simplicial complex are (a) $N = 50, k_1=2, k_2=1, \alpha=0.2, \omega=1$, (b) $N=50, k_1=4, k_2=1, \alpha=0.4, \omega=1$, (c) $N=100, k_1=8, k_2=2, \alpha=0.6, \omega=1.55$, and (d) $N=100, k_1=8, k_2=3, \alpha=0.8, \omega=1.6$. Other parameter values are $\rho_0 = 0.2$ and $\mu = 1$. The results are averaged over five realizations.

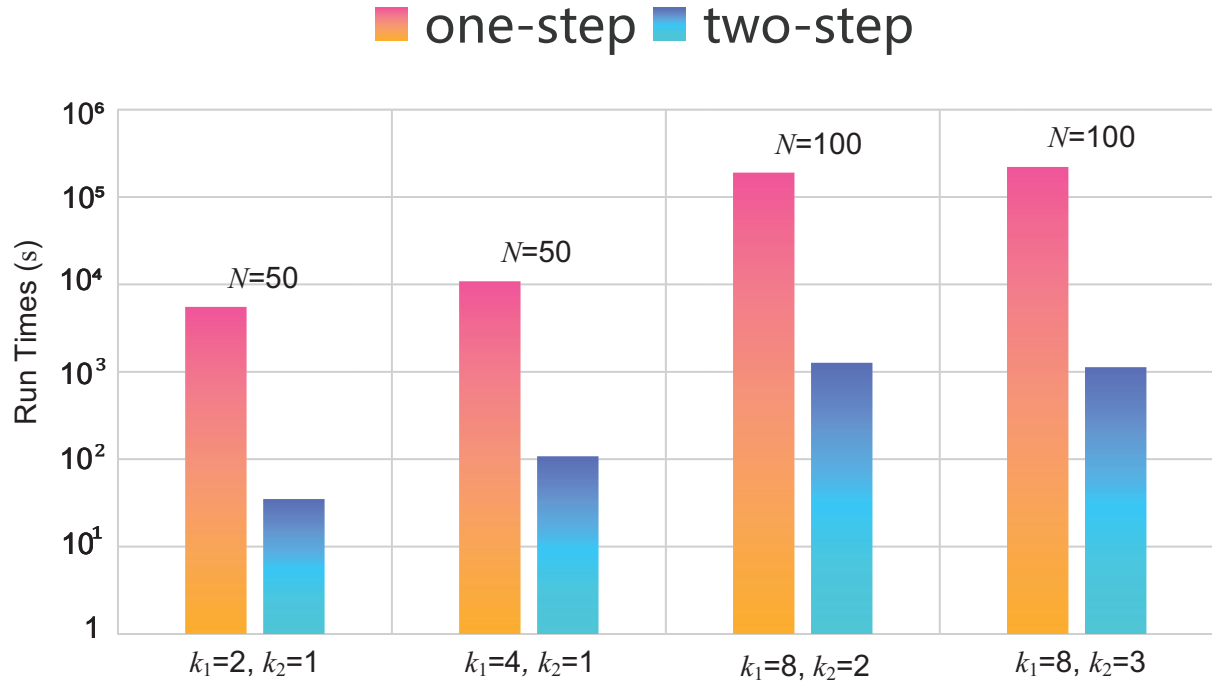


FIG. S4. Comparison of the required computational time of the one-step method and two-step method on SFSC with different sizes and degrees. The parameter values are the same as those in Fig. S3. The running time is calculated based on one realization.

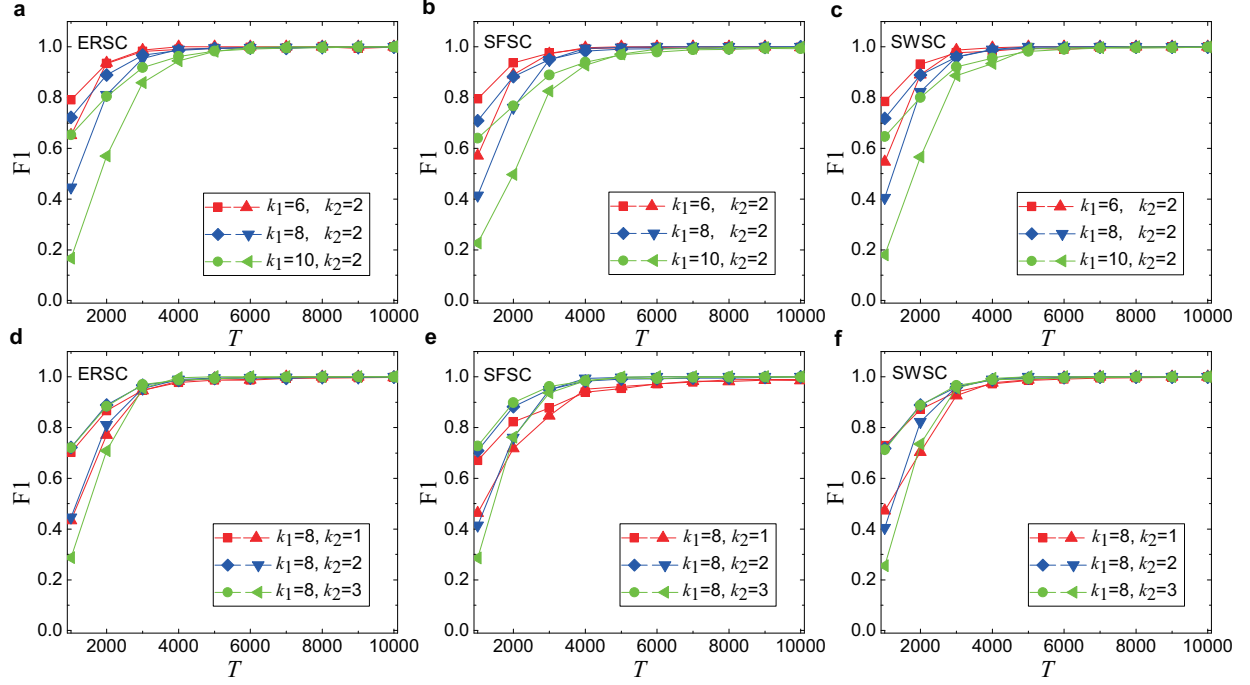


FIG. S5. Reconstruction performance for synthetic 2-simplicial complexes. Shown is F1 score as a function of the length T for three synthetic 2-simplicial complexes with size $N = 100$: ERSC - left column, SFSC - middle column, and SWSC - right column. In each panel, squares, diamonds and circles indicate the performance of reconstructing two-body connections while triangles with different orientations denote the performance of reconstructing three-body connections. Different values of the average degree are distinguished by colors. For $(k_1 = 6, k_2 = 2)$, $(k_1 = 8, k_2 = 2)$, $(k_1 = 10, k_2 = 2)$, $(k_1 = 8, k_2 = 1)$, and $(k_1 = 8, k_2 = 3)$, the parameter values are $(J_1, J_2) = (0.7, 1.2)$, $(0.8, 1.2)$, $(0.86, 1.2)$, $(0.8, 2.4)$ and $(0.8, 0.8)$, respectively. Other parameter values are $\rho_0 = 0.5$ and $\delta = 1/k_1$. Each data point is the result of averaging over five realizations.

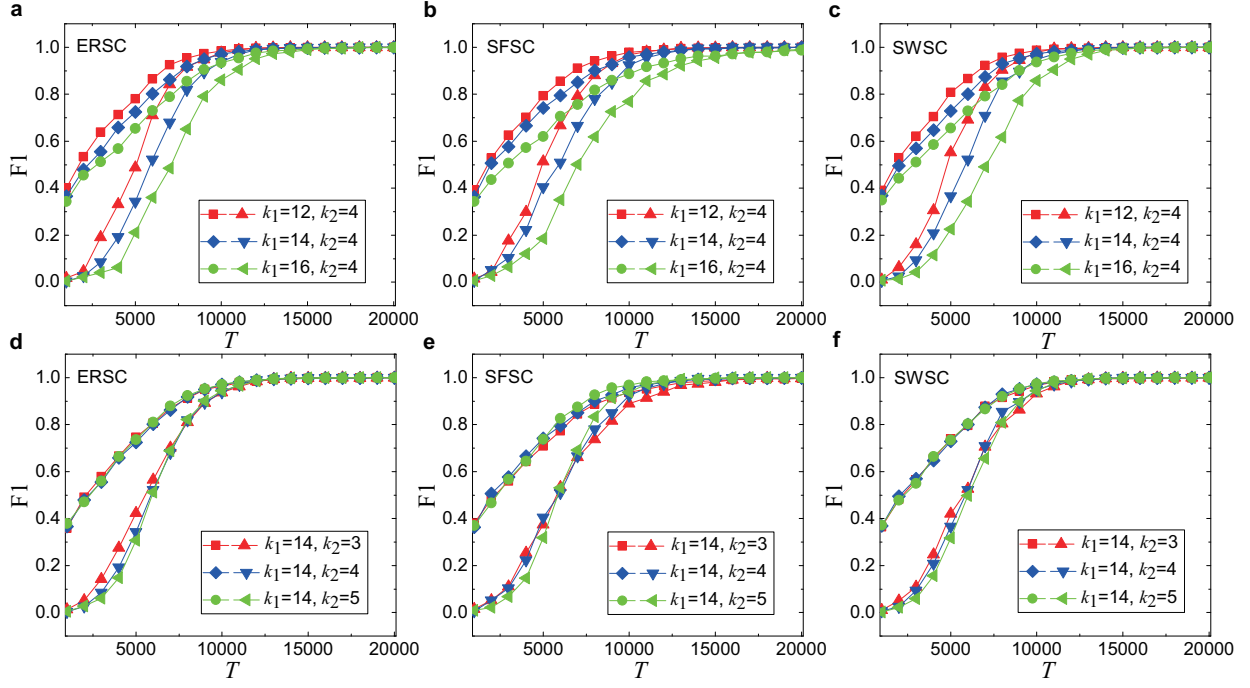


FIG. S6. Reconstruction performance for synthetic 2-simplicial complexes. The legends are the same as those in Fig. S5 except that the network size is $N = 200$. For $(k_1 = 12, k_2 = 4)$, $(k_1 = 14, k_2 = 4)$, $(k_1 = 16, k_2 = 4)$, $(k_1 = 14, k_2 = 3)$, and $(k_1 = 14, k_2 = 5)$, the parameter values are $(J_1, J_2) = (0.7, 1.2)$, $(0.76, 1.2)$, $(0.8, 1.2)$, $(0.76, 1.6)$ and $(0.76, 0.96)$, respectively. Other parameter values are $\rho_0 = 0.5$ and $\delta = 1/k_1$.

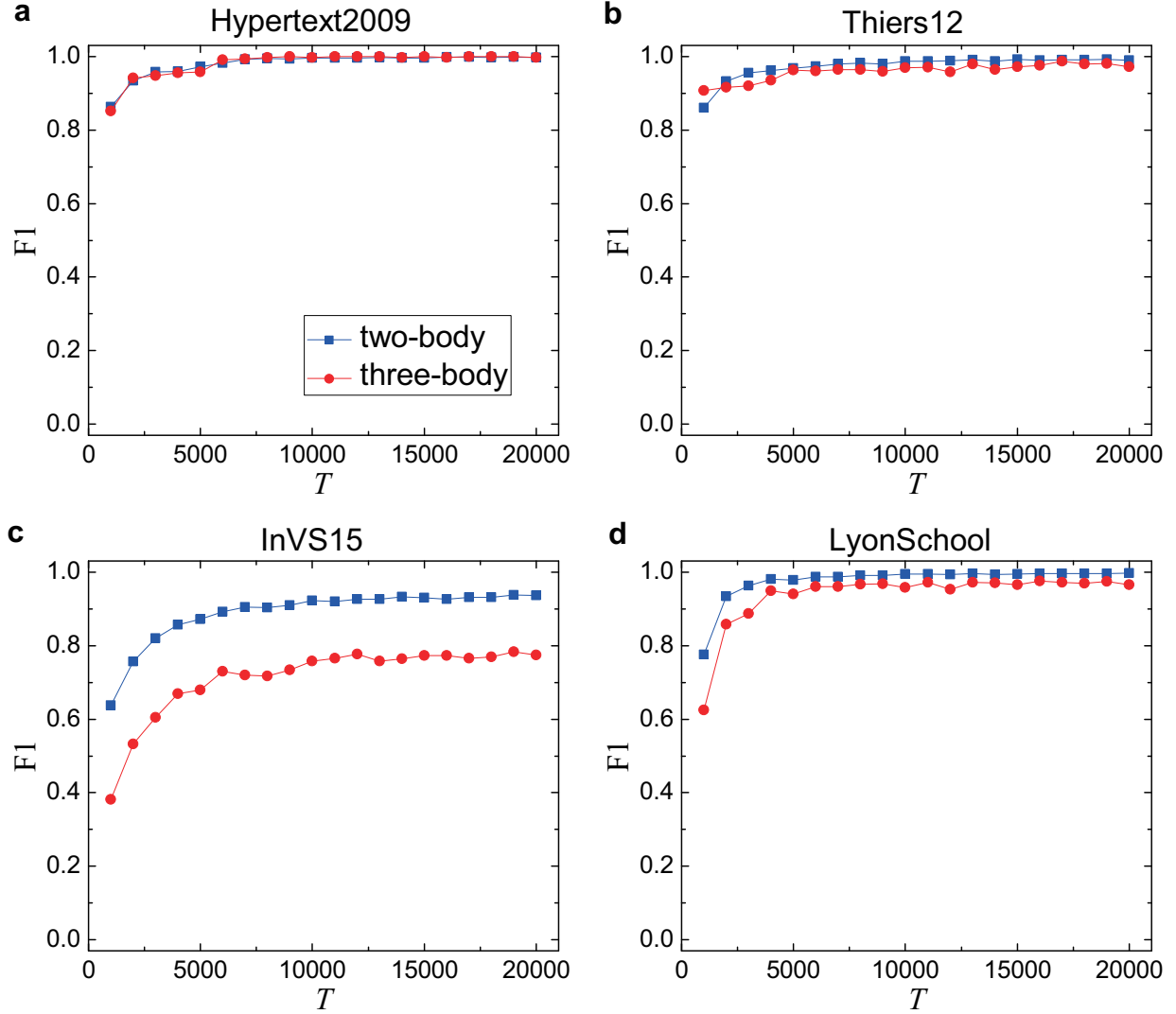


FIG. S7. Reconstruction performance for real-world 2-simplicial complexes. Shown is F1 score versus the time-series length T . The blue squares and red circles indicate the reconstruction performance of two-body and three-body connections, respectively. The parameter values are **(a)** $J_1 = 0.78$, $J_2 = 1.2$, **(b)** $J_1 = 0.8$, $J_2 = 1.2$, **(c)** $J_1 = 0.78$, $J_2 = 1.1$, **(d)** $J_1 = 0.66$, $J_2 = 1.1$. Other parameter values are $\rho_0 = 0.5$ and $\delta = 1/k_1$. Each data point is the result of averaging over five realizations.

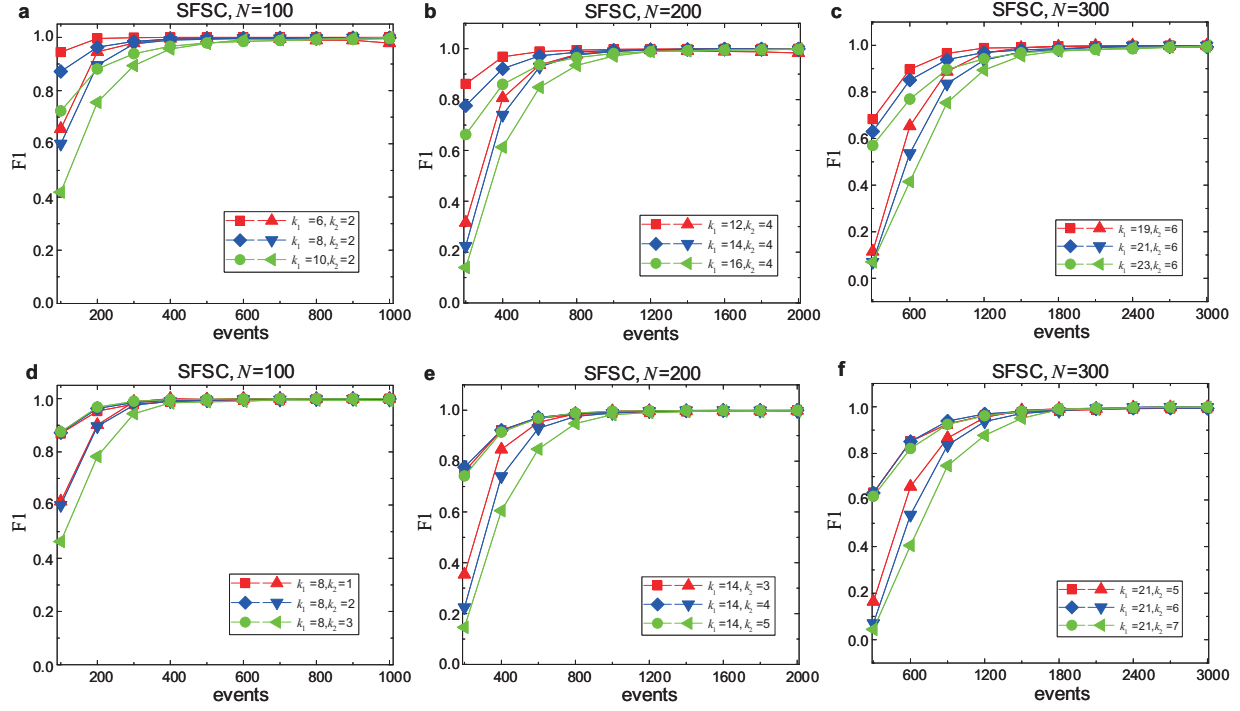


FIG. S8. F1 score as a function of the number of events for SFSC for (a,d) $N = 100$, $\alpha = 0.6$, $\omega = 2$, (b,e) $N = 200$, $\alpha = 0.7$, $\omega = 2.2$, and (c,f) $N = 300$, $\alpha = 0.75$, $\omega = 2.3$. In each panel, squares, diamonds and circles demonstrate the performance of reconstructing two-body connections, and triangles with different orientations depict the performance of reconstructing three-body connections. The average degrees are distinguished by colors. Other parameter values are $\rho_0 = 0.2$ and $\mu = 1$. Each data point is the result of averaging over five realizations.

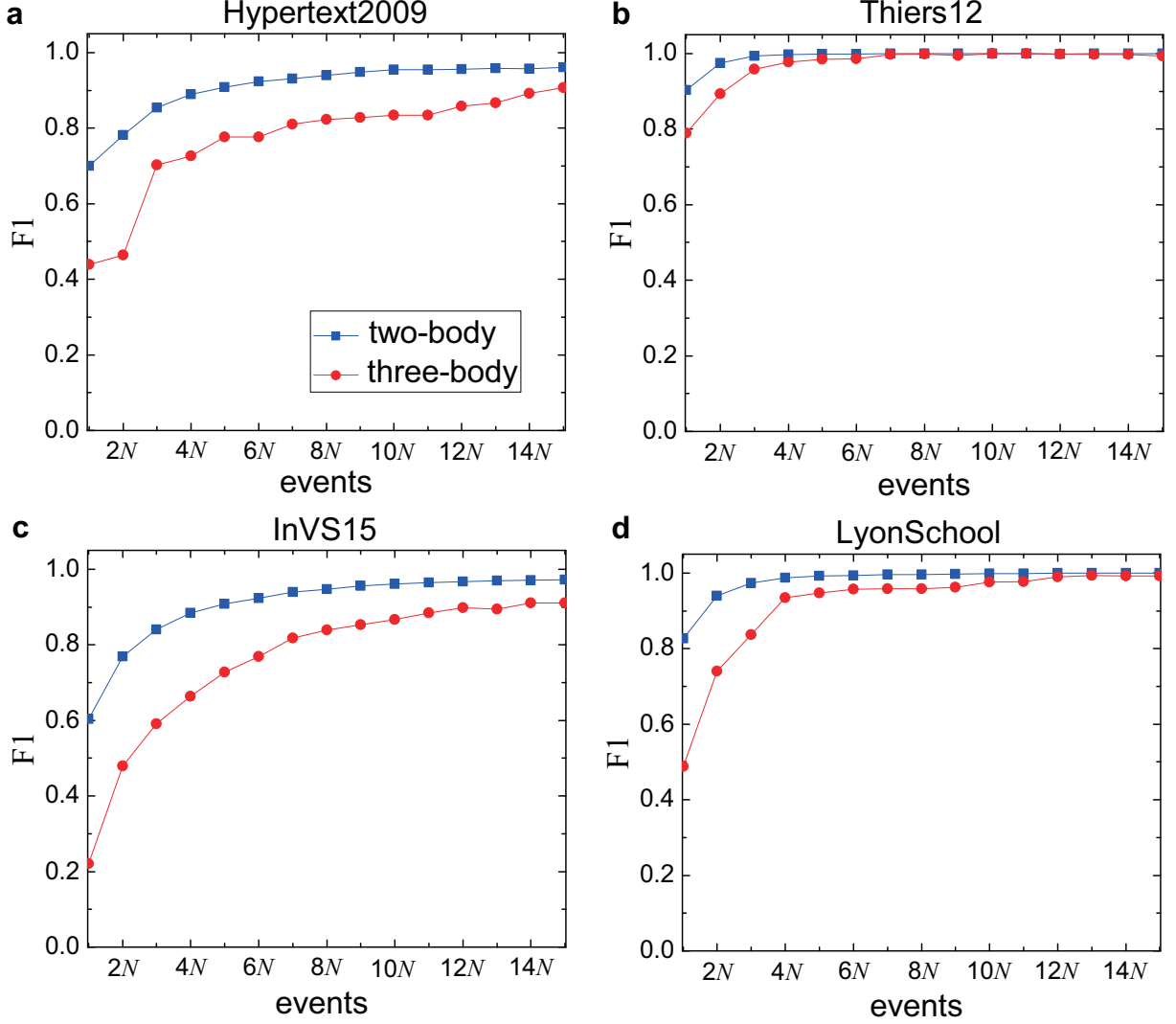


FIG. S9. F1 score as a function of the number of events for four real-world 2-simplicial complexes: (a) Hypertext2009, (b) Thiers12, (c) InVS15, (d) LyonSchool. In each panel, the blue squares and red circles demonstrate the performance of reconstructing two-body and three-body connections, respectively. Parameter values are $\alpha = 0.3$, $\omega = 1$, $\rho_0 = 0.2$ and $\mu = 1$. The “ N ” appeared in the abscissa denotes the size of the specific network. Each data point is the result of averaging over five realizations.

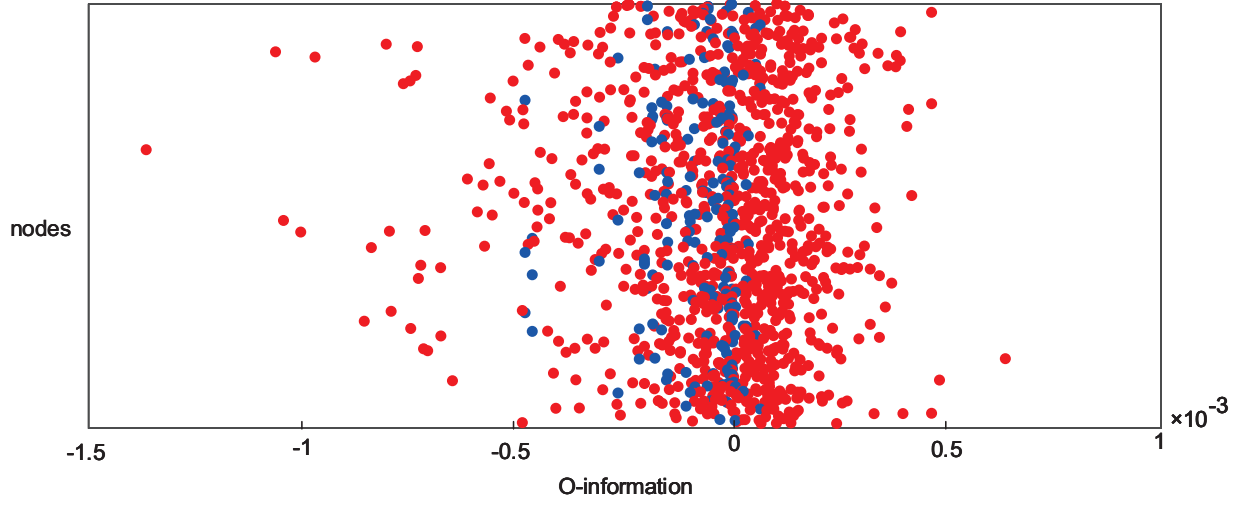


FIG. S10. Performance of O-information based method. Shown are the O-information values of any three points $\Omega(X^3)$ on ERSC. Each row corresponds to the O-information of one node. The blue and red dots denote the existent and nonexistent three-body connections, respectively. The parameter values are $N = 100$, $T = 10000$, $k_1 = 7$, $k_2 = 2$, $\alpha = 0.6$, $\omega = 2.1$, $\rho_0 = 0.2$ and $\mu = 1$. For each node, the number of all possible three-body interactions is about N^2 . Because the number of existent three-body connections is very low, only 0.2% of the red dots are shown in each row for better visualization.

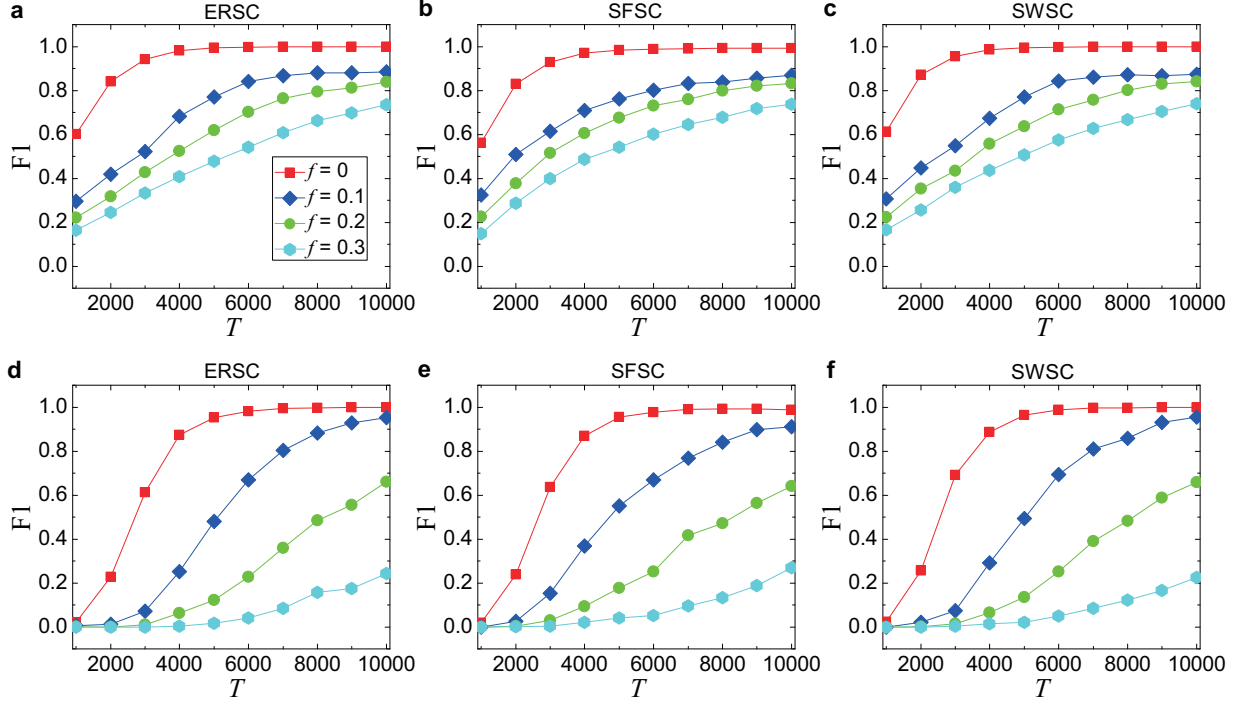


FIG. S11. Reconstruction performance under noise for synthetic 2-simplicial complexes. Shown is F1 score for different values of the flip ratio f as a function of the time-series length T for three synthetic 2-simplicial complexes: random simplicial complex (ERSC - left column), scale-free simplicial complex (SFSC - middle column), and small-world simplicial complex (SWSC - right column): **(a-c)** reconstructing two-body connections and **(d-f)** reconstructing three-body interactions. The results for different values of f are distinguished by symbols and colors. All simplicial complexes have the same size $N = 200$. Other parameter values are $k_1 = 12$, $k_2 = 4$, $\alpha = 0.8$, $\omega = 2.4$, $\rho_0 = 0.2$, and $\mu = 1$. Each data point is the result of averaging over five realizations.

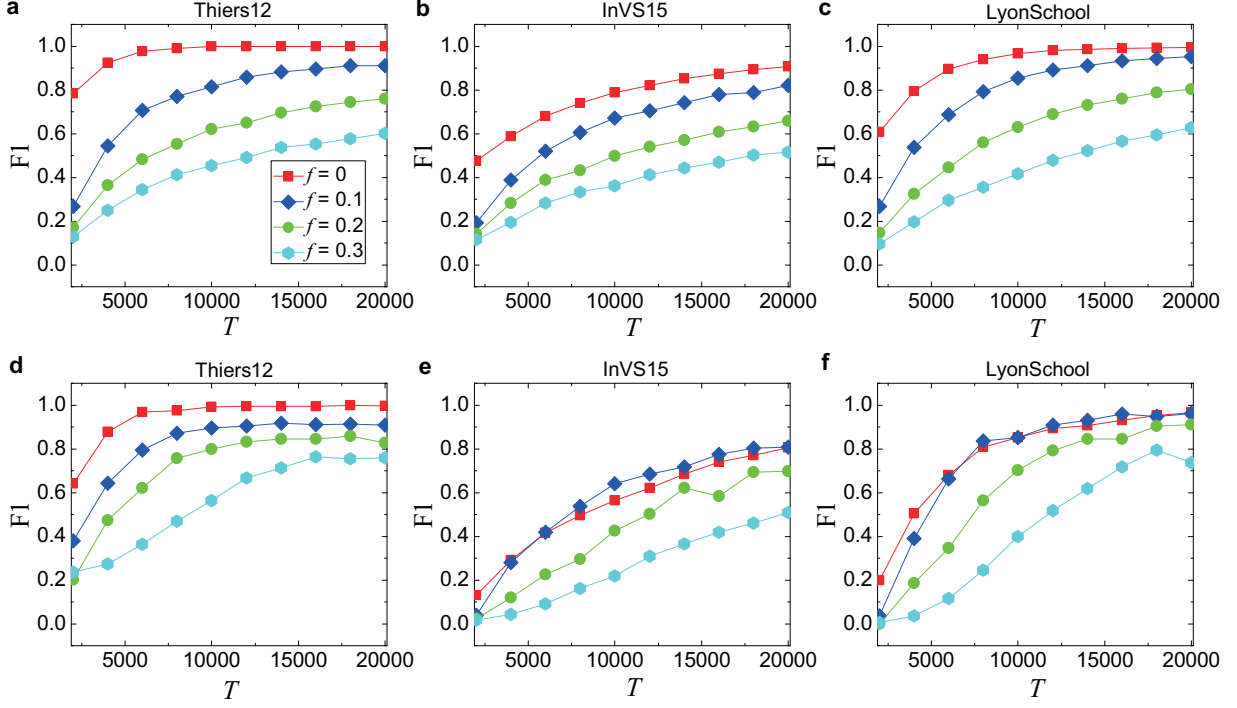


FIG. S12. Reconstruction performance under noise for real-world 2-simplicial complexes. Shown is F1 score for different values of the flip ratio f as a function of the time-series length T for **(a,d)** Thiers12, **(b,e)** InVS15, and **(c,f)** LyonSchool: **(a-c)** reconstructing two-body connections and **(d-f)** reconstructing three-body interactions. Parameter values are $\alpha = 0.3$, $\omega = 1$, $\rho_0 = 0.2$, and $\mu = 1$. Each data point is the result of averaging over five realizations.

VIII. SUPPLEMENTARY REFERENCES

- [1] Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B Method.* **39**, 1–22 (1977).
- [2] Rosas, F. E., Mediano, P. A. M., Gastpar, M. & Jensen, H. J. Quantifying high-order interdependencies via multivariate extensions of the mutual information. *Phys. Rev. E* **100**, 032305 (2019).