




## Hypergraph reconstruction from network data

Jean-Gabriel Young <sup>1,2,3✉</sup>, Giovanni Petri <sup>4</sup> & Tiago P. Peixoto <sup>4,5,6</sup>

Networks can describe the structure of a wide variety of complex systems by specifying which pairs of entities in the system are connected. While such pairwise representations are flexible, they are not necessarily appropriate when the fundamental interactions involve more than two entities at the same time. Pairwise representations nonetheless remain ubiquitous, because higher-order interactions are often not recorded explicitly in network data. Here, we introduce a Bayesian approach to reconstruct latent higher-order interactions from ordinary pairwise network data. Our method is based on the principle of parsimony and only includes higher-order structures when there is sufficient statistical evidence for them. We demonstrate its applicability to a wide range of datasets, both synthetic and empirical.

---

<sup>1</sup>Center for the Study of Complex Systems, University of Michigan, Ann Arbor, MI, USA. <sup>2</sup>Department of Computer Science, University of Vermont, Burlington, VT, USA. <sup>3</sup>Vermont Complex Systems Center, University of Vermont, Burlington, VT, USA. <sup>4</sup>ISI Foundation, Torino, Italy. <sup>5</sup>Department of Network and Data Science, Central European University, Vienna, Austria. <sup>6</sup>Department of Mathematical Sciences, University of Bath, Bath, UK. ✉email: [jean-gabriel.young@uvm.edu](mailto:jean-gabriel.young@uvm.edu)

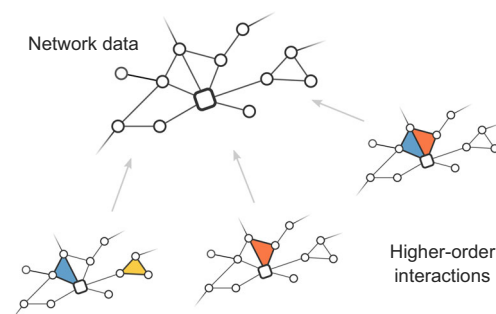
Empirical networks are often locally dense and globally sparse<sup>1</sup>. Whether they are social, biological, or technological<sup>2</sup>, they comprise large groups of densely interconnected nodes, even when only a small fraction of all possible connections exist. This situation leads to delicate modeling challenges: How can we account for two seemingly contradictory properties of networks—density and sparsity—in our models?

Abundant prior work going back to the early days of social network analysis<sup>3,4</sup> and network science<sup>5,6</sup> suggests that higher-order interactions<sup>7</sup> are a possible explanation for the local density of networks<sup>1,6</sup>. According to this reasoning, entities are connected because they have a shared context—a higher-order interaction—within which connections can be created<sup>8</sup>. It is clear that a phenomenon along these lines occurs in many social processes: scientists appear as collaborators in the Web of Science because they co-author papers together; colleagues exchange emails because they are part of the same department or the same division of a company. It is also known that similar phenomena explain tie formation in a broader range of networked systems, including biological, technological, or informational systems<sup>7</sup>.

The ubiquity of higher-order interactions provides a simple and universal explanation for the observed structure of empirical networks. If we assume that most ties are created within contexts of limited scopes, then the resulting networks are locally dense, matching empirical observations<sup>1,9,10</sup>.

Despite their tremendous explanatory power, higher-order interactions are seldom used directly to model empirical systems, due to a lack of data<sup>7</sup>. Indeed, while the context is directly observable for some systems—say, co-authored papers or co-locating species—it is unavailable for several others, including brain data<sup>11</sup>, typical social interaction data<sup>12</sup>, and ecological competitor data<sup>13</sup> to name only a few.

As a specific motivating example, consider one of the empirical social networks gathered as part of the US National Longitudinal Study of Adolescent to Adult Health<sup>14</sup>. This dataset is constructed using surveys, where participants are asked to nominate their friends. Even though there are good reasons to believe that people often interact because of higher-order groups<sup>12</sup>, the survey cannot reveal these groups as it only inquires about pairwise relationships. If we actually need the higher-order interactions to give an appropriate description of the social dynamics at play<sup>12</sup>, what should we do with such inadequate survey data? As we show in Fig. 1, there are many kinds of higher-order interactions that are compatible with the same network data. How can one pick among all these possible higher-order descriptions?



**Fig. 1 Projected higher-order interactions.** We show the extended ego network (circle nodes) of a participant (square node) of the AddHealth study<sup>14</sup>. Friendships are measured between pairs of participants (links and nodes, respectively), even when the fundamental units are groups of friends<sup>12</sup>. Multiple combinations of groups and isolated friendships lead to the same network (gray arrows).

Prior work on higher-order interaction discovery in network data often uses cliques—fully connected subgraphs—to identify the interactions<sup>15–17</sup>. Clique-based methods are straightforward to implement because they rely on clique enumeration, a classical problem for which we have exact<sup>18,19</sup> and sampling<sup>20</sup> algorithms that work well in practice. However, clique decompositions do not offer a satisfactory solution to the recovery problem alone. Networks typically admit many possible clique decompositions, which begs the question of which one to pick. For example, a triangle can be decomposed as a single 2-clique, or as three 1-clique (i.e., as edges) (see Fig. 1). In general, the multiplicity of possible solutions implies that higher-order interaction recovery is an ill-posed inverse problem. It becomes well-posed only once we add further constraints on what constitutes a good solution. Thus, existing approaches have sought to address the ill-posed nature of the higher-order interaction recovery problem in various indirect ways. For instance, in graph theory, it is customary to look for a minimal set of cliques covering the network<sup>21,22</sup>. Other methods appeal to notions of randomness and generative modeling to regularize the problem<sup>1,23–25</sup>. These methods describe an explicit process by which one goes from higher-order data to networks, and can therefore assign a likelihood to possible higher-order data representations, allowing the user to single out representations.

In the present work, we develop a Bayesian method for the inference of higher-order of interactions from the network. Given a network as input, the method identifies the parts of the network best explained by latent higher-order interactions. Our approach is based on the principle of parsimony and directly addresses the ill-posedness of the reconstruction problem with the methods of information theory. We show that the method can find compact descriptions of many empirical networked systems by using latent higher-order interactions, thereby demonstrating that such interactions are in complex systems.

## Results and discussion

**Generative model.** The problem we solve is illustrated in Fig. 1. We have a system we believe is best described with higher-order interactions, but we can only view its structure through the lens of pairwise measurements (an undirected and simple network  $G$ ); our goal is to reconstruct these higher-order interactions from  $G$  only.

For convenience, we encode the higher-order interactions with a hypergraph  $H$ <sup>26</sup>. We represent a higher-order interaction between a set of  $k$  nodes  $i_1, \dots, i_k$  with a hyperedge of size  $k$ . Empirical data often contain repeated interactions between the same group of nodes, so we use hypergraphs with repeated hyperedges and encode the number of hyperedges connecting nodes  $i_1, \dots, i_k$  as  $A_{i_1, \dots, i_k} \geq 0$ .

Our method then makes use of a Bayesian generative model to deduce one such hypergraph  $H$  from some network dataset  $G$ . This generative model gives an explicit description of how the network data  $G$  is generated when there are latent higher-order interactions  $H$ . With a generative model in place, we can compute the posterior probability

$$P(H|G) = \frac{P(G|H)P(H)}{P(G)} \quad (1)$$

that the latent hypergraph is  $H$ , given the observed network  $G$ . In this equation,  $P(G|H)$  and  $P(H)$  define our generative model for the data, and its evidence  $P(G) = \sum_H P(G|H)P(H)$  functions as a normalization constant.

The appeal of such a Bayesian generative formulation is that we can use  $P(H|G)$  to make queries about the hypergraph  $H$ . What was the most likely set of higher-order interactions? What is the

probability that a particular interaction was present in  $H$  based on  $G$ ? How large were the latent higher-order interactions? All of the queries can be answered by computing appropriate averages over  $P(H|G)$ . As is made evident by Eq. (1), however, we first have to introduce two probability distributions so that we may compute  $P(H|G)$  at all. We now define these distributions in detail.

**Projection component.** The first distribution,  $P(G|H)$ , is called the projection component of the model. It tells us how likely a particular network  $G$  is when the latent hypergraph  $H$  is known.

We use a direct projection component and deem two nodes connected in  $G$  if and only if these nodes jointly appear in any of the hyperedges of  $H$ .

This modeling choice is broadly applicable. For instance, when researchers measure the functional connectivity of two brain regions, they record a connection irrespective of whether the regions peaked as a pair or as jointly with many other regions. Likewise, surveyed social networks contain records of friendships that can be attributed to interactions between pairs of individuals, and to interactions that arise from larger groups.

Certain authors use more nuanced projection components<sup>1,27</sup> and do not assume that the joint participation of two nodes in a hyperedge necessarily leads to a measured pairwise interaction (for example, when edges are omitted at random). Doing so blurs the line between community detection and higher-order interaction reconstruction, because there is little difference between noisily measured cliques and communities. Hence, we here treat measurement as a separate issue<sup>28,29</sup>, and assume that the network is reliable.

We formalize the projection component as follows. We set  $P(G|H) = 1$  only when (i) each pair of nodes connected by an edge in  $G$  appears jointly in at least one hyperedge of  $H$ , and (ii) no two disconnected nodes of  $G$  appear together in a hyperedge of  $H$ . If either of these conditions is violated, then we set  $P(G|H) = 0$ . We can express this definition mathematically as

$$P(G|H) = \begin{cases} 1 & \text{if } G = \mathcal{G}(H), \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

where we use  $\mathcal{G}(H)$  to denote the projection of  $H$  and use  $G = \mathcal{G}(H)$  to say that  $H$  projects to  $G$ , or equivalently that (i) and (ii) hold.

Testing  $G \stackrel{?}{=} \mathcal{G}(H)$  might appear unwieldy at first but, thankfully, a factor graph encoding of  $H$  can help us compute the projection component efficiently by highlighting existing relationships between the edges and cliques of  $G$ <sup>30</sup>.

To construct this factor graph, we begin by creating two separate sets of nodes: one representing the edges of  $G$  and the other representing the cliques of  $G$ . Crucially, the second set contains a node for every clique of  $G$ , even the included ones like the edges of a triangle, the triangles of a 4-clique, and so on. We call this set the set of factors and refer to nodes in the first set simply as nodes. We obtain a factor graph, by connecting a factor and a node when the corresponding clique contains the corresponding edge.

This construction is illustrated in Fig. 2 for a simple graph of five nodes. In Fig. 2, we see that, for example, the edge between nodes 1 and 2 is part of the triangle  $\{1, 2, 3\}$  in  $G$ , and it is therefore connected to the factor  $A_{123}$ . This edge is also part of the 2-clique  $\{1, 2\}$ , so it is connected to the factor  $A_{12}$ , too.

The resulting factor graphs can encode particular hypergraphs  $H$  by assigning integers to the factors, corresponding to the number of times every hyperedge appears in  $H$ . For example, by setting  $A_{123} = 1$  and  $A_{23} = A_{24} = A_{34} = A_{45} = 1$ , we can encode a hypergraph with five hyperedges, one of size 3 and four of size 2

(see Fig. 2b). We obtain a simple graph representation of the same data by setting  $A_{123} = 0$  and  $A_{12} = 1$  instead (see Fig. 2a).

It is straightforward to check whether  $G = \mathcal{G}(H)$  holds with this encoding. The first condition—all the connected nodes of  $G$  are connected by at least one hyperedge in  $H$ —can be verified by checking that every node of the factor graph is connected to at least one active factor, defined as  $A_{i_1, \dots, i_k} > 0$ . The second condition—no pairs of disconnected nodes in  $G$  are connected by a hyperedge of  $H$ —is always satisfied by construction, because no factor connects two disconnected nodes of  $G$ , so we never represent these forbidden hyperedges with our factor graph.

We note that the factor graph can be stored relatively efficiently, by first enumerating the maximal cliques—cliques not included in larger cliques—and then constructing an associative array indexed by cliques, which we expand only when included cliques are needed. Even though enumerating maximal clique is technically an NP-hard problem<sup>31</sup>, state-of-the-art enumeration algorithms tend to work well on sparse empirical network data<sup>18,19,32</sup>, and indeed we have found that enumeration is not problematic in our experiments.

**Hypergraph prior.** The second part of Eq. (1),  $P(H)$ , is the hypergraph prior. Empirical hypergraphs generally have a few properties that a reasonable prior should account for<sup>33</sup>: the size of interactions varies; some of these interactions are repeated, and not all nodes are connected by a hyperedge. It turns out that an existing model<sup>34</sup>, known as the Poisson Random Hypergraphs Model (PRHM), reproduces all of these properties. Hence, we adopt it as our hypergraph prior. The PRHM was initially developed to study critical phenomena in hypergraphs<sup>34</sup>; here, we use it to make posterior inferences about networks.

In a nutshell, the PRHM stipulates that the number of hyperedges connecting a set of nodes is a random variable, whose mean  $\lambda_k$  only depends on the size  $k$  of the set. The variable follows a Poisson distribution, such that the number of hyperedges connecting the nodes  $i_1, \dots, i_k$  equals to  $A_{i_1, \dots, i_k}$  with probability

$$P(A_{i_1, \dots, i_k} | \lambda_k) = \frac{\lambda_k^{A_{i_1, \dots, i_k}}}{A_{i_1, \dots, i_k}!} e^{-\lambda_k}, \quad (3)$$

where  $A_{i_1, \dots, i_k}$  is invariant with respect to permutation of the indexes. The PRHM also models all the hyperedges as independent. Hence, the probability of a particular hypergraph can be calculated as

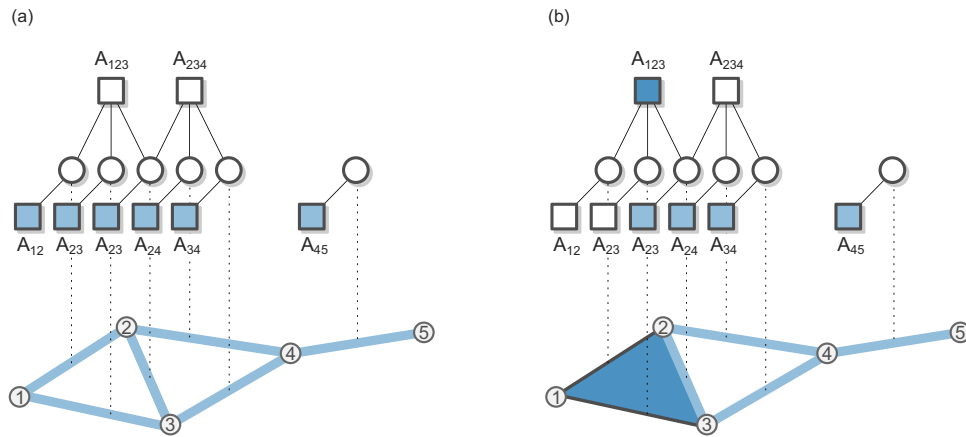
$$\begin{aligned} P(H|\lambda) &= \prod_{k=2}^L \prod_{i_1, \dots, i_k \in C_k^N} P(A_{i_1, \dots, i_k} | \lambda_k) \\ &= \prod_{k=2}^L \prod_{i_1, \dots, i_k \in C_k^N} \frac{\lambda_k^{A_{i_1, \dots, i_k}}}{A_{i_1, \dots, i_k}!} e^{-\lambda_k}, \end{aligned} \quad (4)$$

where  $L$  is the maximal hyperedge size,  $C_k^N$  denotes all possible subsets of size  $k$  of  $\{1, \dots, N\}$ , and where  $\lambda$  refers to all the rates collectively.

Equation (4) expresses the probability of  $H$  in terms of individual hyperedges. To obtain a simpler form, we notice that the number  $E_k$  of hyperedges of size  $k$  can be calculated as

$$E_k = \sum_{i_1, \dots, i_k \in C_k^N} A_{i_1, \dots, i_k} \quad (5)$$

and that there are precisely  $\binom{N}{k}$  terms in the product over all sets of nodes of size  $k$ . We can use these simple observations to



**Fig. 2 Encoding hypergraphs as factor graphs.** The two panels show the factor graph encoding of two different hypergraphs that project to the same network, shown at the bottom of the figure. **a** A hypergraph without higher-order interactions is obtained by associating each edge in  $G$  (blue lines) to a node in the factor graph (empty circles); this correspondence is illustrated with a dashed line. Each node in the factor graph is then connected to a factor  $A_{ij}$  (squares) where  $(i, j)$  is the corresponding edge in  $G$ . Higher-order factors  $A_{ijk}$  corresponding to possible hyperedges of three nodes are also added and connected to the edges  $(i, j), (j, k), (i, k)$  they comprise. Since there is no 4-clique in the graph, the construction stop at this step. A particular combination of hyperedges can then be specified by activating some factors (coloring them blue), here all the factors corresponding to the edges. **b** To encode a hypergraph  $H$  with one higher-order interaction, we mark the factor  $A_{123}$  as active.

rewrite Eq. (4) as

$$P(H|\lambda) = \frac{\prod_{k=2}^L \lambda_k^{E_k} e^{-\binom{N}{k} \lambda_k}}{Z_k}, \tag{6}$$

where we have defined

$$Z_k = \prod_{i_1, \dots, i_k \in C_k^N} A_{i_1, \dots, i_k} = \prod_{m=1}^{\infty} (m!)^{\eta_m^{(k)}}, \tag{7}$$

and where  $\eta_m^{(k)}$  is the number of hyperedges of size  $k$  that are repeated precisely  $m$  times.

In this form, it is clear that the parameters  $\lambda$  control the density of  $H$  at all scales. Hence, they more or less determine the kind of hypergraphs we expect to see a priori, and therefore have a major effect on the model output. How can we choose these important parameters carefully?

We propose to a hierarchical empirical Bayes approach, in which we treat  $\lambda$  as unknowns themselves drawn from prior distributions. We use a maximum entropy, or least informative, prior for  $\lambda$ , because we have no information whatsoever about  $\lambda$  a priori. The only thing we know is that these parameters take values in  $[0, \infty)$  and are modeled with a finite mean<sup>34</sup>. Hence, the maximal entropy prior of interest is the exponential distribution

$$P(\lambda_k | \nu_k) = \frac{e^{-\lambda_k / \nu_k}}{\nu_k}, \tag{8}$$

of mean  $\nu_k$ . We obtain a complete hyperprior for  $\lambda$  by using independent priors for all sizes  $k$ ,  $P(\lambda|\nu) = \prod_{k=2}^L P(\lambda_k | \nu_k)$ . Integrating over the support of  $\lambda$ , we find that the prior for  $H$  is now

$$P(H|\nu) = \int P(H|\lambda) P(\lambda|\nu) d\lambda, \tag{9}$$

$$= \prod_{k=2}^L \frac{E_k!}{Z_k \nu_k} \left[ \frac{1}{\nu_k} + \binom{N}{k} \right]^{-(E_k+1)},$$

with  $\nu$  fixed.

It might appear that we have only pushed our problem further ahead—we got rid of  $\lambda$  but we now have a whole new set of

parameters on our hands. Notice, however, that the new parameters  $\nu$  do not have as direct an effect on  $H$ . A whole range of densities is now compatible with any choice of  $\nu$ . As a result, the model can assign significant probabilities to hypergraphs that project to networks of the correct density, even when the hyperprior is somewhat in error. Hence, we safely fix the new parameters  $\nu$  with empirical Bayes without risking strongly biased results.

With these precautions in place, we use the observed number of edges  $E$  in the network  $G$  to choose  $\nu$ . Our strategy is to equate  $E$  to the expected number of edges  $\langle E(\nu) \rangle$  in the network  $\mathcal{G}(H)$  obtained by projecting  $H$  drawn from  $P(H|\nu)$ . This expected density can be calculated as

$$\langle E(\nu) \rangle = \binom{N}{2} \left[ 1 - \prod_{k=2}^L (e^{-\nu_k} \binom{N}{k-1}) \right], \tag{10}$$

by first computing the reciprocal of the probability that two nodes are not connected by any hyperedge in the hypergraph and then multiplying the result by the total number of node pairs. To set the individual values of  $\nu_k$ , we further require that all sizes contribute equally to the final density, with  $\nu_k \binom{N}{k} = \mu$  for a constant  $\mu$ . Substituting these equalities in Eq. (10), we obtain

$$\mu = (L-1) \log \left( \frac{1}{1 - E / \binom{N}{2}} \right), \tag{11}$$

and the prior of Eq. (9) becomes

$$P(H) = \prod_{k=2}^L \frac{E_k!}{Z_k \binom{N}{k} \mu} \left[ \frac{N-k+1}{k} + \frac{1}{\mu} \right]^{-(E_k+1)}. \tag{12}$$

We note that  $\mu$  diverges as the density  $E / \binom{N}{2}$  of  $G$  approaches one, correctly reflecting the fact that even an infinitely dense hypergraph could have generated the data. This divergence is a

sign that our empirical prior is not well-defined in the extremely dense limit. Since the empirical networks we typically encounter are sparse by construction—we need not worry about this limit in practice.

**Properties of the posterior distribution.** The model defined in Eqs. (2) through (12) has two crucial properties.

The first noteworthy property is that the model assigns a higher posterior probability to hypergraphs without repeated hyperedges, even though the prior  $P(H)$  allows for duplicates. An explicit calculation of how  $P(H|G)$  scales with the number of duplicated hyperedges can illustrate this fact. Indeed, consider a hypergraph  $H_0$  with no repeated hyperedges, for which  $P(G|H_0) = 1$ . Write as  $\alpha$  the fraction of  $k$ -cliques connected by a hyperedge in  $H_0$ , and consider an experiment in which an average of  $\beta \geq 0$  additional hyperedges are placed on top of the hyperedges of size  $k$  already present in  $H_0$ . In these hypergraphs, the expected number of hyperedges of size  $k$  is

$$E_k = \alpha(1 + \beta) \binom{N}{k} \text{ and } \log Z_k \text{ is approximated by}$$

$$\sum_{i_1, \dots, i_k} \log A_{i_1, \dots, i_k} \approx \alpha \binom{N}{k} \log(1 + \beta),$$

see Eq. (7). Substituting our various formula in the logarithm of  $P(H)$ , and using the Stirling approximation  $\log n! \approx n \log n - n$ , we find that

$$\log P(H|G) \sim -\alpha(1 + \beta) \binom{N}{k} \log \left( \frac{N+k-1 + \frac{1}{\mu}}{\alpha} \right).$$

This equation tells us that the log-posterior  $\log P(H|G)$  decreases with growing  $\beta$ , because the argument of the logarithm is at least one. Furthermore, we have  $P(G|H) = 1$  by construction, which implies that the scaling of the prior determines the scaling of the posterior. Hence, the hypergraphs  $H$  generated by adding duplicated hyperedges to  $H_0$ —that is by increasing  $\beta$ —are less likely than  $H_0$ .

A second noteworthy property of the model is that it favors sparser hypergraphs: as long as  $P(G|H) = 1$ , the fewer hyperedges, the better. To make this observation precise, suppose we have a hypergraph  $H_m$  that can be termed minimal for  $G$ : every edge of  $G$  is covered by exactly one hyperedge of  $H_m$  and no more. We observe that we cannot improve on the posterior probability of  $H_m$  by adding a hyperedge, even when this new hyperedge does not fully repeat an existing one. Indeed, consider the hypergraph  $H'_m$  created by adding a hyperedge of size  $k$  to  $H_m$ . For example, we could add a hyperedge of size 3 on a triangle whose sides were already covered by edges, but did not yet participate in any larger hyperedge together. By direct calculation, the ratio of posterior probability for  $H'_m$  and  $H_m$  equals

$$\frac{P(H'_m|G)}{P(H_m|G)} = \frac{E_k + 1}{\binom{N}{k} \left( \frac{N+k-1 + \frac{1}{\mu}}{\alpha} \right)}.$$

This ratio is smaller than one: the minimal property of  $H_m$  implies that  $E_k < \binom{N}{k}$ , and the term in the parenthesis is  $>1$  because  $N \gg k$ . As a result, adding a spurious hyperedge to a minimal hypergraph decreases the posterior probability.

As a corollary of the two above observations, we conclude that the minimal hypergraphs are high-quality local maxima of  $P(H|G)$ . We cannot simply pick one of these optima as our reconstruction, however, because there may exist multiple ones of comparable quality. Further, nonoptimal hypergraphs may account for a significant fraction of the posterior probability in

principle. Instead, we handle these possibly conflicting descriptions by combining them.

**Posterior estimation.** In the Bayesian formulation of hypergraph inference, estimating a given quantity of interests always amount to computing expectations over the posterior distribution  $P(H|G)$ . For example, the expected number of hyperedges of size  $k$  can be computed as  $\langle E_k \rangle = \sum_H E_k(H) P(H|G)$ . More generally, we are interested in averages of the form

$$\langle f(H) \rangle = \sum_H f(H) P(H|G) \tag{13}$$

for arbitrary functions  $f$  that map hypergraphs to vectors or scalars.

The summation in Eq. (13) is unfortunately intractable: the set of possible hypergraphs grows exponentially in size with both the number of nodes and the maximal size of the hyperedges. Hence, we propose a Markov Chain Monte Carlo (MCMC) algorithm to evaluate Eq. (13). This kind of approach generates a random walk over the space of all hypergraphs, with a limiting distribution identical to  $P(H|G)$ . We use the Metropolis–Hastings (MH) construction to implement the random walk. As is usual, the algorithm consists of proposing a move from  $H$  to  $H'$  with probability  $Q(H \leftarrow H')$  and accepting it with probability<sup>35</sup>

$$a = \min \left\{ 1, \frac{Q(H \leftarrow H') P(H'|G)}{Q(H' \leftarrow H) P(H|G)} \right\},$$

$$= \min \left\{ 1, \frac{Q(H \leftarrow H') P(G|H') P(H')}{Q(H' \leftarrow H) P(G|H) P(H)} \right\}. \tag{14}$$

We use the factor graph representation of  $H$  to define these Monte Carlo moves this encoding facilitates checking the value of  $P(G|H')$ . Hence, we can state the moves as modifications to the value of the factors  $A_{i_1, \dots, i_k}$ , i.e., the number of hyperedges connecting particular sets of nodes.

The specific set of moves we use goes as follows. For every move, we begin by choosing a maximal factor node uniformly at random from the set of all such factors. We select a size  $\ell$  uniformly at random from  $\{2, 3, \dots, k\}$ , where  $k$  is the size of the clique corresponding to the current maximal factor. Then, we select one of the subfactors  $A_{i_1, \dots, i_\ell}$  of size  $\ell$  uniformly at random, among the  $\binom{k}{\ell}$  factors of that size, and we update the selected factors as either  $A' = A_{i_1, \dots, i_\ell} + 1$  or  $A' = A_{i_1, \dots, i_\ell} - 1$  (with probability 1/2). If  $A_{i_1, \dots, i_\ell}$  was already equal to zero, we force  $A' = A_{i_1, \dots, i_\ell} + 1$ . Therefore, we have that

$$\frac{Q(H \leftarrow H')}{Q(H' \leftarrow H)} = \begin{cases} 1 & \text{if } A_{i_1, \dots, i_\ell} > 0, \\ 1/2 & \text{if } A_{i_1, \dots, i_\ell} = 0, \\ 2 & \text{if } A' = 0. \end{cases} \tag{15}$$

Finally, we check whether  $P(G|H) = 1$  using the factor representation, and compute the ratio  $P(H')/P(H)$  to obtain the acceptance probability  $a$ . We test for acceptance and, if the move is accepted, we record the update. Otherwise, we do nothing.

The posterior distribution is rugged so the initialization of the MCMC algorithm matters a great deal in practice. Building on our observations about the properties of  $P(H|G)$ , we select as our initialization the hypergraph with one hyperedge for every maximal clique of  $G$ . This starting point is not a known optimum of  $P(H|G)$ , but it is close to many of them. Hence, chains initialized at this point have a fairly good chance of converging to a good optimum. Indeed, in our experiments, we find that the maximal clique initialization works much better than a random initialization, an edge initialization, or an empty one.

**Recovery of planted higher-order interactions in synthetic data.** To develop an intuition for the workings of our method, we first use our algorithm to uncover higher-order interactions in synthetic data generated by the model appearing in Eqs. (2)–(12), altered slightly to facilitate the interpretation of the results. In this experiment, we create a hypergraph that comprises a few large disconnected hyperedges, and we add several random edges (chosen uniformly from the set of all edges) to create a noisy hypergraph  $\tilde{H}$ . We then project this noisy hypergraph to obtain a network  $\mathcal{G}(\tilde{H})$ , which we feed to our recovery algorithm as input. Our goal in this experiment is to find the hypergraph  $H^*$  that maximizes the posterior probability  $P(H|\mathcal{G}(\tilde{H}))$  (we do not use the full samples given by our MCMC algorithm just yet). We can consider the experiment successful if  $H^*$  contains all the higher-order interactions planted in  $\tilde{H}$ .

The results of this experiment are reported in Fig. 3. At the bottom of Fig. 3a, we show a typical example of what the projected networks  $\mathcal{G}(\tilde{H})$  look like when there are very few added random edges. In this regime, the recovered higher-order interactions (in blue) correspond perfectly to those planted in  $\tilde{H}$ . For the sake of comparison, we also generate an equivalent random network, obtained by completely rewiring the edges of  $\mathcal{G}(\tilde{H})$ , see the top of Fig. 3a. (Equivalently, we generate an Erdős–Rényi graph with an equal number of edges<sup>36</sup>.) This network has the same number of edges as  $\mathcal{G}(\tilde{H})$ , but is otherwise unstructured. As expected, we find no higher-order interactions beyond the random triangles that occur at this density<sup>37</sup>.

If we add many more random edges, we obtain the results shown in Fig. 3c. Again, we can recover the planted higher-order interactions, but we also start to find additional ones, due to the appearance of random triangles formed by triplets of edges added at random<sup>38</sup>. To understand this behavior we turn to the minimum description length (MDL) interpretation of Bayesian inference<sup>39,40</sup>.

In a nutshell, the description length is the number of bits that a receiver and a sender with shared knowledge of the model  $P(G, H)$  would need to communicate the network  $G$  to one another. This communication costs can be minimized by finding a hypergraph  $H^*$  that is cheap to communicate and that projects to  $G$ ; receivers who know  $P(G, H)$  also know that they can project

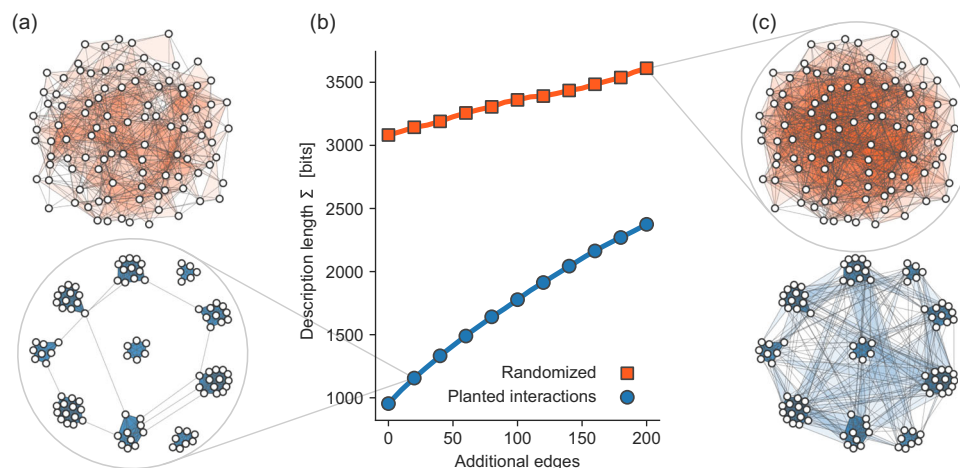
$H^*$  to find  $G = \mathcal{G}(H)^*$ . From this communication perspective, hypergraphs with as few hyperedges as possible are good candidates because they are cheaper to send<sup>24</sup>. The connection with Bayesian inference is that  $H^*$  happens to coincide with the hypergraph which maximizes the posterior probability  $P(H|G)$  (see Supplementary Notes 1 and 2 for a detailed discussion). Hence, maximum a posteriori inference is equivalent to compression.

Reviewing our experiment with the MDL interpretation in mind illuminates the results. In Fig. 3b, we plot the description length provided by our model, for levels of randomness that interpolate between the easy regime shown in Fig. 3a, and the much more random regime appearing in Fig. 3c. We find that the model compresses those networks that have planted interactions much better than their randomized equivalents. These results make intuitive sense: networks with planted interactions contain large cliques, and these cliques can be harnessed to communicate regularities in  $G$ . As can be expected, these savings disappear once the large cliques are destroyed by rewiring.

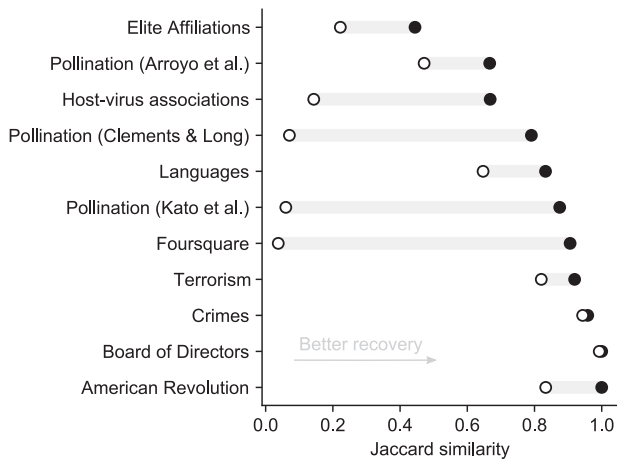
### Recovery of planted higher-order interactions in empirical data.

Having verified that the method works when the higher-order network possesses little structure beyond disjoint planted cliques, we turn to more complicated problems. We ask: can our method identify relevant higher-order interactions when the data are (plausibly) more structured? To answer this question, we use empirical bipartite networks and create higher-order networks, by representing the bipartite networks as hypergraphs  $H^2$ . We then project the hypergraphs with Eq. (2), and attempt to recover the planted higher-order interactions in  $\mathcal{G}(H)$  with our method.

In Fig. 4, we report the results of this experiment for 11 hypergraph constructed with empirical networks datasets<sup>41–45,45–50</sup>. (See also Supplementary Table 1 for a detailed numerical account of the results.) The figure depicts the accuracy of the reconstruction, as quantified by the Jaccard similarity  $J$ , defined as the number of hyperedges found in both the original and the reconstructed hypergraph, divided by the number of hyperedges found in either of them. A similarity of  $J = 1$  denotes perfect agreement, while  $J = 0$  would mean that the hypergraphs share no hyperedges. (When computing  $J$ , we ignore duplicate



**Fig. 3 Reconstruction in random networks with and without higher-order interactions.** **a** Two networks, one obtained by projecting a hypergraph of ten disjoint hyperedges of unequal sizes (blue shades), and the other obtained by drawing uniformly from all networks with the same number of edges (orange shades). **b** Description length  $\Sigma$  of the networks as a function of the number of additional edges, chosen uniformly at random from the set of non-edges. **c** The two networks of panel (a), with 200 additional edges. In **a**, **c**, we show the group interactions uncovered by our method with shaded colors. The description lengths are averaged over ten independent realizations of the network generation and inference processes. Error bars of 1 standard deviations are too narrow to see with the naked eye.



**Fig. 4 Quality of the planted interaction reconstruction in projected bipartite networks.** Empty symbols show the Jaccard similarity of the higher-order interactions reconstructed with the maximal clique decomposition. Filled symbols depict the same quantity for the reconstruction given by our method (best model fit). Our method improves on the baseline in every case. Detailed numerical results are reported in Supplementary Table 1.

hyperedges since they are impossible to distinguish from the projection. For instance, if the board of many companies comprises the exact same directors, then we encode their association with a single hyperedge.)

To obtain a baseline, we also attempt a reconstruction by identifying the maximal cliques of the projected graph to hyperedges—a maximal clique reconstruction. We find that the reconstruction given by our method is good but imperfect, which is expected as the problem is under-determined. However, we also find that our method systematically outperforms the maximal clique decomposition, often by a sizable margin. In many cases, the maximal clique decomposition recovers nearly none of the common interactions, whereas our method reconstructs the interactions to a great extent.

**Detailed case study of higher-order interactions in an empirical network.** To understand why our method works well in practice, it is useful to analyze a small empirical dataset in detail. For this example, we will consider the well-known football network<sup>51</sup>. The nodes of this network represent teams playing in Division I-A of the NCAA (now the NCAA Division I Football Bowl Subdivision), and two teams are connected if they played at least one game during the regular season of Fall 2000. The relationships between teams are viewed through the lens of pairwise interactions, but higher-order phenomena shape the system. For example, the teams of a conference all play each other during a season. Other higher-order phenomena such as geography also intervene: teams in different conferences are likely to meet during the regular season if they are in close-by states. There might also be more subtle phenomena like historical rivalries that survived conference changes. Which of these higher-order organizing principles best determine the structure is not that clear, so there is no single natural bipartite representation of the system—it is best to work with the projected network and let the data guide us.

**Best model fit.** In Fig. 5 we show the interactions that our method uncovers when we look for the single best higher-order description  $H^*$ . We find a large number of interactions that are not pairwise: 86 of the hyperedges of  $H^*$  involve more than two nodes.

The higher-order interactions uncovered by our method are not merely the maximal cliques of  $G$  (see Fig. 5a). We argue that interlocked maximal cliques—cliques that share edges—are the reason why these descriptions differ. When two maximal cliques interlock, the hypergraph constructed directly from maximal cliques contains two overlapping hyperedges. This choice is wasteful from a compression perspective: the edges in the intersection of the two cliques are part of two hyperedges, and therefore contribute twice to the description length  $\Sigma = -\log P(H)$ . Our method instead looks for a more parsimonious description of the data. In doing so, it can identify trade-offs and, for example, represent one of the two cliques as a higher-order interaction and break down the other as a series of smaller interactions, thereby avoiding redundancies. These trade-offs culminate into much better compression: we find a hypergraph  $H^*$  with a description length of 4123.3 bits, which represents a 33.6% saving over the description length of the maximal clique hypergraph (6208.5 bits). The interactions in the optimal hypergraphs do not necessarily map to obvious suspects like subdivisions or geographical clusters; instead, they interact in nonobvious ways and reveal, for example, that one of the subdivisions (top left of Fig. 5b) is best described as two interlocking large hyperedges with a few interactions.

**Probabilistic descriptions.** Being Bayesian, our method provides complete estimation procedures, beyond maximum a posteriori estimation. For example, a quantity of particular interest is the posterior probability that a set of nodes is connected by at least one hyperedge<sup>28</sup>, once we account for the full distribution over hypergraphs  $P(H|G)$ . By computing this probability for all sets of nodes with a non-negligible connection probability, we can encode the probabilistic structure of  $H$  in a compact way<sup>52</sup>, with a few probabilities only.

In practice, we evaluate the connection probabilities by generating samples from  $P(H|G)$  and counting the samples in which a set of interest is connected by at least one hyperedge (recall that the model defines a distribution over hypergraph with repeated hyperedges). Mathematically this is computed as

$$P(X_{i_1, \dots, i_k} = 1|G) = \frac{1}{n} \sum_{\ell=1}^n X_{i_1, \dots, i_k}(H_\ell), \quad (16)$$

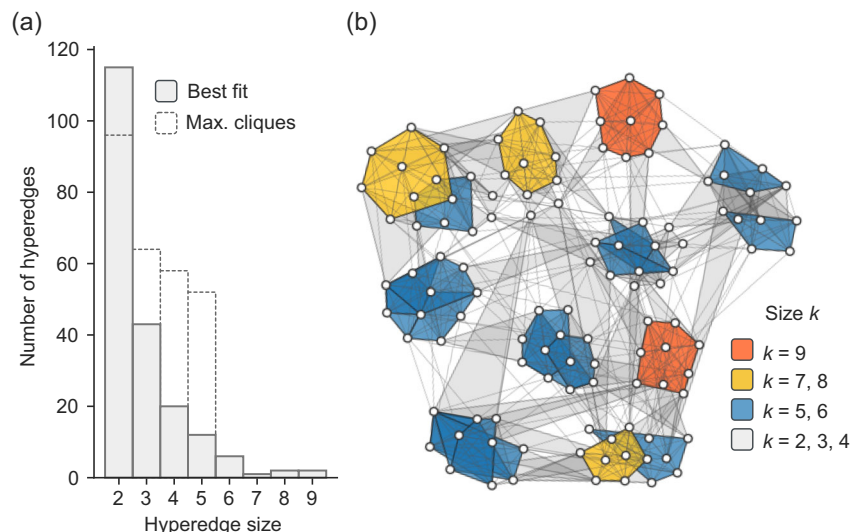
where  $H_1, \dots, H_n$  are  $n$  hypergraph sampled from  $P(H|G)$ , and where  $X_{i_1, \dots, i_k}(H) = \mathbb{1}_{A_{i_1, \dots, i_k} \neq \emptyset}$  is a presence/absence variable, equal to 1 if and only if there is at least one hyperedge connecting nodes  $i_1, \dots, i_k$  in hypergraph  $H$ .

Applying this technique to the Football data, we find that many of the hyperedges of  $H^*$  have a presence probability close to 1, even once we account for the full distribution over hypergraphs. The hypergraph is not reconstructed with absolute certainty, however. Observing that probabilities  $P(X_{i_1, \dots, i_k} = 1|G)$  close to 1 or 0 both indicate confidence in the presence/absence of edge, we define a certainty threshold  $\alpha$  and classify all hyperedges with existence probabilities in  $[\alpha/2, 1 - \alpha/2]$  as uncertain. With a threshold of  $\alpha = 0.05$ , we find six uncertain triangles (hyperedges on three nodes) and five uncertain edges in the Football data.

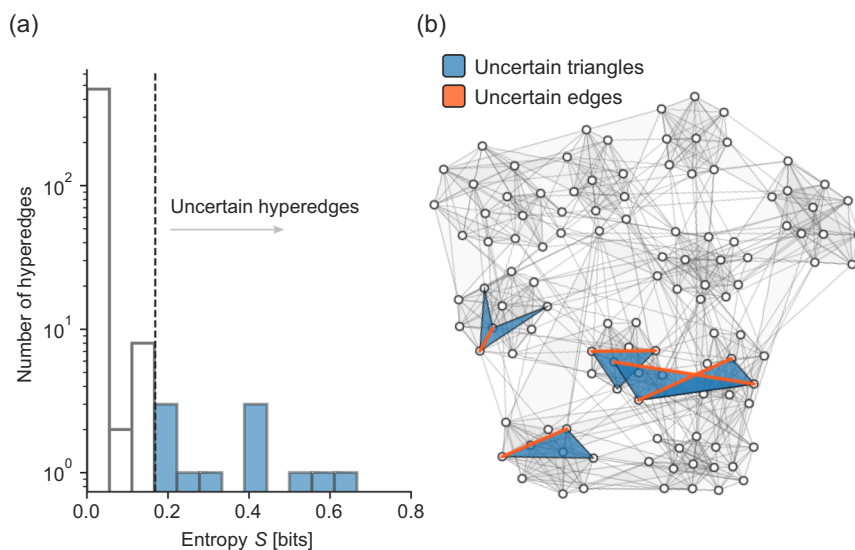
To go beyond a simple threshold analysis, we compute the entropy of the probabilities  $\hat{p} := P(X_{i_1, \dots, i_k} = 1|G)$ , defined as

$$S(\hat{p}) = -\hat{p} \log_2 \hat{p} - (1 - \hat{p}) \log_2 (1 - \hat{p}). \quad (17)$$

The entropy provides a useful transformation of  $\hat{p}$  because it grows as  $\hat{p}$  moves away from the extremes  $\hat{p} = 0, 1$ , with a maximum of  $S = 1$  at  $\hat{p} = 1/2$ , the point of maximal uncertainty. The distribution of entropy is shown in Fig. 6a for the Football data. The figure shows that while the majority of hyperedges are certain (i.e., their entropy is greater than  $S^* \approx 0.169$



**Fig. 5 Higher-order interactions uncovered in the network of American football games.** **a** Size distribution of the hyperedges of the hypergraph  $H^*$  that maximizes the posterior probability or, alternatively, minimizes the description length of the network of football games  $G$ . Also shown is the distribution of hyperedge sizes for the hypergraph constructed by assigning a hyperedge to every maximal clique of  $G$  (dashed histogram). **b** Visualization of the hyperedges present in  $H^*$ , color-coded by size.



**Fig. 6 Uncertain higher-order interactions uncovered in the network of American football games.** **a** Distribution of the entropy for the presence/absence of hyperedges. The entropy quantifies the variability of hyperedges: sets of nodes that are connected in nearly all—or almost none—of the samples have low entropy. We deem as uncertain a hyperedge that has a probability  $p \in [\alpha/2, 1 - \alpha/2]$  of being present, with  $\alpha = 0.05$ . Note that we only show the entropy for the sets of nodes that were connected at least once in our Monte Carlo samples; a large number of hyperedges, of entropy zero, are never seen in our samples. **b** Visualization of the uncertain hyperedges. Uncertain edges are highlighted in orange and uncertain triangles are shown in blue. All results are computed with 2000 Monte Carlo samples from the posterior distribution each separated by 2000 complete sweep of the factor graph.

corresponding to  $\alpha = 0.05$ ), making the certainty criterion slightly more stringent would add many more uncertain hyperedges to the ones we already have.

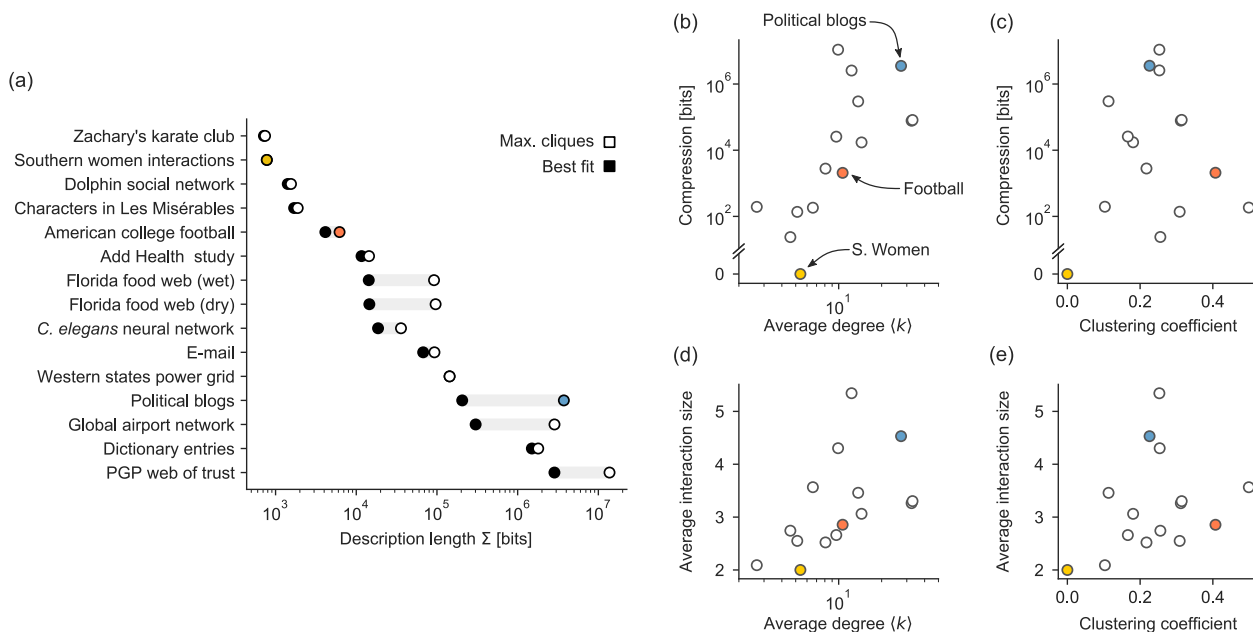
In Fig. 6b, we show the location of the uncertain hyperedges in  $H$ . We observe that these uncertain hyperedges are co-located. The minimality properties of  $P(H|G)$  discussed above can explain these results. Hypergraphs that have a sizable posterior probability are typically sparse and include as few hyperedges as possible. But they also need to cover the whole graph, meaning that every edge of  $G$  needs to appear as a subset of at least one hyperedge  $H$  (due to the constraint  $G = \mathcal{G}(H)$ ). Co-located uncertain triangles and hyperedges are hence the result of competing solutions of roughly equal qualities, which cover

a specific part of the hypergraph with hyperedges of different sizes.

**Systematic analysis of higher-order interactions in empirical networks.** For our fourth and final example, we apply our method to 15 network datasets, taken from various representative scientific domains and structural classes<sup>11,14,51,53–64</sup>.

For each empirical network in our list, we first search for the hypergraph  $H^*$  that maximizes  $P(H|G)$ , as we have done in our two previous examples. This search gives us an MDL  $\Sigma$ . For the sake of comparison, we also compute the description length  $\Sigma'$  that we would obtain if we were to use the maximum clique





**Fig. 7 Higher-order interaction in empirical networks.** A few datasets are highlighted with colors: the Southern women interaction data<sup>53</sup> (yellow), the Football data<sup>51</sup> (orange), and the political blogs<sup>54</sup> (blue). **a** Description length of the hypergraph  $H$  whose hyperedges are the maximal cliques of the input network  $G$ , compared with the description length found with our method. **b, c** Compression, defined as the difference in description length, as a function of the average degree and clustering coefficient<sup>2</sup>. **d, e** Interaction size, averaged over all interactions in  $H^*$ , as a function of the average degree and clustering coefficient. Detailed numerical results are reported in Supplementary Table 2.

decomposition to construct  $H$  naively. We note that  $\Sigma'$  cannot be smaller than  $\Sigma$  because it is the description length of the starting point of the MCMC algorithm—at best, the algorithm cannot improve on  $\Sigma'$ , and we then have  $\Sigma = \Sigma'$ . The difference  $\Sigma' - \Sigma$  gives the compression factor or, in other words, the number of bits we save by using the best hypergraph instead of a hypergraph of maximal cliques.

In Fig. 7a, we show the description lengths of the networks in our collection of datasets. We observe a broad range of outcomes. Compression of multiple orders of magnitude is possible in some cases, like with the political blogs data<sup>54</sup> highlighted in blue, while the best description is directly the maximal cliques in others, like with the Southern women interaction data<sup>53</sup> highlighted in yellow. We find that the average degree of the nodes correlates with compression (Kendall's  $\tau = 0.53$ ) (see Fig. 7b.) This result is expected: the denser a network, the more likely it is that interlocking cliques are present, and therefore that a parsimonious description can be obtained by optimizing over  $P(H|D)$ . The average local clustering coefficient  $\langle C \rangle^2$  is not correlated with compression, however ( $\tau = -0.07$ ) (see Fig. 7c). Local clustering quantifies the density of closed triangles in the neighborhood of a node and is, as such, a proxy for the density of cliques. However, as our results show,  $\langle C \rangle$  fails to capture the correct type of redundancy necessary for good compression with our model.

We note that clustering, nonetheless, predicts the absence of compression well: If  $\langle C \rangle = 0$ , then there are no closed triangles in  $G$ , and it is impossible to compress the network with our method—there are no cliques, and therefore no higher-order interactions in the data. The Southern Women<sup>53</sup> falls in this category because it is a bipartite network.

In Fig. 7d, e, we show the size of the higher-order interactions found by our method, averaged over the hyperedges of  $H^*$ . We again observe a wide range of outcomes. As a sanity check, we can confirm that the incompressible network has an average interaction size of 2. All hypergraphs are just networks in this case and therefore have no higher-order interactions. Other datasets yield hypergraphs with large interactions on average,

involving as many as five nodes in the airport network. The correlation between local properties and interaction size is not as strong as with compression, but there are some dependencies ( $\tau = 0.40$  and  $\tau = 0.27$  for the degree and local clustering, respectively). These might be partly explained by constraints on the possible values that the average interaction size  $\langle s \rangle$  can adopt. For instance, to have an average size  $\langle s \rangle$ , a network must have an average degree of at least  $\langle s \rangle - 1$ . Likewise, some level of clustering is required to obtain large interactions.

Summarizing these results, we find that some level of compression is always possible, except when the network has no clustering whatsoever. Furthermore, we find that a high average degree is related to more compression and larger higher-order interactions. Finally, we find that some minimal level of clustering is necessary for compression, but that results vary otherwise.

### Conclusion

Higher-order interactions shape most relational datasets<sup>7,26</sup>, even when they are not explicitly encoded. In this work, we have shown that it is possible to recover these interactions from data. We have argued that while the problem is ill-defined, one can introduce regularization in the form of a Bayesian generative model, and obtain a principled recovery method.

The framework we have presented is close in spirit to precursors who have used a generative model to find small patterns in networks, so it is worth pointing out where it differs, both in its methodological details and philosophical underpinning. Closely related work includes that of Wegner<sup>23</sup>, who used a notion of probabilistic subgraph covers to induce distributions over possible decomposition in motifs, and more recent works in graph machine learning that solve graph compression by decomposing the network in small building blocks<sup>24,25</sup>. Unlike these authors, however, we focused on higher-order interactions, so we considered decompositions in hyperedges rather than in general motif grammars. We also differ on a methodological ground: we

embraced the complexity of the problem and proposed a fully Bayesian method that can account for the multiplicity of descriptions, in contrast with the greedy optimization favored in previous work<sup>23–25</sup>. As a result of these methodological choices, our work is perhaps closest to that of Williamson and Tec<sup>1</sup>, who also solved a similar problem by using Bayesian nonparametric techniques<sup>1</sup>, and view a network as collections of overlapping cliques. Unlike these authors, however, we have formalized network data as uncorrupted; in our framework, latent higher-order interactions always show up in network data as fully connected cliques. In contrast, they think of this process as noisy, so latent higher-order interactions can translate into relatively sparsely connected sets of nodes. Their proposed methods thus bear a resemblance to community detection techniques that formalize communities as noisily measured cliques<sup>27,65–69</sup>.

The method we have proposed here is undoubtedly one of the simplest instantiations of the broader idea of uncovering higher-order interactions in empirical relational data. There are many ways in which one could expand on the method. On the modeling front, for example, it would be worthwhile to study the interplay of the projection component  $P(G|H)$  of Eq. (2) and inference: can it be defined in a way that does not turn higher-order interaction discovery into overlapping community detection? The hypergraph prior, too, will have to be expanded as the PRHM we have used is pretty simple. Interesting models could include degree heterogeneity as part of the reconstruction<sup>70–72</sup>, or community structure<sup>73</sup>. One could also envision a simplicial analog to these models, leading to probabilistic simplicial complex recovery<sup>74,75</sup>. Finally, it would be interesting to explore the connection between different forms of regularizations that make the problem well-defined.

On the technical front, it will be interesting to see whether more refined MCMC methods can lead to more robust convergence and faster mixing. Our proposed move-set is among the simplest ones that can propose for the problem and could be improved. Another interesting avenue of research will be to harness the known properties of  $P(H|G)$  to construct efficient inference algorithms and perhaps connect the method to algorithms in the graph theory of clique covers.

The higher-order interaction data we need to inform the development of higher-order network science<sup>7</sup> are often inaccessible. Our methods provide the tools needed to extract higher-order structures from much more accessible and abundant relation data. With this work, we hope to have shown that moving to principled techniques is possible, and that ad hoc reconstruction methods should be avoided, in favor of those based on information-theoretic parsimony and statistical evidence.

### Data availability

The network data that support the findings of this study are available online in the Netzschleuder network repository<sup>76</sup>.

### Code availability

A fast implementation of the Markov Chain Monte Carlo algorithm described in this study is freely available as part of the `graph-tool` Python library<sup>77</sup>.

Received: 8 October 2020; Accepted: 25 May 2021;

Published online: 15 June 2021

### References

- Williamson, S. A. & Tec, M. Random clique covers for graphs with local density and global sparsity. *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference, in Proceedings of Machine Learning Research*, **115**, 228–238 (2020).
- Newman, M. *Networks* 2nd edn (Oxford Univ. Press, 2018).
- Frank, O. & Strauss, D. Markov graphs. *J. Am. Stat. Assoc.* **81**, 832–842 (1986).
- Iacobucci, D. & Wasserman, S. Social networks with two sets of actors. *Psychometrika* **55**, 707–720 (1990).
- Watts, D. J., Dodds, P. S. & Newman, M. E. J. Identity and search in social networks. *Science* **296**, 1302–1305 (2002).
- Newman, M. E. J. Properties of highly clustered networks. *Phys. Rev. E* **68**, 026121 (2003).
- Battiston, F. et al. Networks beyond pairwise interactions: structure and dynamics. *Phys. Rep.* **874**, 1 (2020).
- Latapy, M., Magnien, C. & Del Vecchio, N. Basic notions for the analysis of large two-mode networks. *Soc. Netw.* **30**, 31–48 (2008).
- Pollner, P., Palla, G. & Vicsek, T. Preferential attachment of communities: the same principle, but a higher level. *Europhys. Lett.* **73**, 478 (2005).
- Hébert-Dufresne, L., Laurence, E., Allard, A., Young, J.-G. & Dubé, L. J. Complex networks as an emerging property of hierarchical preferential attachment. *Phys. Rev. E* **92**, 062809 (2015).
- White, J. G., Southgate, E., Thomson, J. N. & Brenner, S. The structure of the nervous system of the nematode *Caenorhabditis elegans*. *Philos. Trans. R. Soc. Ser. B* **314**, 1–340 (1986).
- Atkin, R. *Mathematical Structure in Human Affairs* (Heinemann, 1974).
- Grilli, J., Barabás, G., Michalska-Smith, M. J. & Allesina, S. Higher-order interactions stabilize dynamics in competitive network models. *Nature* **548**, 210–213 (2017).
- Resnick, M. D. et al. Protecting adolescents from harm: findings from the national longitudinal study on adolescent health. *J. Am. Med. Assoc.* **278**, 823–832 (1997).
- Patania, A., Vaccarino, F. & Petri, G. Topological analysis of data. *EPJ Data Sci.* **6**, 7 (2017).
- Petri, G., Scolamiero, M., Donato, I. & Vaccarino, F. Networks and cycles: a persistent homology approach to complex networks. In *Proc. European Conference on Complex Systems 2012*, 93–99 (2013).
- Petri, G., Scolamiero, M., Donato, I. & Vaccarino, F. Topological strata of weighted complex networks. *PLoS ONE* **8**, e66506 (2013).
- Bron, C. & Kerbosch, J. Algorithm 457: finding all cliques of an undirected graph. *Commun. ACM* **16**, 575–577 (1973).
- Tomita, E., Tanaka, A. & Takahashi, H. The worst-case time complexity for generating all maximal cliques and computational experiments. *Theor. Comput. Sci.* **363**, 28–42 (2006).
- Jain, S. & Seshadhri, C. A fast and provable method for estimating clique counts using Turán’s theorem. In *Proc. 26th International Conference on World Wide Web*, 441–449 (2017).
- Erdős, P., Goodman, A. W. & Pósa, L. The representation of a graph by set intersections. *Can. J. Math.* **18**, 106–112 (1966).
- Coutinho, B. C., Wu, A.-K., Zhou, H.-J. & Liu, Y.-Y. Covering problems and core percolations on hypergraphs. *Phys. Rev. Lett.* **124**, 248301 (2020).
- Wegner, A. E. Subgraph covers: an information-theoretic approach to motif analysis in networks. *Phys. Rev. X* **4**, 041026 (2014).
- Koutra, D., Kang, U., Vreeken, J. & Faloutsos, C. Vog: summarizing and understanding large graphs. In *Proc. 2014 SIAM International Conference on Data Mining*, 91–99 (SIAM, 2014).
- Liu, Y., Safavi, T., Shah, N. & Koutra, D. Reducing large graphs to small supergraphs: a unified approach. *Soc. Netw. Anal. Min.* **8**, 17 (2018).
- Torres, L., Blevins, A. S., Bassett, D. S. & Eliassi-Rad, T. The why, how, and when of representations for complex systems. Preprint at <https://arxiv.org/abs/2006.02870> (2020).
- Barber, D. Clique matrices for statistical graph decomposition and parameterising restricted positive definite matrices. Preprint at <https://arxiv.org/abs/1206.3237> (2012).
- Young, J.-G., Cantwell, G. T. & Newman, M. Bayesian inference of network structure from unreliable data. *J. Complex Netw.* **8**, cnaa046 (2020).
- Peixoto, T. P. Reconstructing networks with unknown and heterogeneous errors. *Phys. Rev. X* **8**, 041011 (2018).
- Bishop, C. M. *Pattern Recognition and Machine Learning* (Springer, 2006).
- Karp, R. M. In *Complexity of Computer Computations: Proc. of a Symp. on the Complexity of Computer Computations* (eds. Miller, R. E. & Thatcher, J. W.) The IBM Research Symposia Series, 85–103 (Plenum Press, 1972).
- Fox, J., Roughgarden, T., Seshadhri, C., Wei, F. & Wein, N. Finding cliques in social networks: a new distribution-free model. *SIAM J. Comput.* **49**, 448–464 (2020).
- Aksoy, S. G., Joslyn, C., Ortiz Marrero, C., Praggastis, B. & Purvine, E. Hypernetwork science via high-order hypergraph walks. *EPJ Data Sci.* **9**, 16 (2020).
- Darling, R. W. & Norris, J. R. Structure of large random hypergraphs. *Ann. Appl. Probab.* **15**, 125–152 (2005).
- Andrieu, C., De Freitas, N., Doucet, A. & Jordan, M. I. An introduction to mcmc for machine learning. *Mach. Learn.* **50**, 5–43 (2003).

36. Erdős, P. & Rényi, A. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.* **5**, 17–60 (1960).
37. Bollobás, B. & Erdős, P. Cliques in random graphs. *Math. Proc. Cambridge Philos. Soc.* **80**, 419–427 (1976).
38. Shi, X., Adamic, L. A. & Strauss, M. J. Networks of strong ties. *Phys. A* **378**, 33–47 (2007).
39. MacKay, D. J. C. *Information Theory, Inference and Learning Algorithms* 1st edn (Cambridge Univ. Press, 2003).
40. Grünwald, P. D. *The Minimum Description Length Principle* (MIT Press, 2007).
41. Barnes, R. C. Structural redundancy and multiplicity within networks of us corporate directors. *Crit. Sociol.* **43**, 37–57 (2017).
42. Arroyo, M. T. K., Armesto, J. J. & Primack, R. B. Community studies in pollination ecology in the high temperate andes of central Chile ii. Effect of temperature on visitation rates and pollination possibilities. *Plant Syst. Evol.* **149**, 187–203 (1985).
43. Olival, K. J. et al. Host and viral traits predict zoonotic spillover from mammals. *Nature* **546**, 646–650 (2017).
44. Clements, F. E. & Long, F. L. *Experimental Pollination: An Outline of the Ecology of Flowers and Insects*, No. 336 (Carnegie Institution of Washington, 1923).
45. Kunegis, J. KONECT: the Koblenz network collection. In *Proc. 22nd International Conference on World Wide Web*, 1343–1350 (2013).
46. Kato, M., Kakutani, T., Inoue, T. & Itino, T. Insect-flower relationship in the primary beech forest of Ashu, Kyoto: an overview of the flowering phenology and the seasonal pattern of insect visits. *Contributions Biol. Lab., Kyoto Univ.* **27**, 309–376 (1990).
47. Yang, D., Zhang, D., Yu, Z. & Yu, Z. Fine-grained preference-aware location search leveraging crowdsourced digital footprints from lbsns. In *Proc. 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 479–488 (2013).
48. Gerdes, L. M., Ringler, K. & Autin, B. Assessing the Abu Sayyaf group's strategic and learning capacities. *Stud. Confl. Terrorism* **37**, 267–293 (2014).
49. University of Missouri--St. Louis, Saint Louis (Mo.), Saint Louis (Mo.). Metropolitan Police Department, Missouri. Department of Health. *The St. Louis Homicide Project: Local Responses to a National Problem* (University, 1991).
50. Seierstad, C. & Opsahl, T. For the few not the many? The effects of affirmative action on presence, prominence, and social capital of women directors in Norway. *Scand. J. Manag.* **27**, 44–54 (2011).
51. Girvan, M. & Newman, M. E. J. Community structure in social and biological networks. *Proc. Natl Acad. Sci. USA* **99**, 7821–7826 (2002).
52. Parchas, P., Gullo, F., Papadias, D. & Bonchi, F. Uncertain graph processing through representative instances. *ACM Trans. Database Syst.* **40**, 1–39 (2015).
53. Davis, A., Gardner, B. B. & Gardner, M. R. *Deep South: A Social Anthropological Study of Caste and Class* (Univ. South Carolina Press, 2009).
54. Adamic, L. A. & Glance, N. The political blogosphere and the 2004 US election: divided they blog. In *Proc. 3rd international workshop on Link discovery*, 36–43 (2005).
55. Zachary, W. W. An information flow model for conflict and fission in small groups. *J. Anthropol. Res.* **33**, 452–473 (1977).
56. Lusseau, D. et al. The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations. *Behav. Ecol. Sociobiol.* **54**, 396–405 (2003).
57. Knuth, D. E. *The Stanford GraphBase: A Platform for Combinatorial Computing* 1st edn (Addison-Wesley, 1993).
58. Ulanowicz, R. E. & DeAngelis, D. L. Network analysis of trophic dynamics in South Florida ecosystems. In *US Geological Survey Program on the South Florida Ecosystem*, Vol. 114, 45 (1999).
59. Newman, M. E. J. Modularity and community structure in networks. *Proc. Natl Acad. Sci. USA* **103**, 8577–8582 (2006).
60. Guimera, R., Danon, L., Diaz-Guilera, A., Giralt, F. & Arenas, A. Self-similar community structure in a network of human interactions. *Phys. Rev. E* **68**, 065103 (2003).
61. Watts, D. J. & Strogatz, S. H. Collective dynamics of 'small-world' networks. *Nature* **393**, 440 (1998).
62. Peixoto, T. P. Hierarchical block structures and high-resolution model selection in large networks. *Phys. Rev. X* **4**, 011047 (2014).
63. Batagelj, V., Mrvar, A. & Zaversnik, M. *Network Analysis of Texts* 143–148 (Language Technologies, 2002).
64. Richters, O. & Peixoto, T. P. Trust transitivity in social networks. *PLoS ONE* **6**, e18384 (2011).
65. Davis, G. B. & Carley, K. M. Clearing the fog: fuzzy, overlapping groups for social networks. *Soc. Netw.* **30**, 201–212 (2008).
66. Airoldi, E. M., Blei, D. M., Fienberg, S. E. & Xing, E. P. Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.* **9**, 1981–2014 (2008).
67. Xie, J., Kelley, S. & Szymanski, B. K. Overlapping community detection in networks: the state-of-the-art and comparative study. *ACM Comput. Surv.* **45**, 1–35 (2013).
68. Verzelen, N. & Arias-Castro, E. et al. Community detection in sparse random networks. *Ann. Appl. Probab.* **25**, 3465–3510 (2015).
69. Fortunato, S. Community detection in graphs. *Phys. Rep.* **486**, 75–174 (2010).
70. Peixoto, T. P. Latent poisson models for networks with heterogeneous density. *Phys. Rev. E* **102**, 012309 (2020).
71. Stasi, D., Sadeghi, K., Rinaldo, A., Petrović, S. & Fienberg, S. E.  $\beta$  models for random hypergraphs with a given degree sequence. Preprint at <https://arxiv.org/abs/1407.1004> (2014).
72. Chodrow, P. S. Configuration models of random hypergraphs. *J. Complex Netw.* **8**, cnaa018 (2020).
73. Chodrow, P. S., Veldt, N. & Benson, A. R. Hypergraph clustering: from blockmodels to modularity. Preprint <https://arxiv.org/abs/2101.09611> (2021).
74. Young, J.-G., Petri, G., Vaccarino, F. & Patania, A. Construction of and efficient sampling from the simplicial configuration model. *Phys. Rev. E* **96**, 032312 (2017).
75. Courtney, O. T. & Bianconi, G. Generalized network structures: the configuration model and the canonical ensemble of simplicial complexes. *Phys. Rev. E* **93**, 062311 (2016).
76. Peixoto, T. P. The Netzschleuder network catalogue and repository. <https://networks.skewed.de> (2020).
77. Peixoto, T. P. The graph-tool python library. figshare <https://graph-tool.skewed.de> (2014).

### Acknowledgements

This work was funded, in part, by the James S. McDonnell Foundation (J.-G.Y.), the Sanpaolo Innovation Center (G.P.), and the Compagnia San Paolo via the AdnD project (G.P.).

### Author contributions

J.-G.Y., G.P. and T.P.P. conceptualized the study. T.P.P. and J.-G.Y. developed the model and implemented the algorithm. J.-G.Y. performed the numerical experiments. J.-G.Y., G.P. and T.P.P. analyzed the results and contributed to writing the manuscript.

### Competing interests

The authors declare no competing interests.

### Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s42005-021-00637-w>.

**Correspondence** and requests for materials should be addressed to J.-G.Y.

**Peer review information** *Communications Physics* thanks the anonymous reviewers for their contribution to the peer review of this work.

**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) 2021