NETWORK SCIENCE

Community detection in large hypergraphs

Nicolò Ruggeri[1,2]*, Martina Contisciani[1], Federico Battiston[3], Caterina De Bacco[1]*

Hypergraphs, describing networks where interactions take place among any number of units, are a natural tool to model many real-world social and biological systems. Here, we propose a principled framework to model the organization of higher-order data. Our approach recovers community structure with accuracy exceeding that of currently available state-of-the-art algorithms, as tested in synthetic benchmarks with both hard and overlapping ground-truth partitions. Our model is flexible and allows capturing both assortative and disassortative community structures. Moreover, our method scales orders of magnitude faster than competing algorithms, making it suitable for the analysis of very large hypergraphs, containing millions of nodes and interactions among thousands of nodes. Our work constitutes a practical and general tool for hypergraph analysis, broadening our understanding of the organization of real-world higher-order systems.

INTRODUCTION

Over the last decades, most relational data, from biological to social systems, have found a successful representation in terms of networks, where nodes describe the basic units of the system, and link their pairwise interactions (*1*). Nevertheless, such a modeling approach cannot properly encode the presence of group interactions, describing associations among three or more system units at a time (*2*–*5*). Such higher-order interactions have been observed in a wide variety of systems, including collaboration networks (*6*), cellular networks (*7*), drug recombination (*8*), human (*9*) and animal (*10*) face-to-face interactions, and structural and functional mapping of the human brain (*11*–*13*). In addition, the higher-order organization of many interacting systems is associated with the generation of new phenomena and collective behavior across many different dynamical processes, such as diffusion (*14*), synchronization (*15*–*20*), spreading (*21*–*23*), and evolutionary games (*24*–*26*).

Networked systems with higher-order interactions are better described by different mathematical frameworks from networks, such as hypergraphs, where hyperedges encode interactions among an arbitrary number of system units (*2*, *27*). In the last few years, several tools have been developed for higher-order network analysis. These include higher-order centrality scores (*28*, *29*), clustering (*30*), and motif analysis (*31*, *32*), as well as higher-order approaches to network backboning (*33*, *34*), link prediction (*35*), and methods to reconstruct nondyadic relationships from pairwise interaction records (*36*). A variety of approaches have been suggested to detect communities in hypergraphs, including nonparametric methods with hypergraphons (*37*), tensor decompositions (*38*), latent space distance models (*39*), latent class models (*40*), flow-based algorithms (*41*, *42*), spectral clustering (*43*–*45*), and spectral embeddings (*46*). A different line of works focuses on deriving theoretical detectability limits (*47*–*49*).

Recently, statistical inference frameworks have been proposed to capture in a principled way the mesoscale organization of hypergraphs (*35*, *50*, *51*). Despite their success, current approaches suffer from a number of notable drawbacks. For instance, the method in (*51*) is restricted to using very small hypergraphs and hyperedges, due to its high computational complexity. Also, the approach in (*50*) suffers from a high computational complexity in the general case and needs to make strong assumptions to scale to real-life datasets. Finally, the model in (*35*) is constrained to work only with assortative community structures.

Here, we propose a framework to model the organization of higher-order systems. Our method allows detecting communities in hypergraphs with accuracy exceeding that of state-of-the-art approaches, in the cases of both hard and mixed community assignments, as we show on synthetic benchmarks with known ground-truth partitions. Furthermore, its flexibility allows capturing general configurations that could not be previously studied, such as disassortative community interactions.

Finally, overcoming the computational thresholds of previous methods, our model is extremely efficient, making it suitable to study hypergraphs containing millions of nodes and interactions among thousands of system units not accessible to alternative tools. We illustrate the advantages of our approach through a variety of experiments on synthetic and real data. Our results showcase the wide applicability of the proposed method, contributing to broaden our understanding of the organization of higher-order real-world systems.

GENERATIVE MODEL

A hypergraph consists of a set of nodes $V = \{1, ..., N\}$ and a set of hyperedges E. Each hyperedge e is a subset of V, representing a higher-order interaction between a number $|e|$ of nodes. We denote by D the maximum possible hyperedge size, which can be arbitrarily imposed up to a maximum value of $D = N$, and Ω the set of all possible hyperedges among nodes in V. We represent the hypergraph via an adjacency vector $\mathbf{A} \in \mathbb{N}^{\Omega}$, with entry A_e being the weight of $e \in \Omega$. We assume the weights A_e to be nonnegative and discrete. For real-world systems, $\mathbf{A}$ is typically sparse. The number $|E|$ of nonzero entries is typically linear in N, and thus much smaller than the dimension $|\Omega|$.

We model hypergraphs probabilistically, assuming an underlying arbitrary community structure with K overlapping groups, similarly to a mixed-membership stochastic block model. Each node i

[1]Max Planck Institute for Intelligent Systems, Cyber Valley, 72076 Tübingen, Germany. [2]Department of Computer Science, ETH, 8004 Zürich, Switzerland. [3]Department of Network and Data Science, Central European University, 1100 Vienna, Austria.
*Corresponding author. Email: nicolo.ruggeri@tuebingen.mpg.de (N.R.); caterina.debacco@tuebingen.mpg.de (C.D.B.)

can potentially belong to multiple groups, as specified by a K-dimensional membership vector $\mathbf{u_i}$ with nonnegative entries. We collect all the membership assignments in a $N \times K$ matrix u. The density of interactions within and between communities is regulated by a symmetric nonnegative $K \times K$ affinity matrix w. These two main parameters, u and w, control the Poisson distributions of the hyperedge weights

$$p(A_e; u, w) = \text{Pois}\left(A_e; \frac{\lambda_e}{\kappa_e}\right) \tag{1}$$

where

$$\begin{aligned}\lambda_e &= \textstyle\sum_{i<j:i,j\in e} \mathbf{u_i}^T w\, u_j \\ &= \textstyle\sum_{i<j:i,j\in e} \sum_{k,q=1}^{K} u_{ik}\, u_{jq}\, w_{kq}\end{aligned} \tag{2}$$

Here, $\kappa_e = \kappa_{|e|}$ is a normalization factor that solely depends on the hyperedge size $|e|$. We develop our theory for a general form of κ_n. While in principle any choice $\kappa_n > 0$ is possible, in our experiments we use the form $\kappa_n = \frac{n(n-1)}{2}\binom{N-2}{n-2}$, for every hyperedge of size n (*52*). Because of the fact that $\kappa_2 = 1$, if the hypergraph contains only pairwise interactions our model is similar to existing mixed-membership block models for dyadic networks (*53*, *54*). Intuitively, given two nodes i, j, the term $\binom{N-2}{n-2}$ normalizes for the number of possible choices of the remaining $n-2$ nodes in the hyperedge. The term $n(n-1)/2$ averages among the number of possible pairwise interactions among the n nodes in the hyperedge. Note that previous generative models for hypergraphs were limited to detect only assortative community interactions (*35*, *50*). By contrast, in our model, each entry w_{kq} distinctly specifies the strength of the interactions between each k, q community pair. Hence, for the first time, our method allows encoding more general community structures, without the need to impose a priori assumptions to ensure computational and theoretical feasibility. In particular, the bilinear form in Eq. 2 allows for a tractable and scalable inference, regardless of the structure of w. Another relevant feature of the model is that the size of the affinity matrix w does not vary with maximum hyperedge size D nor with the number of hyperedges, making it memory efficient also for hypergraphs with large interactions. We name our model Hy-MMSBM, for hypergraph mixed-membership stochastic block model, and provide an open-source implementation at http://github.com/nickruggeri/Hy-MMSBMgithub.com/nickruggeri/Hy-MMSBM. We have also incorporated our algorithm inside the open-source library Hypergraphx (*55*).

INFERENCE

Optimization procedure

In real-life scenarios, practitioners observe a list of hyperedges, encoded in the vector $\mathbf{A}$, and aim to learn the node memberships u and affinity matrix w that best fit the data. To this end, we start by considering the likelihood of $\mathbf{A}$ given the parameters $\theta = (u, w)$. Using Eqs. 1 and 2, this is given by

$$p(\mathbf{A}\,|\,\theta) = \prod_{e\in\Omega} \text{Pois}\left(A_e; \frac{\lambda_e}{\kappa_e}\right) \tag{3}$$

where the hyperedge weights are assumed to be conditionally independent given (u, w). Its logarithm is given by

$$\begin{aligned}\log p(\mathbf{A}\,|\,\theta) = &\textstyle\sum_{e\in\Omega} -\frac{1}{\kappa_e} \sum_{i<j\in e} \mathbf{u_i}^T w\, u_j \\ &+ \textstyle\sum_{e\in E} A_e \log \sum_{i<j\in e} \mathbf{u_i}^T w\, u_j\end{aligned} \tag{4}$$

where we discarded constant terms not depending on the parameters. The first summation over $|\Omega|$ terms appears intractable due to the exploding size of the configuration space. However, one important feature of our model is that this high dimensionality can be treated analytically, as the likelihood conveniently simplifies. The summand $\sum_{e\in\Omega} -\frac{1}{\kappa_e}\sum_{i<j\in e}\mathbf{u_i}^T w\, u_j$ is simply taking the interaction term $\mathbf{u_i}^T w u_j$ as many times as it appears in all the possible hyperedges, each weighted by the factor $1/\kappa_e$. This reasoning yields the count $C = \sum_{n=2}^{D} \frac{1}{\kappa_n}\binom{N-2}{n-2}$ and the following simplified log-likelihood

$$\begin{aligned}\log p(\mathbf{A}\,|\,\theta) = &-C \textstyle\sum_{i<j\in V} \mathbf{u_i}^T w\, u_j \\ &+ \textstyle\sum_{e\in E} A_e \log \sum_{i<j\in e} \mathbf{u_i}^T w\, u_j\end{aligned} \tag{5}$$

obtaining a tractable sum of terms. To maximize Eq. 5 with respect to u and w, we use a standard variational approach via Jensen's inequality $\log\mathbb{E}[x] \geq \mathbb{E}[\log x]$ to lower bound the second summand as

$$\begin{aligned}&\textstyle\sum_{e\in E} A_e \log \sum_{i<j\in e} \mathbf{u_i}^T w\, u_j \geq \\ &\textstyle\sum_{e\in E} A_e \sum_{i<j\in e} \sum_{k,q=1}^{K} \rho_{ijkq}^{(e)} \log\left(\frac{u_{ik}\, u_{jq}\, w_{kq}}{\rho_{ijkq}^{(e)}}\right)\end{aligned} \tag{6}$$

Here, the variational distribution is specified by the $\rho_{ijkq}^{(e)}$ values, which can be any configuration of strictly positive probabilities such that $\sum_{i<j\in e}\sum_{k,q=1}^{K}\rho_{ijkq}^{(e)} = 1$. The equality in Eq. 6 is achieved when

$$\rho_{ijkq}^{(e)} = \frac{u_{ik}u_{jq}w_{kq}}{\sum_{i<j\in e}\sum_{k,q=1}^{K} u_{ik}u_{jq}w_{kq}} = \frac{u_{ik}u_{jq}w_{kq}}{\lambda_e} \tag{7}$$

Hence, maximizing $\log p(\mathbf{A}|\theta)$ is equivalent to maximizing

$$\begin{aligned}\mathcal{L}(u, w, \rho) = &-C \textstyle\sum_{i<j\in V} \mathbf{u_i}^T w\, u_j \\ &+ \textstyle\sum_{e\in E} A_e \sum_{i<j\in e} \sum_{k,q=1}^{K} \rho_{ijkq}^{(e)} \log\left(\frac{u_{ik}\, u_{jq}\, w_{kq}}{\rho_{ijkq}^{(e)}}\right)\end{aligned}$$

with respect to both (u, w) and ρ. This can be done by alternating between updating ρ and (u, w), as in the expectation-maximization (EM) algorithm.

The update for $\theta \in \{u, w\}$ is obtained by setting the partial derivative $\partial\mathcal{L}(\theta, \rho)/\partial\theta$ to 0, which yields the following expressions

$$u_{ik} = \frac{\sum_{e\in E: i\in e} A_e \rho_{ik}^{(e)}}{C \sum_q w_{kq} \sum_{j\neq i\in V} u_{jq}} \tag{8}$$

$$w_{kq} = \frac{\sum_{e\in E} A_e \rho_{kq}^{(e)}}{C \sum_{i<j\in V} u_{ik}u_{jq}} \tag{9}$$

The terms $\rho_{ik}^{(e)}, \rho_{kq}^{(e)}$ are defined as

$$\rho_{ik}^{(e)} = \sum_{j \in e: j \neq i} \sum_{q} \rho_{ijkq}^{(e)}$$

$$\rho_{kq}^{(e)} = \sum_{i<j \in e} \rho_{ijkq}^{(e)}$$

and obtained after updating $\rho_{ijkq}^{(e)}$ according to Eq. 7. These updates presented in this section are based on maximum likelihood estimation, where we do not set any prior for (u, w). However, we can get maximum a posteriori estimates (MAP) with similar derivations and complexity by arbitrarily setting prior distributions for the parameters, as we show in the Supplementary Materials [Appendix Maximum-a-Posteriori (MAP) estimation]. We comment on how to obtain efficient matrix operations that implement the updates in Eqs. 8 and 9 in the "Practical implementation and efficiency" section.

Identifiability, interpretation, and theoretical implications

In the following, we make some observations on relevant aspects regarding the identifiability, interpretation, and theoretical implications of the proposed generative model. First, the log-likelihood in Eq. 5 is invariant under permutations of the groups and under the rescaling $u \to c\,u$ and $w \to w/c^2$, for any constant $c > 0$. This observation may raise questions about identifiability of the parameters. However, both permutation and rescaling do not change the composition of the communities or the relative magnitude of the entries of w; thus, the mesoscale structure is not affected by them. Nevertheless, one can easily make the model identifiable by setting a prior probability on w and considering MAP estimates (see Appendix Identifiability in the Supplementary Materials for details).

Second, for similar invariance reasons, the constant C can be neglected and absorbed after convergence, by either rescaling $u' = \sqrt{C}\,u$ or $w' = C\,w$. While the forms of the rescaling constants κ_e play no role during inference, as they only enter the updates through the C term, they do instead affect the generative process when sampling hypergraphs from it (*52*). For instance, calculations similar to those in the Supplementary Materials (Appendix Average degree) allow getting a closed-form expression for the average weighted degree when only considering interactions of size k. The resulting formula $\mathbb{E}[d_k^w] = \binom{N-2}{k-2} \frac{k}{\kappa_k N} \sum_{i<j \in V} \mathbf{u_i}^T\, w\, u_j$ shows that rescaling the constant κ_k translates into a rescaling of the average degree. Similar considerations apply to the expected number of hyperedges of a given size and show that the normalization constants κ_e play an important role in determining the expected statistics of the model and hence of the samples they produce. Generally, the sampling procedure from the generative model in Eq. 3, allows determining the degree sequence (i.e., the degree array of the single nodes) as well as the size sequence (i.e., the count of hyperedges for every specified size), which depend on the Poisson parameters and hence on the κ_e normalizers. Alternatively, the sampling procedure from our generative model can be conditioned to respect such sequences (*52*).

Third, it is possible to obtain the analytical expressions of the expected degree of a node i, which evaluates to

$$\mathbb{E}[d_i^w] = \sum_{e \in \Omega: i \in e} \mathbb{E}[A_e]$$
$$= C \mathbf{u_i}^T w \sum_{j \in V: j \neq i} u_j + C' \sum_{j<m \in V: j,m \neq i} u_j^T\, w u_m$$

where $C' = \sum_{d=3}^{D} \frac{\binom{N-3}{d-3}}{\kappa_d}$ is a constant similar to C (see Appendix Average degree in the Supplementary Materials). This expression has a relevant interpretation, as it reveals a fundamental difference between simple networks and higher-order systems. Since in dyadic systems $C' = 0$, we can think of the rightmost summand as a term contributing only to higher-order interactions, while the leftmost one is a shift of the expected degree coming from binary interactions only. One can also observe an analogy with networks of interactions in physical systems. In this context, the leftmost summand can be seen as a mean-field acting on node i in a cavity system where the node is hypothetically removed, while the rightmost term acts as a background field generated by all interactions involving any pair of nodes that does not include node i. This background term is peculiar to higher-order systems, as remarked above. Its presence has a relevant effect of building higher-order interactions between nodes in different groups. This can be illustrated with a simple example of a system with assortative w and node i belonging to a different community than all the other nodes. While the leftmost summand yields expected degree zero in dyadic systems, the background field allows i to form on average nonzero edges. Intuitively, this difference is due to the bilinear form in Eq. 2, which allows observing hyperedges that are not completely homogeneous, where there could be a minor fraction of nodes that are in different communities than the majority. Notice that such a generation, allowing for mixed hyperedges, is a desirable feature. On the one hand, it is appropriate to model contexts where individuals have multiple preferences and thus are expected to belong to multiple groups. On the other hand, recent work (*56*) proves the combinatorial unfeasibility of hypergraphs where all nodes exhibit majority homophily—implying rather uniform hyperedges contained in single communities—and encourages the development of more flexible generative models.

Practical implementation and efficiency

From an optimization perspective, the EM algorithm starts by initializing u and w at random and then repeatedly alternating between the Eq. 8 and Eq. 9 updates until convergence of $\mathscr{L}(u, w, \rho)$. This does not guarantee to reach the global optimum, but only a local one. In practice, one runs the algorithm several times, each time from a different random initialization, and outputs the parameters corresponding to the realization with highest log-likelihood $\mathscr{L}(u, w, \rho)$. We provide a pseudocode description of the whole inference procedure in Algorithm 1. For all our experiments, we perform MAP inference on the affinity w, setting a factorized exponential prior with rate 1, and maximum likelihood inference on the assignment u. This choice corresponds to the half-Bayesian model presented in the Supplementary Materials [Appendices Maximum-1-Posteriori (MAP) estimation and Identifiability]. The updates have linear computational cost, obtained by exploiting the sparsity of most real-world datasets with efficient matrix operations, as we show in Appendix Computational considerations in the

Supplementary Materials. Overall, the complexity scales as $O(N K + |E|)$, allowing to tackle inference on hypergraphs whose number of nodes and hyperedges was previously prohibitive (see the "Modeling of real data" section). Another advantage of our inference procedure is that it is stable and reliable for extremely large hyperedges. Because of computational and numerical constraints, previous models were also limited to considering hyperedges with maximal size $D = 25$ (*35*, *50*). As we illustrate in the "Modeling of real data" section with an Amazon and a Gene-Disease dataset, large interactions (respectively $D = 9350$ and $D = 1074$) should not be neglected as they provide useful information and substantially boost the quality of inference.

Algorithm 1: Hy-MMSBM EM inference

Input: Hypergraph A, training rounds r
Result: Inferred parameters (u, w)

BestLoglik $= -\infty$
BestParams $=$ None
`> Train model r times and choose`
`> realization with best likelihood`
for $t = 1, \ldots, r$ **do**
  `> Initialize at random`
  $u, w \leftarrow \text{init}(u, w)$
  `> convergence is attained for a max number of EM steps, or below a certain change in parameter values`
  while *not converged* **do**
    $u \leftarrow \text{update}(u)$ Eq. (8)
    $w \leftarrow \text{update}(w)$ Eq. (9)
  end
  $\text{L} = \text{loglik}(u, w)$ Eq. (5)
  if $L >$ *BestLoglik* **then**
    BestLoglik $\leftarrow$ L
    BestParams $\leftarrow (u, w)$
  end
end

RECOVERY OF GROUND-TRUTH COMMUNITIES

A standard way to assess the effectiveness of a community detection algorithm is to check if the inferred node memberships match those of a given ground truth. Such ground truth is generally not available for real-world systems (*57*), while it can be imposed as a planted configuration for synthetic data. For this reason, we consider a recently developed sampling method to produce structured synthetic hypergraphs with flexible structures specified in input (*52*). For further details, see Appendix Recovery of community assignments in the Supplementary Materials.

In Fig. 1, we generate hypergraphs with an underlying diagonal affinity matrix w (assortative structure) and show the recovery performance for the cases of hard (left) and mixed-membership (right) community assignments. The detailed description of the data generation process is provided in Appendix Recovery of community assignments in the Supplementary Materials. We compare our approach with Hypergraph-MT (*35*), an inference algorithm designed to detect overlapping community assignments and assortative interactions; Spectral Clustering (*43*), which recovers hard communities via hypergraph cut optimization; and Hypergraph AON-MLL (*50*), which performs a modularity-like optimization based on a Poisson generative model with hard memberships. For our comparisons, we compute the cosine similarity between the ground truth and the inferred communities, which is appropriate to measure the similarity for both hard and mixed-membership vectors. A value of zero represents no similarity, while a value of one is attained by completely overlapping vectors. In both cases, we find that our model successfully recovers the ground-truth communities as more information is made available in terms of hyperedges of increasing sizes. This is somehow expected because the generating process of these data reflects the one of our method, and is a sanity check of our maximum likelihood approach. Spectral Clustering and Hypergraph-MT attain comparable cosine similarity scores on hard-membership data (left), while their performances differ when detecting mixed memberships (right), with Hypergraph-MT performing better. This is because Spectral Clustering performs an approximate combinatorial search and can only recover hard communities, while Hypergraph-MT allows for overlapping communities via maximum likelihood inference. The low performance of Hypergraph AON-MLL is explained by its generative assumptions. AON-MLL assigns the same probability to all the hyperedges containing nodes from more than one community. As most of the hyperedges in this synthetic data are made of nodes from more than one community, the recovery of hypergraph modularity on such systems is close to random. Altogether, such results highlight the effectiveness of the inference procedure, making our model suitable for networked systems with higher-order interactions. Although relevant, the results in Fig. 1 are just one possible comparison among algorithms with different generative assumptions. Such assumptions are expected to yield better or worse results depending on the data, and in general, the no-free-lunch theorem implies that no algorithm will consistently outperform all others on all types of data. As a case for this argument, in Appendix Additional experiments on ground truth recovery in the Supplementary Materials, we present additional results on different synthetic data.

DETECTABILITY OF COMMUNITY CONFIGURATION

Previous inference algorithms rely on the strong assumption of assortative community interactions, hampering their ability to model more complex mesoscale patterns observed in the real world. By contrast, our model allows detecting a variety of different regimes, as it assumes a more flexible w.

Here, we investigate the detection—and detectability—of different assortative and disassortative community structures in hypergraphs, generalizing previous work on pairwise systems (*58*). In particular, we generate hypergraphs with hard community assignments and different community interactions. We take affinity matrices w with diagonal values c_{in} and out-diagonal values c_{out}, and vary both c_{in} and the ratio $c_{\text{out}}/c_{\text{in}}$. By fixing the value of $c_{\text{out}}/c_{\text{in}}$, we expect higher detectability with increasing c_{in}, as this term regulates the expected degree and consequently the information contained in the data. On the contrary, for a fixed value of c_{in}, we expect the disassortative model to attain better recovery as the ratio $c_{\text{out}}/c_{\text{in}}$ increases, due to the stronger intercommunity interactions. Details on data generation are provided in Appendix Detection of community structure in the Supplementary Materials.

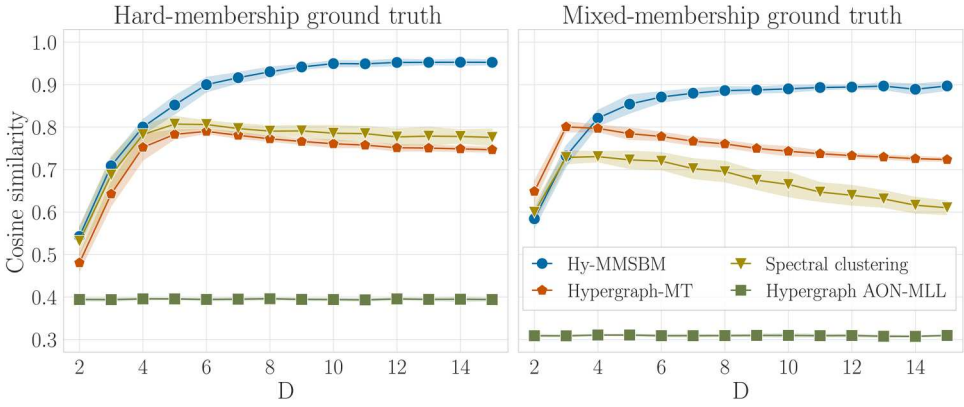

Fig. 1. Recovery of ground-truth community assignments. We measure the cosine similarity between the ground truth and the inferred assignments. We vary the maximum hyperedge size D in synthetic data and study the cases of hard (left) and mixed (right) ground-truth memberships. When information is scarce, represented by few hyperedges of small maximum size D, our method is comparable to the most efficient approaches currently available. However, as larger hyperedges are considered, our method outperforms competing algorithms, on both hard and mixed-membership planted partitions.

We compare the log-likelihoods obtained by the model when the affinity matrix w is initialized as diagonal or full, which we refer to as assortative and disassortative, respectively. Notice that the multiplicative updates in Eq. 9 guarantee that, if w is initialized as diagonal, it will remain as such during training. It is also possible that a full matrix will converge to diagonal during inference. Nonetheless, the strong bias of a diagonal initialization restricts the parameter space of the assortative model, facilitating the convergence to better optima for the detection of assortative structures.

Given the log-likelihood of the assortative ($\mathscr{L}_a$) and disassortative ($\mathscr{L}_d$) models, we measure the difference $\mathscr{L}_a - \mathscr{L}_d$ while varying the values of c_{in} and $c_{\text{out}}/c_{\text{in}}$. Positive values denote stronger performance of the assortative model, as its likelihood is higher, while negative values favor the disassortative one. We observe that the assortative model attains higher likelihood for low values of $c_{\text{out}}/c_{\text{in}}$, when within-community interactions are stronger, as shown in Fig. 2A. Its performance deteriorates as we increase $c_{\text{out}}/c_{\text{in}}$, with the disassortative one taking over with higher likelihood values. Furthermore, we can notice an inflexion point at $c_{\text{out}}/c_{\text{in}} = 1$, where the difference in likelihood between the models is null. While one would expect the disassortative model to perform better in such a scenario, we highlight that this regime is a challenging and noisy one, as the affinity matrix is the uniform matrix of ones. Hence, recovery is difficult and not guaranteed, regardless of the model. We finally notice an increase of $\mathscr{L}_a - \mathscr{L}_d$ with c_{in}, which regulates the strength of the signal and makes it easier to separate the two regimes.

While we expect recovery to improve at more detectable regimes, this may not be observed by only looking at the $\mathscr{L}_a - \mathscr{L}_d$ difference. For this reason, in Fig. 2B, we complement our analysis by plotting only the log-likelihood $\mathscr{L}_d$ attained via the disassortative initialization. In this case, we notice that the performance of the disassortative model increases with both $c_{\text{out}}/c_{\text{in}}$ and c_{in}, as the intercommunity interactions get stronger and the expected degree gets higher. Altogether, our algorithm provides a principled way to extract arbitrary community interactions from higher-order data with varying structural organizations.

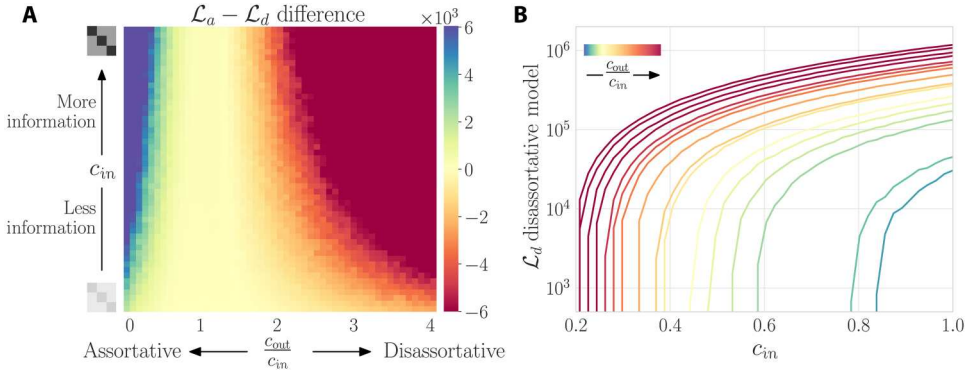

Fig. 2. Detection of assortative and disassortative community interactions. We generate data where the affinity matrices contain diagonal values c_{in} and out-diagonal c_{out} and measure the ability of our model to detect different assortative and disassortative regimes. (**A**) Positive (negative) differences in log-likelihood values indicate that the assortative (disassortative) model attains a better fit. An intermediate regime, highlighted in yellow, also emerges. Here, the detectability is compromised due to not having enough structure ($c_{\text{out}} \approx c_{\text{in}}$) or enough information (low c_{in}). (**B**) Log-likelihood of the disassortative model. In this case, the model attains better fit for data with marked disassortative structure (darker red).

CORE-PERIPHERY STRUCTURE

Many real-world systems are characterized by a different mesoscale organization known as core-periphery (CP) structure (*59*, *60*). Networks characterized by such structure present a group of core of nodes connected among themselves, and often with high degree (*61*, *62*), and a separate periphery of weakly connected nodes. Recently, methods to study and detect the existence of such patterns in hypergraphs have been proposed (*63*, *64*). Conceptually, Hy-MMSBM has not been developed with the purpose of CP detection. Nevertheless, we can show its ability in capturing CP structures in hypergraphs through the generation of synthetic data that resemble the core structures of the input dataset.

To measure the recovery of CP structures, we use the method developed by Tudisco and Higham (*64*), HyperNSM, that assigns to each node of a hypergraph a core-score quantifying how close the node is to the core, where higher values denote stronger participation. HyperNSM achieved good performance on synthetic and real-world data, and its implementation is extremely efficient.

We analyze the Enron email dataset (*65*). Notably, the dataset comes with metadata information identifying a group of core nodes, employees of the organization who send batch emails to the periphery, which in turn only receive emails. This allows us to evaluate the ability of a model to recover a CP structure. In our study, we use the dataset used by Tudisco and Higham (*64*) with a planted core set that arises directly from the data collection process, as discussed by Amburg *et al.* (*63*) (it is preprocessed by keeping only hyperedges of size $D \leq 25$). The dataset has $N = 4423$ nodes and a core composed by 132 nodes. We apply HyperNSM to quantify the CP structure of the input Enron email dataset, as well as of the samples generated with Hy-MMSBM. To generate the samples, we first run our inference procedure on the Enron email dataset and then sample hypergraphs distributed according to the obtained u, w parameters. Further details on how to generate the samples are provided in Appendix Core-periphery experiments in the Supplementary Materials. For comparison, we also generate samples with a configuration model for hypergraphs (*66*) and obtain their core-score vectors with HyperNSM as well.

To evaluate the quality of the CP assignments in the different samples, we use the CP profile, the metric defined in (*64*) as

$$\gamma(S) = \frac{\#\text{ hyperedges with all nodes in } S}{\#\text{ hyperedges with at least one node in } S}, S \subseteq V \quad (10)$$

For any $k \in \{1, \ldots, N\}$, we calculate the value $\gamma[S_k(x)]$, where $S_k(x)$ is the set of k nodes with smallest core-score in x. Given its definition, $\gamma(S)$ is small if S is largely contained in the periphery of the hypergraph and it should increase drastically as k crosses some threshold value k_0, which indicates that the nodes in $V \backslash S_{k_0}(x)$ form the core.

In Fig. 3A we show the CP profiles corresponding to the core-scores computed with HyperNSM on the different datasets, i.e., the input Enron email, the samples generated with Hy-MMSBM, and the samples generated with the configuration model for hypergraphs. We plot 600 nodes with the highest core-score in decreasing order, and for all datasets, we notice a sharp drop, which highlights the existence of a CP structure. The main difference is given by the threshold k_0 at which this drop happens. This determines the dimension of the core. Remember that the data have a core composed by 132 nodes, and when applying HyperNSM on the input data, we obtain a core dimension equal to 117, validating the good core-detection performance of this algorithm. The samples generated with the configuration model present a core with an average of 530.6 nodes, quite far from what observed in the input dataset. On the other hand, Hy-MMSBM generates samples that better resemble the property of the Enron email dataset, with an average core dimension of 195.7 nodes.

To understand the impact of nonpairwise interactions on higher-order CP structure, we also study the connection between hyperedge size and CP score. In Fig. 3B, we plot the CP score of a given node against the mean size of the hyperedges it belongs to. While we can observe a strong relationship between these two quantities at low CP scores, such regularity disappears in the center of the plot, which contains core nodes and presents a high scattering of hyperedge size values. This unexplained variance is justified by the rich information encoded in the CP score, which jointly depends on different factors related to the topology of the hypergraph. Yet, the scatter plots obtained on the Enron email dataset and the samples generated with Hy-MMSBM have higher similarity than the samples generated with the configuration model. Quantitatively, we measure the similarity between the core-scores of the different datasets for the 132 core nodes with the Pearson correlation, a measure $\rho \in [-1,1]$ of linear correlation between two sets of data. The CP scores of the data have a Pearson correlation equal to 0.81 ± 0.01 with the samples generated with Hy-MMSBM, and of 0.76 ± 0.03 with the samples generated with the configuration model. Similar results are found on the relation between CP score and another structural property, namely, the degree of a node (see fig. S2 in Appendix Additional results on the Enron email dataset in the Supplementary Materials).

MODELING OF REAL DATA

In this section, we perform an extensive investigation of higher-order real-world systems. As explained in the "Inference" section and in the Supplementary Materials (Appendix Computational considerations), the linear-cost EM updates, together with a careful implementation that exploits the sparsity of most datasets, make our method suitable for the analysis of a variety of hypergraphs that were previously inaccessible due to computational constraints. Our method proves to be scalable with respect to both the number of system units and the size of the interactions, improving substantially on competing algorithms currently available in the literature. Moreover, our model is based on a probabilistic formulation, allowing it to perform additional operations and extract information that is not viable via other approaches, such as spectral clustering. First, we evaluate the quality of fit of various community detection methods based on their hyperedge prediction capabilities on a Gene Disease dataset, where nodes are genes, and interactions contain genes that are associated with a disease. To this end, we use the area under the curve (AUC) measure, a link prediction metric defined as follows: Given a randomly selected observed edge, and a randomly selected nonobserved one, the AUC $\in [0,1]$ computes the number of times that the generative model assigns a higher probability to the observed edge. Here, we split the datasets into train and test subsets, where the train sets are used to estimate the parameters, and we evaluate the prediction performance in terms of AUC on the test sets (see Appendix Experiments on real data in the

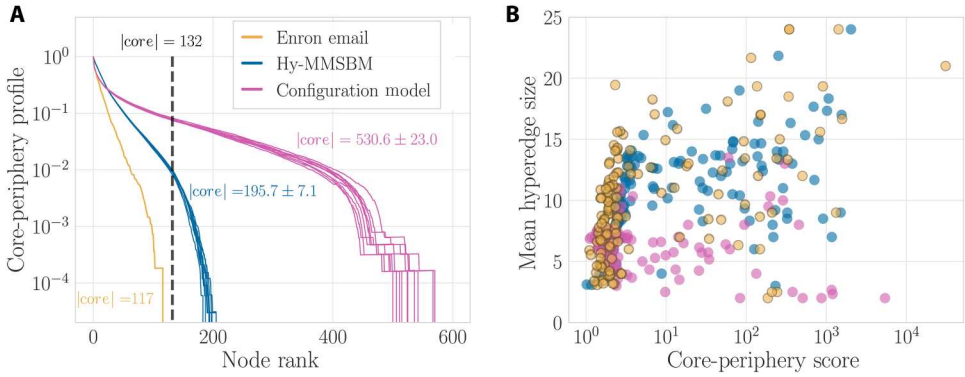

Fig. 3. Recovery of structural CP information. (**A**) CP profile (Eq. 10) corresponding to the core-scores computed with HyperNSM on the input Enron email (yellow), 10 synthetic samples generated with Hy-MMSBM (blue), and 10 synthetic samples generated with a configuration model for hypergraphs (magenta). We plot 600 nodes with the highest core-score in decreasing order and report the averages and standard deviations of the core dimension for the different datasets. Our method generates samples that closely resemble the property of the input dataset, with an average core dimension close to 132 nodes. (**B**) Mean size of the hyperedges a node belongs to against its CP score. We observe higher agreement between the data and the inference-based sample generated with Hy-MMSBM. This is also highlighted by the Pearson correlation of the 132 core nodes that is equal to 0.81 ± 0.01 for Hy-MMSBM versus the value of 0.76 ± 0.03 for the samples generated with the configuration model.

Supplementary Materials for details). Scalability with respect to hyperedge size is a crucial aspect of models for higher-order data. However, due to computational and numerical constraints, previous methods are limited to considering interactions of moderate size only, possibly causing a loss of information and a biased representation of the full system. In contrast, our model is able to efficiently process all the information provided in the dataset, reliably scaling to hyperedges of size of the order of the thousands. In Fig. 4A, we compare our method with other probabilistic approaches with hyperedge prediction capabilities. When only small interactions are considered, our model outperforms the competitive algorithms. At the computational limit of other approaches $D = 25$, Hypergraph-MT and our model attain a similar score, signaling the importance of considering large interactions. Beyond this computational threshold, our method continues to exploit the information provided by interactions among a growing number of units up to the maximum size observed of $D = 1074$, which results in an AUC score of 0.79.

We then extend our analysis to a variety of datasets from different domains, as described in Fig. 4B. For each dataset, we show the inference running time as a function of the number of nodes N and the size of the largest hyperedge D. The AUC scores, reported in Table 1 and ranging from 0.74 to 0.98, show that the model generally yields a good fit and predicts the existence of hyperedges reliably. While these scores are on average aligned with those of other existing algorithms (*35*), the running time of our model is orders of magnitude lower. This allows studying very large hypergraphs such as the Arxiv, Trivago 2core, and Amazon datasets, containing up to millions of nodes and hyperedges. Overcoming the resulting computational challenges, our method allows the efficient modeling of a variety of previously unexplored datasets, which, to the best of our knowledge, could not be tackled by competing higher-order community detection algorithms.

Taken all together, these results show the effectiveness of our model in tackling datasets of small and large dimensions, in terms of both quantitative performance and computational scalability, and make Hy-MMSBM a valid tool for the study of complex higher-order systems.

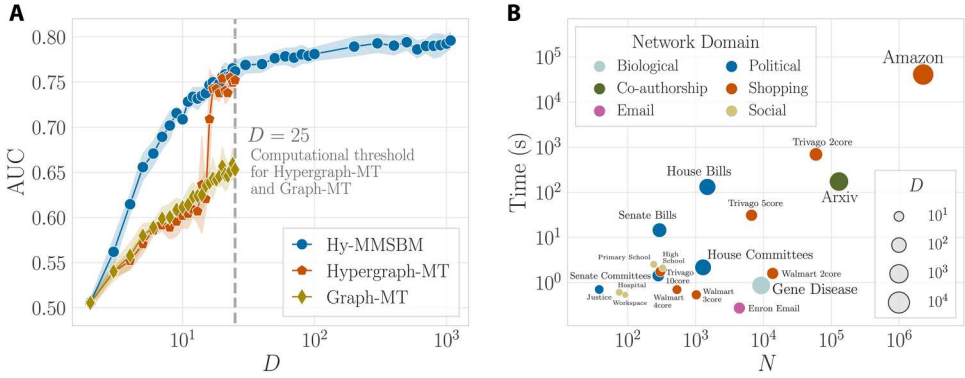

Fig. 4. Modeling of real data: hyperedge prediction and running time. (**A**) Quality of hyperedge prediction measured by the AUC score on a Gene Disease dataset, where nodes are genes and hyperedges contain genes that are associated with a disease. For Hypergraph-MT and Graph-MT, the plot shows a computational threshold at the maximum hyperedge size $D = 25$. Hy-MMSBM attains the highest scores and is able to model the entire hypergraph, up to $D = 1074$. (**B**) Running time of Hy-MMSBM for a variety of real-world datasets. The node represents the data domain. Both N and D are in log scale. The corresponding AUC scores are reported in Table 1.

DISCUSSION

Here, we have developed a probabilistic framework to model hypergraphs. Our method allows performing inference on very large hypergraphs, detecting their community structure, and reliably predicting the existence of higher-order interactions of arbitrary size. When compared to other available methods on synthetic hypergraphs with known ground truth, for both hard and mixed assignments, our model attains the most efficient recovery of the planted partitions. Moreover, compared to previous proposals, Hy-MMSBM relies on less restrictive assumptions on the latent community structure in the data and is thus able to detect configurations, such as disassortative community interactions, which could not be previously identified. Furthermore, our method is extremely fast. Its efficient numerical implementation exploits optimized closed-form updates and dataset sparsity and has linear cost in the number of nodes and hyperedges. The resulting formulas are also numerically stable, not resulting in under- or overflows during the computations. Such numerical stability carries over to extremely large interactions, a substantial improvement over the computational threshold of previous methods, allowing to explore higher-order datasets with millions of nodes and interactions among thousands of units, that could not be previously tackled.

There are several directions for future work. From a theoretical perspective, our proposed likelihood function is based on a bilinear form for capturing dependencies within the hyperedges, a key ingredient for ensuring both mixed-membership nodes and fast inference. A possible extension would be to consider alternative likelihood definitions where the probability of the hyperedges is determined by multilinear forms, which would in principle allow capturing more complex interactions within the hyperedges. Similarly, here, we have assumed the hyperedges to be independent conditioned on the latent variables. Relaxing this assumption may ameliorate the expressiveness of the model, allowing to capture topological properties that involve more than two hyperedges, as already observed in the case of networks (*67*–*69*). From an algorithmic perspective, there are different questions that may allow further stabilizing and improving the inference procedure. Among these, the propensity of different initial conditions to be trapped in local optima during EM or MAP inference has not yet been investigated. Devising suitable initialization procedures or parameter priors to favor different membership types, as done in other works (*70*), offers a promising path in this direction. Finally, we have considered here a standard scenario where the input data are a list of hyperedges, and these are provided all at once. Other approaches may be needed in case of availability of extra information such as node attributes (*71*, *72*) or for dynamic data (*73*).

Altogether, our work provides an accurate, flexible, and scalable tool for the modeling of very large hypergraphs, advancing our ability to tackle and study the organization of real-world higher-order systems.

Table 1. AUC scores on real datasets. We report the number of nodes *N*, number of hyperedges |*E*|, maximum hyperedge size *D*, number of communities *K*, and AUC scores attained by our method on 19 large-scale real-world hypergraphs. The results are averages and standard deviations over 10 random test sets, and the value of *K* is chosen via cross-validation (see Appendix Experiments on real data in the Supplementary Materials).

	N	|*E*|	*D*	*K*	AUC
Justice	38	2,826	9	4	0.909 ± 0.008
Hospital	75	1,825	5	2	0.767 ± 0.013
Workspace	92	788	4	5	0.741 ± 0.015
Primary School	242	12,704	5	10	0.832 ± 0.002
Senate Committees	282	301	31	30	0.926 ± 0.023
Senate Bills	294	21,721	99	13	0.921 ± 0.002
Trivago 10core	303	3,162	14	11	0.960 ± 0.005
High School	327	7,818	5	17	0.879 ± 0.007
Walmart 4core	532	2,292	10	4	0.837 ± 0.013
Walmart 3core	1,025	3,553	11	4	0.825 ± 0.010
House Committees	1,290	335	81	25	0.939 ± 0.015
House Bills	1,494	54,933	399	19	0.946 ± 0.001
Enron Email	4,423	5,734	25	2	0.835 ± 0.009
Trivago 5core	6,687	33,963	26	30	0.962 ± 0.001
Gene Disease	9,262	3,128	1,074	2	0.828 ± 0.010
Walmart 2core	13,706	19,869	25	2	0.788 ± 0.004
Trivago 2core	59,536	140,698	52	100	0.863 ± 0.002
Arxiv	130,024	172,173	2,097	10	0.884 ± 0.001
Amazon	2,268,231	4,242,421	9,350	29	0.978 ± 0.002

Supplementary Materials

This PDF file includes:

Supplementary Text
Figs. S1 and S2
References

REFERENCES AND NOTES

1. S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.-U. Hwang, Complex networks: Structure and dynamics. *Phys. Rep.* **424**, 175–308 (2006).
2. F. Battiston, G. Cencetti, I. Iacopini, V. Latora, M. Lucas, A. Patania, J. G. Young, G. Petri, Networks beyond pairwise interactions: Structure and dynamics. *Phys. Rep.* **874**, 1–92 (2020).
3. L. Torres, A. S. Blevins, D. Bassett, T. Eliassi-Rad, The why, how, and when of representations for complex systems. *SIAM Rev.* **63**, 435–485 (2021).
4. F. Battiston, E. Amico, A. Barrat, G. Bianconi, G. Ferraz de Arruda, B. Franceschiello, I. Iacopini, S. Kéfi, V. Latora, Y. Moreno, M. M. Murray, T. P. Peixoto, F. Vaccarino, G. Petri, The physics of higher-order interactions in complex systems. *Nat. Phys.* **17**, 1093–1098 (2021).
5. F. Battiston, G. Petri, *Higher-Order Systems* (Springer, 2022).

6. A. Patania, G. Petri, F. Vaccarino, The shape of collaborations. *EPJ Data Sci.* **6**, 18 (2017).
7. S. Klamt, U.-U. Haus, F. Theis, Hypergraphs and cellular networks. *PLOS Comput. Biol.* **5**, e1000385 (2009).
8. A. Zimmer, I. Katzir, E. Dekel, A. E. Mayo, U. Alon, Prediction of multidimensional drug dose responses based on measurements of drug pairs. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 10442–10447 (2016).
9. G. Cencetti, F. Battiston, B. Lepri, M. Karsai, Temporal properties of higher-order interactions in social networks. *Sci. Rep.* **11**, 7028 (2021).
10. F. Musciotto, D. Papageorgiou, F. Battiston, D. R. Farine, Beyond the dyad: Uncovering higher-order structure within cohesive animal groups. bioRxiv 2022.05.30.494018 [Preprint]. 30 May 2022. https://doi.org/10.1101/2022.05.30.494018.
11. G. Petri, P. Expert, F. Turkheimer, R. Carhart-Harris, D. Nutt, P. J. Hellyer, F. Vaccarino, Homological scaffolds of brain functional networks. *J. R. Soc. Interface* **11**, 20140873 (2014).
12. C. Giusti, R. Ghrist, D. S. Bassett, Two's company, three (or more) is a simplex. *J. Comput. Neurosci.* **41**, 1–14 (2016).
13. A. Santoro, F. Battiston, G. Petri, E. Amico, Higher-order organization of multivariate time series. *Nat. Phys.* **19**, 1–9 (2023).
14. T. Carletti, F. Battiston, G. Cencetti, D. Fanelli, Random walks on hypergraphs. *Phys. Rev. E* **101**, 022308 (2020).
15. C. Bick, P. Ashwin, A. Rodrigues, Chaos in generically coupled phase oscillator networks with nonpairwise interactions. *J. Nonlin. Sci.* **26**, 094814 (2016).
16. P. S. Skardal, A. Arenas, Higher order interactions in complex networks of phase oscillators promote abrupt synchronization switching. *Commun. Phys.* **3**, 218 (2020).
17. A. P. Millán, J. J. Torres, G. Bianconi, Explosive higher-order kuramoto dynamics on simplicial complexes. *Phys. Rev. Lett.* **124**, 218301 (2020).
18. M. Lucas, G. Cencetti, F. Battiston, Multiorder laplacian for synchronization in higher-order networks. *Phys. Rev. Res.* **2**, 033410 (2020).
19. L. V. Gambuzza, F. Di Patti, L. Gallo, S. Lepri, M. Romance, R. Criado, M. Frasca, V. Latora, S. Boccaletti, Stability of synchronization in simplicial complexes. *Nat. Commun.* **12**, 1255 (2021).
20. Y. Zhang, M. Lucas, F. Battiston, Higher-order interactions shape collective dynamics differently in hypergraphs and simplicial complexes. *Nat. Commun.* **14**, 1605 (2023).
21. I. Iacopini, G. Petri, A. Barrat, V. Latora, Simplicial models of social contagion. *Nat. Commun.* **10**, 2485 (2019).
22. S. Chowdhary, A. Kumar, G. Cencetti, I. Iacopini, F. Battiston, Simplicial contagion in temporal higher-order networks. *J. Phys. Complex.* **2**, 035019 (2021).
23. L. Neuhäuser, A. Mellor, R. Lambiotte, Multibody interactions and nonlinear consensus dynamics on networked systems. *Phys. Rev. E* **101**, 032310 (2020).
24. U. Alvarez-Rodriguez, F. Battiston, G. F. de Arruda, Y. Moreno, M. Perc, V. Latora, Evolutionary dynamics of higher-order interactions in social networks. *Nat. Hum. Behav.* **5**, 586–595 (2021).
25. A. Civilini, N. Anbarci, V. Latora, Evolutionary game model of group choice dilemmas on hypergraphs. *Phys. Rev. Lett.* **127**, 268301 (2021).
26. A. Civilini, O. Sadekar, F. Battiston, J. Gómez-Gardeñes, V. Latora, Explosive cooperation in social dilemmas on higher-order networks. arXiv:2303.11475 [physics.soc-ph] (20 March 2023).
27. C. Berge, *Graphs and Hypergraphs* (North-Holland Pub. Co., 1973).
28. A. R. Benson, Three hypergraph eigenvector centralities. *SIAM J. Math. Data Sci.* **1**, 293–312 (2019).
29. F. Tudisco, D. J. Higham, Node and edge nonlinear eigenvector centrality for hypergraphs. *Commun. Phys.* **4**, 201 (2021).
30. A. R. Benson, R. Abebe, M. T. Schaub, A. Jadbabaie, J. Kleinberg, Simplicial closure and higher-order link prediction. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E11221–E11230 (2018).
31. Q. F. Lotito, F. Musciotto, A. Montresor, F. Battiston, Higher-order motif analysis in hypergraphs. *Commun. Phys.* **5**, 79 (2022).
32. Q. F. Lotito, F. Musciotto, F. Battiston, A. Montresor, Exact and sampling methods for mining higher-order motifs in large hypergraphs. arXiv:2209.10241 [cs.SI] (21 September 2022).
33. F. Musciotto, F. Battiston, R. N. Mantegna, Detecting informative higher-order interactions in statistically validated hypergraphs. *Commun. Phys.* **4**, 218 (2021).
34. F. Musciotto, F. Battiston, R. N. Mantegna, Identifying maximal sets of significantly interacting nodes in higher-order networks. arXiv:2209.12712 [physics.soc-ph] (26 September 2022).
35. M. Contisciani, F. Battiston, C. De Bacco, Inference of hyperedges and overlapping communities in hypergraphs. *Nat. Commun.* **13**, 7229 (2022).
36. J.-G. Young, G. Petri, T. P. Peixoto, Hypergraph reconstruction from network data. *Commun. Phys.* **4**, 135 (2021).
37. K. Balasubramanian, D. Gitelman, H. Liu, Nonparametric modeling of higher-order interactions via hypergraphons. *J. Mach. Learn. Res.* **22**, 146 (2021).
38. Z. T. Ke, F. Shi, D. Xia, Community detection for hypergraph networks via regularized tensor power iteration. arXiv:1909.06503 [stat.ME] (14 September 2019).
39. K. Turnbull, S. Lunagomez, C. Nemeth, E. Airoldi, Latent space modelling of hypergraph data. arXiv:1909.00472 [stat.ME] (1 September 2019).
40. T. L. J. Ng, T. B. Murphy, Model-based clustering for random hypergraphs. *Adv. Data Anal. Classif.* **16**, 691–723 (2022).
41. T. Carletti, D. Fanelli, R. Lambiotte, Random walks and community detection in hypergraphs. *J. Phys. Complex.* **2**, 015011 (2021).
42. A. Eriksson, D. Edler, A. Rojas, M. de Domenico, M. Rosvall, How choosing random-walk model and network representation matters for flow-based community detection in hypergraphs. *Commun. Phys.* **4**, 133 (2021).
43. D. Zhou, J. Huang, B. Schölkopf, Learning with hypergraphs: Clustering, classification, and embedding. *Adv. Neural Inf. Process. Syst.* **19**, 1601–1608 (2006).
44. D. Ghoshdastidar, A. Dukkipati, A provable generalized tensor spectral method for uniform hypergraph partitioning, in *International Conference on Machine Learning* (PMLR, 2015), pp. 400–409.
45. M. C. Angelini, F. Caltagirone, F. Krzakala, L. Zdeborová, Spectral detection on sparse hypergraphs, in *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)* (IEEE, 2015), pp. 66–73.
46. X. Gong, D. J. Higham, K. Zygalakis, Generative hypergraph models and spectral embedding. *Sci. Rep.* **13**, 540 (2023).
47. D. Ghoshdastidar, A. Dukkipati, Consistency of spectral partitioning of uniform hypergraphs under planted partition model. *Adv. Neural Inf. Process. Syst.* **27**, (2014).
48. C.-Y. Lin, I. E. Chien, I.-H. Wang, On the fundamental statistical limit of community detection in random hypergraphs, in *2017 IEEE International Symposium on Information Theory (ISIT)* (IEEE, 2017), pp. 2178–2182.
49. K. Ahn, K. Lee, C. Suh, Community recovery in hypergraphs. *IEEE Trans. Inf. Theory* **65**, 6561–6579 (2019).
50. P. S. Chodrow, N. Veldt, A. R. Benson, Generative hypergraph clustering: From blockmodels to modularity. *Sci. Adv.* **7**, eabh1303 (2021).
51. L. Brusa, C. Matias, Model-based clustering in simple hypergraphs through a stochastic blockmodel. arXiv:2210.05983 [stat.ME] (12 October 2022).
52. N. Ruggeri, F. Battiston, C. De Bacco, A framework to generate hypergraphs with community structure. arXiv:2212.08593 [cs.SI] (22 June 2023).
53. E. M. Airoldi, D. Blei, S. Fienberg, E. Xing, Mixed membership stochastic blockmodels. *Adv. Neural Inf. Process. Syst.* **9**, 1981–2014 (2008).
54. C. De Bacco, E. A. Power, D. B. Larremore, C. Moore, Community detection, link prediction, and layer interdependence in multilayer networks. *Phys. Rev. E* **95**, 042317 (2017).
55. Q. F. Lotito, M. Contisciani, C. de Bacco, L. di Gaetano, L. Gallo, A. Montresor, F. Musciotto, N. Ruggeri, F. Battiston, Hypergraphx: A library for higher-order network analysis. *Journal of Complex Networks* **11**, cnad019 (2023).
56. N. Veldt, A. R. Benson, J. Kleinberg, Combinatorial characterizations and impossibilities for higher-order homophily. *Sci. Adv.* **9**, eabq3200 (2023).
57. L. Peel, D. B. Larremore, A. Clauset, The ground truth about metadata and community detection in networks. *Sci. Adv.* **3**, e1602548 (2017).
58. A. Decelle, F. Krzakala, C. Moore, L. Zdeborová, Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Phys. Rev. E* **84**, 066106 (2011).
59. S. P. Borgatti, M. G. Everett, Models of core/periphery structures. *Soc. Netw.* **21**, 375–395 (2000).
60. P. Csermely, A. London, L.-Y. Wu, B. Uzzi, Structure and dynamics of core/periphery networks. *J. Complex Netw.* **1**, 93–123 (2013).
61. V. Colizza, A. Flammini, M. A. Serrano, A. Vespignani, Detecting rich-club ordering in complex networks. *Nat. Phys.* **2**, 110–115 (2006).
62. A. Ma, R. J. Mondragón, Rich-cores in networks. *PLOS ONE* **10**, e0119678 (2015).
63. I. Amburg, J. Kleinberg, A. R. Benson, Planted hitting set recovery in hypergraphs. *J. Phys. Complex* **2**, 035004 (2021).
64. F. Tudisco, D. J. Higham, Core-periphery detection in hypergraphs. *SIAM J. Math. Data Sci.* **5**, 1–21 (2023).
65. B. Klimt, Y. Yang, *European Conference on Machine Learning* (Springer, 2004), pp. 217–226.
66. P. S. Chodrow, Configuration models of random hypergraphs. *Networks* **8**, cnaa018 (2020).
67. H. Safdari, M. Contisciani, C. De Bacco, Generative model for reciprocity and community detection in networks. *Phys. Rev. Res.* **3**, 023209 (2021).
68. M. Contisciani, H. Safdari, C. De Bacco, Community detection and reciprocity in networks by jointly modelling pairs of edges. *Networks* **10**, cnac034 (2022).
69. H. Safdari, M. Contisciani, C. De Bacco, Reciprocity, community detection, and link prediction in dynamic networks. *J. Phys. Complex* **3**, 015010 (2022).

70. N. Nakis, A. Çelikkanat, M. Mørup, *Complex Networks and Their Applications XI: Proceedings of The Eleventh International Conference on Complex Networks and Their Applications: COMPLEX NETWORKS 2022–Volume 1* (Springer, 2023), pp. 350–363.
71. M. E. Newman, A. Clauset, Structure and inference in annotated networks. *Nat. Commun.* **7**, 11863 (2016).
72. M. Contisciani, E. A. Power, C. De Bacco, Community detection with node attributes in multilayer networks. *Sci. Rep.* **10**, 15736 (2020).
73. C. Matias, V. Miele, Statistical clustering of temporal networks through a dynamic stochastic block model. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **79**, 1119–1141 (2017).
74. E. L. Lehmann, G. Casella, *Theory of Point Estimation* (Springer Science & Business Media, 2006).
75. D. B. Larremore, A. Clauset, A. Z. Jacobs, Efficiently inferring community structure in bipartite networks. *Phys. Rev. E* **90**, 012805 (2014).
76. N. W. Landry, M. Lucas, I. Iacopini, G. Petri, A. Schwarze, A. Patania, L. Torres, Xgi: A python package for higher-order interaction networks. *J. Open Source Softw.* **8**, 5162 (2023).

Acknowledgments
Funding: N.R. acknowledges support from the Max Planck ETH Center for Learning Systems. M.C. and C.D.B. were supported by the Cyber Valley Research Fund. M.C. acknowledges support from the International Max Planck Research School for Intelligent Systems (IMPRS-IS). F.B. acknowledges support from the Air Force Office of Scientific Research under award number FA8655-22-1-7025. **Author contributions:** All authors conceived the project. N.R. developed the code implementation and performed the simulations and analysis. All the authors contributed to the development of models and experiments and to the writing and revision of the paper. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All synthetic data needed to evaluate the conclusions of the paper are explained in detail for reproduction. All real data are properly referenced and publicly available.

Submitted 30 January 2023
Accepted 12 June 2023
Published 12 July 2023
10.1126/sciadv.adg9159

ScienceAdvances

Community detection in large hypergraphs

Nicol Ruggeri, Martina Contisciani, Federico Battiston, and Caterina De Bacco

Sci. Adv., **9** (28), eadg9159.
DOI: 10.1126/sciadv.adg9159