Revealing Consensus and Dissensus between Network Partitions

Tiago P. Peixoto[*]

Department of Network and Data Science, Central European University, 1100 Vienna, Austria,
ISI Foundation, Via Chisola 5, 10126 Torino, Italy,
and Department of Mathematical Sciences, University of Bath,
Claverton Down, Bath BA2 7AY, United Kingdom

(Received 23 June 2020; revised 22 December 2020; accepted 12 January 2021; published 5 April 2021)

Community detection methods attempt to divide a network into groups of nodes that share similar properties, thus revealing its large-scale structure. A major challenge when employing such methods is that they are often degenerate, typically yielding a complex landscape of competing answers. As an attempt to extract understanding from a population of alternative solutions, many methods exist to establish a consensus among them in the form of a single partition "point estimate" that summarizes the whole distribution. Here, we show that it is, in general, not possible to obtain a consistent answer from such point estimates when the underlying distribution is too heterogeneous. As an alternative, we provide a comprehensive set of methods designed to characterize and summarize complex populations of partitions in a manner that captures not only the existing consensus but also the dissensus between elements of the population. Our approach is able to model mixed populations of partitions, where multiple consensuses can coexist, representing different competing hypotheses for the network structure. We also show how our methods can be used to compare pairs of partitions, how they can be generalized to hierarchical divisions, and how they can be used to perform statistical model selection between competing hypotheses.

DOI: 10.1103/PhysRevX.11.021003

Subject Areas: Complex Systems,
Interdisciplinary Physics,
Statistical Physics

I. INTRODUCTION

One of the most important tools in network analysis is the algorithmic division of an unannotated network into groups of similar nodes—a task broadly known as network clustering or community detection [1]. Such divisions allow researchers to provide a summary of the large-scale structure of a network and, in this way, obtain fundamental insight about its function and underlying mechanism of formation. Within this broad umbrella, many community detection methods have been developed, based on different mathematical definitions of the overall task [2]. What most methods share in common is that they are based on some objective function defined over all possible partitions of the network, which, if optimized, yields the most adequate partition for that particular network. Another universal property of community detection methods is that when they are applied to empirical networks, they exhibit at least some degree of degeneracy, in that even if there exists a single partition with the largest score among all others, there is usually an abundance of other solutions that possess a very similar score, making a strict optimization among them somewhat arbitrary [3]. This issue is compounded with the fact that instances of the community detection problem are generically computationally intractable, such that no known algorithm that guarantees the correct solution can perform substantially better than an exhaustive search over all answers [4,5], which is not feasible for networks with more than very few nodes. As a consequence, most available methods rely on stochastic heuristics that give only approximations of the optimum and end up being especially susceptible to the degenerate landscape, yielding different answers whenever they are employed.

In response to this inherent degeneracy, many authors have emphasized the need to collectively analyze many outputs of any given community detection method, not only the best scoring result [6–9]. In this direction, one particularly interesting proposition is to recover the task of detecting a single partition but doing so in a manner that incorporates the consensus over many different alternatives [7,9–15]. If most results are aligned with the same general solution, the consensus among them allows us, in fact, to profit from the degeneracy since small distortions due to irrelevant details or statistical fluctuations are averaged out,

[*] peixotot@ceu.edu

leading to a more robust answer than any of the individual solutions. However, consensus clustering cannot provide a full answer to the community detection problem because any kind of approach based on point estimates possesses an Achilles' heel in situations where the competing answers do not all point in a cohesive direction and instead amount to incompatible results. A consensus between diverging answers is inconsistent in the same manner as the mean of a bimodal distribution is not a meaningful representation of the corresponding population. Therefore, extracting understanding from community detection methods requires more than simply finding a consensus, as we also need to characterize the *dissensus* among the competing partitions. In fact, we need robust methods that give us a complete picture of the entire population of partitions.

Some authors have previously considered the problem of fully characterizing the landscape of possible partitions. Good *et al.* [3] have used nonlinear dimensionality reduction to project the space of partitions in two dimensions, thereby revealing degeneracies. Closer to what is proposed in this work, Calatayud *et al.* [8] have used an *ad hoc* algorithm to cluster partitions, in order to determine how many samples are necessary to better characterize a distribution. Although these previous works effectively demonstrate the role of partition heterogeneity in empirically relevant situations, the approaches developed so far are implemented outside of a well-defined theoretical framework and rely on many seemingly arbitrary choices, such as projection dimension, similarity function used, cluster forming criterion, etc. Thus, it is difficult to interpret, in simple terms, the structures found by those methods and also to evaluate if they are meaningful and statistically significant or are merely artifacts of the provisional choices made.

In this work, we develop a round set of methods to comprehensively characterize a population of network partitions, in a manner that reveals both the consensus and dissensus between them. Our methods start from the formulation of interpretable probabilistic generative models for arbitrary collections of partitions that are based on explicit definitions of the notion of unique group labelings and clusters of partitions. From these models, we are able to derive principled Bayesian inference algorithms that are efficient and effective at characterizing heterogeneous sets of partitions, according to their statistical significance. Importantly, our methods are nonparametric and do not require *a priori* choices to be made, such as distance thresholds or even the number of existing clusters, with the latter being uncovered by our method from the data alone. Our method also bypasses dimensionality reduction [16,17], as required by some data clustering techniques, and operates directly on a collection of partitions. Since it is grounded in a broader statistical framework, our method also allows potential generalizations and principled comparison with alternative modeling assumptions.

We approach our characterization task by first providing a solution to the community label identification problem, which allows us to unambiguously identify groups of nodes between partitions even when their node compositions are not identical. This approach allows us to perform the basic (but, until now, not fully solved) task of computing marginal distributions of group memberships for each node in the network, and it also naturally leads to a way of comparing partitions based on the maximum overlap distance, which has a series of useful properties that we demonstrate. Our method yields a simple way to characterize the consensus between a set of partitions, acting in a way analogous to a maximum *a posteriori* (MAP) estimation of a categorical distribution. We also highlight the pitfalls of consensus estimation in community detection, which fails when the ensemble of solutions is heterogeneous. Finally, we provide a more powerful alternative, consisting of the generalization of our method to the situation where multiple consensuses are possible, such that groups of partitions can align in different directions. The identification of these partition "modes" yields a compact and understandable description of the heterogeneous landscape of community detection results, allowing us to assess their consistency and weigh the alternative explanations they offer to the network data.

This work is divided as follows. We begin in Sec. II with a description of the label identification problem, which serves as a motivation for our approach on consensus clustering developed in Sec. III, based on the inference of what we call the random label model. In Sec. IV, we discuss how we can extract consensus from network partitions via "point estimates" and how this leads to inconsistencies in situations when the different partitions disagree. We then show how we can find both consensus and dissensus in Sec. V, by generalizing the random label model, thus obtaining a comprehensive description of multimodal populations of partitions, including how partitions may agree and disagree with each other. In Sec. VI, we show how our ideas can be easily generalized to ensembles of hierarchical partitions, and finally, in Sec. VII, we show how our methods allow us to perform more accurate Bayesian model selection, which requires a detailed depiction of the space of solutions that our approach is able to provide. We end in Sec. VIII with a conclusion.

II. GROUP IDENTIFICATION PROBLEM IN COMMUNITY DETECTION

In this work, we focus on the approach to community detection that is based on the statistical inference of generative models [18]. Although our techniques can be used with arbitrary community detection methods (or, in fact, for any data clustering algorithm), those based on inference lend themselves more naturally to our analysis since they formally define a probability distribution over partitions. More specifically, if we consider a generative model for a network conditioned on a node partition $\boldsymbol{b} = \{b_i\}$, where b_i is the group label of node i, such that each network $\boldsymbol{A}$ occurs with a probability $P(\boldsymbol{A}|\boldsymbol{b})$, we obtain the posterior distribution of network partitions by employing Bayes' rule,

$$P(\boldsymbol{b}|\boldsymbol{A}) = \frac{P(\boldsymbol{A}|\boldsymbol{b})P(\boldsymbol{b})}{P(\boldsymbol{A})}, \tag{1}$$

where $P(\boldsymbol{b})$ is the prior probability of partitions and $P(\boldsymbol{A}) = \sum_{\boldsymbol{b}} P(\boldsymbol{A}|\boldsymbol{b})P(\boldsymbol{b})$ is the model evidence. There are many ways to compute this probability, typically according to one of the many possible parametrizations of the stochastic block model (SBM) [19] and the corresponding choice of prior probabilities for their parameters. Since our analysis does not depend on any particular choice, we omit their derivations and instead point the reader to Ref. [18] for a summary of the most typical alternatives. For our present goal, it is sufficient to establish that such a posterior distribution can be defined, and we have mechanisms to either approximately maximize or sample partitions from it.

The first central issue we seek to address is that, for this class of problems, the actual numeric values of the group labels have no particular significance, as we are simply interested in the division of the nodes into groups, not in their particular placement in named categories. Thus, the posterior probability above is invariant to label permutations. More specifically, if we consider a bijective mapping of the labels $\mu(r) = s$, such that its inverse $\mu^{-1}(s) = r$ recovers the original labels, then a label permutation $\boldsymbol{c} = \{c_i\}$, where $c_i = \mu(b_i)$, has the same posterior probability,

$$P(\boldsymbol{b}|\boldsymbol{A}) = P(\boldsymbol{c}|\boldsymbol{A}), \tag{2}$$

for any choice of $\boldsymbol{\mu}$. Very often, this detail is considered to be unimportant since many inference methods break this label permutation symmetry intrinsically. For example, if we try to find a partition that maximizes the posterior distribution with a stochastic algorithm, we will invariably find one of the many possible label permutations, in an arbitrary manner that usually depends on the initial conditions, and we can usually move on with the analysis from there. Methods like belief propagation [5], which can be employed in the special case where the model parameters other than the partition $\boldsymbol{b}$ are known, yield marginal distributions over partitions that, due to random initialization, also break the overall label permutation symmetry and yield a distribution centered around one particular group labeling. The same occurs also for some Markov chain Monte Carlo (MCMC) algorithms, for example, those based on the movement of a single node at a time [20,21], which will often get trapped inside one particular choice of labels. This happens because the swap of two labels can only occur if the respective groups exchange all their nodes one by one, a procedure that invariably moves the Markov chain through low probability states and thus is never observed in practice. Although this spontaneous label symmetry breaking can be seen as a helpful property in these cases, strictly speaking, it is a failure of the inference procedure in faithfully representing the overall label symmetry that exists in the posterior distribution. In fact, this symmetry guarantees that the marginal posterior group membership probability of any node must be the same for all N nodes, i.e.,

$$\pi_i(r) = \sum_{\boldsymbol{b}} \delta_{b_i,r} P(\boldsymbol{b}|\boldsymbol{A}) = \sum_{B=r}^{N} \frac{P(B)}{B}, \tag{3}$$

where $P(B)$ is the marginal distribution of the number of labels (nonempty groups), and we assume that the labels always lie in a contiguous range from 1 to B. Therefore, the true answer to the question "what is the probability of a node belonging to a given group?" is always an unhelpful one since it is the same one for every node and carries no information about the network structure. Far from being a pedantic observation, we encounter this problem directly when employing more robust inference methods such as the merge-split MCMC of Ref. [22]. In that algorithm, the merge and split of groups are employed as direct move proposals, which significantly improve the mixing time and the tendency of the Markov chain to get trapped in metastable states, when compared to single-node moves. However, as a consequence, the merge and split of groups result in the frequent sampling of the same partition where two group labels have been swapped, after a merge and split. In fact, the algorithm of Ref. [22] also includes a joint merge-split move, where the memberships of the nodes belonging to two groups are redistributed in a single move, which often results in the same exact partition but with the labels swapped. Such an algorithm will rapidly cycle through all possible label permutations, leading to the correct, albeit trivial, uniform marginal probabilities given by Eq. (3).

In Fig. 1, we show how the label permutation invariance can affect community detection for a network of copurchases of political books [23], for which we used the Poisson degree-corrected SBM (DC-SBM) [24], with the parametrization of Ref. [25] and the merge-split MCMC of Ref. [22]. Although the individual partitions yield seemingly meaningful divisions, they are observed with a random permutation of the labels, preventing an aggregate statistics at the level of single nodes to yield useful information.

At first, we might think of a few simple strategies that can alleviate the problem. For example, instead of marginal distributions, we can consider the pairwise co-occurrence probabilities $c_{ij} = \sum_{\boldsymbol{b}} \delta_{b_i,b_j} P(\boldsymbol{b}|\boldsymbol{A}) \in [0,1]$, which quantify how often two nodes belong to the same

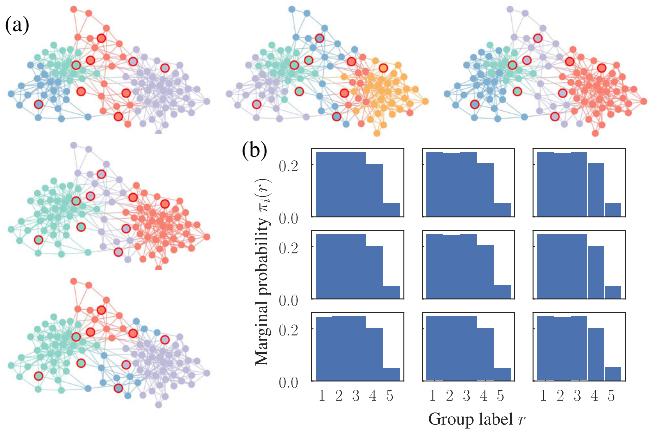


FIG. 1. (a) Five sampled partitions from the posterior distribution of a network of political books, with the group labels represented as colors, using the Poisson DC-SBM and the MCMC algorithm of Ref. [22] (b) Marginal posterior distribution of the group memberships of the nodes highlighted in red in (a), obtained for 10^5 samples from the posterior distribution. The same asymptotic distribution is obtained for every single node in the network.

group and thus are invariant with respect to label permutations. However, this approach gives us a large, dense matrix of size N^2, which is harder to interpret and manipulate than marginal distributions—indeed, the usual approach is to try to cluster this matrix [10] by finding groups of nodes that have similar co-occurrences with other nodes, but this method just brings us back to the same kind of problem. Another potential option is to choose a canonical naming scheme for the group labels, for example, by indexing groups according to their size, such that $r < s$ if $n_r < n_s$, where n_r is the number of nodes with group label r. However, this idea quickly breaks down if we have groups of the same size or if the group sizes vary significantly in the posterior distribution. An alternative canonical naming is one based on an arbitrary ordering of the nodes, forcing the labels to be confined to a contiguous range so that $b_j > b_i$ for $j > i$ whenever b_j corresponds to a group label previously unseen for nodes $k \le i$. In this way, every partition corresponds to a single canonical labeling, which we can generate before collecting statistics on the posterior distribution. Unfortunately, this approach is not straightforward to implement since the marginal distributions will depend strongly on the chosen ordering of the nodes. For example, if the first node happens to be one that can belong to two groups with equal probability, whenever this node changes membership, it will incur the relabeling of every other group, thus spuriously causing the marginal distribution of every other node to be broader, even if they always belong to the "same" group. It seems intuitive, therefore, to order the nodes according to the broadness of their marginal distribution, with the most stable nodes first, but since determining the marginal distribution depends on the ordering itself, it leads to a circular problem.

In the following, we provide a different solution to this problem, based on a generative model of labeled partitions, which is both satisfying and easy to implement; in the end, it allows us to obtain marginal distributions in an unambiguous manner.

III. ESTABLISHING CONSENSUS: THE RANDOM LABEL MODEL

If we have, as an objective, the estimation of the marginal probability $\pi_i(r)$ of node i belonging to group r, given M partitions $\{\boldsymbol{b}\} = \{\boldsymbol{b}^{(1)}, \ldots, \boldsymbol{b}^{(M)}\}$ sampled from a posterior distribution, this is done by computing the mean

$$\pi_i(r) = \frac{1}{M}\sum_{m=1}^{M} \delta_{b_i^m, r}. \tag{4}$$

This approach is fully equivalent to fitting a factorized "mean-field" model on the same samples, given by

$$P_{\mathrm{MF}}(\boldsymbol{b}|\boldsymbol{p}, B) = \prod_i p_i(b_i), \tag{5}$$

where $p_i(r)$ is the probability of node i belonging to group $r \in \{1, \ldots, B\}$. Given the same partitions, the maximum likelihood estimate of the above model corresponds exactly to how we estimate marginal distributions, i.e.,

$$\hat{p}_i(r) = \underset{p_i(r)}{\operatorname{argmax}} \prod_{m=1}^{M} P_{\mathrm{MF}}(\boldsymbol{b}^{(m)}|\boldsymbol{p}, B) = \pi_i(r). \tag{6}$$

Although this computation is common practice, it is important to note that this model is inconsistent with our posterior distribution in Eq. (1) since it is, in general, not invariant to label permutations; i.e., if we swap two labels r and s, we have the same distribution only if $p_i(r) = p_i(s)$ for every node i. Therefore, in order to tackle the label symmetry problem, we may modify this inference procedure by making it also label symmetric. We do so by assuming that our partitions are initially sampled from the above model, but then the labels are randomly permuted. In other words, we have

$$P(\boldsymbol{b}|\boldsymbol{p}, B) = \sum_{\boldsymbol{c}} P(\boldsymbol{b}|\boldsymbol{c}) P_{\mathrm{MF}}(\boldsymbol{c}|\boldsymbol{p}, B), \tag{7}$$

where the intermediary partition $\boldsymbol{c}$ is relabeled into $\boldsymbol{b}$ with a uniform probability

$$P(\boldsymbol{b}|\boldsymbol{c}) = \frac{[\boldsymbol{b} \sim \boldsymbol{c}]}{q(\boldsymbol{c})!}, \tag{8}$$

where we make use of the symmetric indicator function

$$[\boldsymbol{b}\sim\boldsymbol{c}]=\begin{cases}1 & \text{if } \boldsymbol{b} \text{ is a label permutation of } \boldsymbol{c}\\ 0 & \text{otherwise,}\end{cases}\tag{9}$$

and where $q(\boldsymbol{c})$ is the number of labels actually present in partition $\boldsymbol{c}$ [not to be confused with the total number of group labels B in the underlying model since some groups may end up empty, so that $q(\boldsymbol{c})\le B$], and $q(\boldsymbol{c})!$ in the total number of label permutations of $\boldsymbol{c}$. Now, inferring the probabilities $\boldsymbol{p}$ from the model above involves finding a single underlying canonical labeling that is erased at each sample but, after it is identified, allows us to obtain marginal distributions. This canonical labeling itself is not unique since every permutation of its labels is equivalent, but we do not care about the identity of the labels, just an overall alignment, which is what the inference will achieve.

We proceed with the inference of the above model in the following way. Suppose we observe M partitions $\{\boldsymbol{b}\}=\{\boldsymbol{b}^{(1)},\dots,\boldsymbol{b}^{(M)}\}$ sampled from the posterior distribution as before. Our first step is to infer the hidden labels $\{\boldsymbol{c}\}=\{\boldsymbol{c}^{(1)},\dots,\boldsymbol{c}^{(M)}\}$ from the posterior

$$P(\{\boldsymbol{c}\},B|\{\boldsymbol{b}\})=\frac{P(\{\boldsymbol{b}\}|\{\boldsymbol{c}\})P(\{\boldsymbol{c}\}|B)P(B)}{P(\{\boldsymbol{b}\})},\tag{10}$$

with the marginal likelihood integrated over all possible probabilities $\boldsymbol{p}$, and given by

$$P(\{\boldsymbol{c}\}|B)=\int P_{\mathrm{MF}}(\{\boldsymbol{c}\}|\boldsymbol{p})P(\boldsymbol{p}|B)\mathrm{d}\boldsymbol{p}\tag{11}$$

$$=\prod_i\frac{(B-1)!}{(M+B-1)!}\prod_r n_i(r)!,\tag{12}$$

where

$$n_i(r)=\sum_{m=1}^{M}\delta_{c_i^m,r}\tag{13}$$

is the number of relabeled partitions where node i has hidden label r, and we have used an uninformative prior

$$P(\boldsymbol{p}|B)=\prod_i(B-1)!,\tag{14}$$

corresponding to a constant probability density for every node over a B-dimensional simplex, each with volume $1/(B-1)!$, which is also equivalent to a Dirichlet prior with unit hyperparameters. Therefore, up to an unimportant multiplicative constant, we have that the posterior distribution of hidden relabelings is given by

$$\begin{aligned}&P(\{\boldsymbol{c}\},B|\{\boldsymbol{b}\})\\&\quad\propto\left(\prod_{m=1}^{M}[\boldsymbol{b}^{(m)}\sim\boldsymbol{c}^{(m)}]\right)\prod_i\frac{(B-1)!}{(M+B-1)!}\prod_r n_i(r)!,\end{aligned}\tag{15}$$

where have assumed a uniform prior $P(B)=1/N$, which does not contribute to the above. We proceed by considering the conditional posterior distribution of a single partition $\boldsymbol{c}^{(m)}$,

$$\begin{aligned}&P(\boldsymbol{c}^{(m)}|\{\boldsymbol{b}\},\{\boldsymbol{c}^{(m'\neq m)}\},B)\\&\quad\propto\prod_i\prod_r[n'_i(r)+\delta_{c_i^m,r}]!\\&\quad\propto\prod_i\prod_r\{[n'_i(r)+1]!\}^{\delta_{c_i^m,r}}\{[n'_i(r)]!\}^{1-\delta_{c_i^m,r}}\\&\quad\propto\prod_i\prod_r[n'_i(r)+1]^{\delta_{c_i^m,r}},\end{aligned}\tag{16}$$

where $n'_i(r)=\sum_{m'\neq m}\delta_{c_i^{(m')},r}$ is the label count excluding $\boldsymbol{c}^{(m)}$, and we have dropped the indicator function for conciseness, but without forgetting that $[\boldsymbol{c}^{(m)}\sim\boldsymbol{b}^{(m)}]=1$ must always hold. If we seek to find the most likely hidden labeling $\boldsymbol{c}^{(m)}$, we need to maximize the above probability, or equivalently its logarithm, which is given by

$$\ln P(\boldsymbol{c}^{(m)}|\{\boldsymbol{b}\},\{\boldsymbol{c}_{m'\neq m}\},B)=\sum_{i,r}\delta_{c_i^m,r}\ln[n'_i(r)+1],\tag{17}$$

up to an unimportant additive constant. The maximization involves searching through all $q(\boldsymbol{b}^{(m)})!$ possible relabelings of $\boldsymbol{b}^{(m)}$. Unfortunately, this number grows too fast for an exhaustive search to be feasible, unless the number of labels is very small. Luckily, as we now show, it is possible to reframe the optimization in a manner that exposes its feasibility. We begin by representing the mapping between the labels of $\boldsymbol{b}^{(m)}$ and $\boldsymbol{c}^{(m)}$ via the bijective function $\mu(r)$, chosen so that

$$\mu(b_i^m)=c_i^m,\quad\forall\, i.\tag{18}$$

Now, by introducing the matrix

$$w_{rs}=\sum_i\delta_{b_i,r}\ln[n'_i(s)+1],\tag{19}$$

we can express the log-likelihood as

$$\ln P(\boldsymbol{c}^{(m)}|\{\boldsymbol{b}\},\{\boldsymbol{c}_{m'\neq m}\},B)=\sum_r w_{r,\mu(r)}.\tag{20}$$

Therefore, if we consider the matrix w_{rs} as the weighted adjacency matrix of a bipartite graph, where the group labels of $\boldsymbol{b}^{(m)}$ and $\boldsymbol{c}^{(m)}$ form the nodes on each partition (see Fig. 2), the above log-likelihood corresponds to the sum of

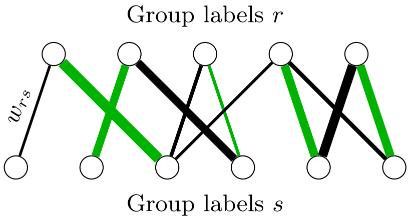


FIG. 2. Relabeling a partition corresponds to finding the solution of a maximum bipartite weighted matching problem, where the partition labels are the nodes of a bipartite graph with weights w_{rs} on the edges. The matching is a bijection $\mu(r)$ that needs to be chosen so that the total sum $\sum_r w_{r,\mu(r)}$ is maximized. In this illustration, the edge thickness corresponds to the weight w_{rs}, and the edges in green correspond to the maximum matching.

the weights of the edges selected by $\boldsymbol{\mu}$. Finding such a bijection is an instance of a very well-known combinatorial optimization problem called maximum bipartite weighted matching, also known as the assignment problem, which corresponds to finding a "matching" on a bipartite graph, defined as a subset of the edges that share no common nodes, such that the sum of the weights of the edges belonging to the matching is maximized. This case corresponds precisely to the sum given in Eq. (20), where a given choice of $\boldsymbol{\mu}$ corresponds to a particular matching. In particular, we are interested in the unbalanced and imperfect version of the matching problem, where the number of groups on both sides might be different, and groups on either side might be left unmatched [26]—in which case, for each unmatched group, we give it a label of a new group. Luckily, fast polynomial algorithms for this problem have been long known. For example, using the "Hungarian" or Kuhn-Munkres algorithm [27,28], this problem can be solved with a worst-case running time of $O(q(\boldsymbol{b})^3)$, which is substantially better than an exhaustive search, rendering our approach not only feasible but also efficient.

Having found the maximum of Eq. (20), we are still left with maximizing the value of B according to Eq. (15). But, as it is easy to verify, the likelihood is a monotonically decreasing function of B. Therefore, since $q(\boldsymbol{c}) \le B$, this step simply amounts to choosing B so that

$$B = \max_m q(\boldsymbol{c}^{(m)}). \tag{21}$$

Equipped with the above information, we can summarize our whole inference algorithm as follows:

(1) We sample M partitions $\boldsymbol{b}^{(1)}, \ldots, \boldsymbol{b}^{(M)}$ from the posterior distribution $P(\boldsymbol{b}|\boldsymbol{A})$.
(2) We initialize $\boldsymbol{c}^{(m)} = \boldsymbol{b}^{(m)}$ for every sample m.
(3) For each sample m, in random order, we obtain a new relabeling $\boldsymbol{c}^{(m)}$ such that Eq. (20) is maximized.
(4) If any value of $\boldsymbol{c}^{(m)}$ is changed during the last step, we repeat it; otherwise, we stop and return $\{\boldsymbol{c}\}$.
(5) We update the inferred value of B according to Eq. (21).

By the end of this algorithm, we are guaranteed to find a local maximum of Eq. (15), but not a global one; hence, we need to run it multiple times and obtain the result with the largest posterior probability. However, we find that repeated runs of the algorithm give the same result in the vast majority of cases we tried [29].

Computationally, step 3 is the heart of the above algorithm, as it corresponds to the alignment of each partition with the rest. It takes time $O[M(N + B^3)]$ in the worst case, where B is the total number of labels used, since for each partition we need time $O(N)$ to compute the weights w_{rs} and time $O(B^3)$ to solve the maximum bipartite weighted matching problem. We can then use the final values of $\{\boldsymbol{c}\}$ to easily obtain the marginal probabilities via

$$\hat{p}_i(r) = \underset{p_i(r)}{\operatorname{argmax}}\ P(\boldsymbol{p}|\{\boldsymbol{c}\}) = \frac{1}{M}\sum_{m=1}^{M} \delta_{c_i^m, r}. \tag{22}$$

Note that the above procedure is not much more computationally intensive than obtaining the marginals in the naive way, i.e., directly from the originally labeled partitions $\boldsymbol{b}$, which requires a time $O(MN)$ to record the label counts. It does, however, require more memory, with a total $O(MN)$ storage requirement, as we need to keep all M partitions for the whole duration of the algorithm. In practice, however, we do not need to perform the whole procedure above for all M partitions, as it is often sufficient to choose a relatively small subset of them, provided they give a good representation of the ensemble; then, we run steps 1 to 4 only on this subset. Thus, we can simply process each remaining partition by simply finding its relabeling $\boldsymbol{c}^{(m)}$, updating the global label counts $n_i(r)$, and then discarding the partition. Although this process gives only an approximation of the optimization procedure, we find it works very well in practice, yielding results that are often indistinguishable from what is obtained with the full algorithm while requiring less memory.

In Fig. 3, we show the partitions of the political books network considered in Fig. 1 but now relabeled according to the algorithm above. Despite groups changing size and composition, and the appearance and disappearance of groups, the unique labeling allows us to identify them clearly across partitions. In Fig. 4, these relabelings are used to obtain marginal distributions on the nodes, where we can say unambiguously, with each frequency, a node belongs to a given group.

A. Maximum overlap distance

The method described in this section serves as a principled way to disambiguate group labels in an ensemble of partitions, but the ideas articulated in its derivation also lead us to a way of comparing two partitions with each

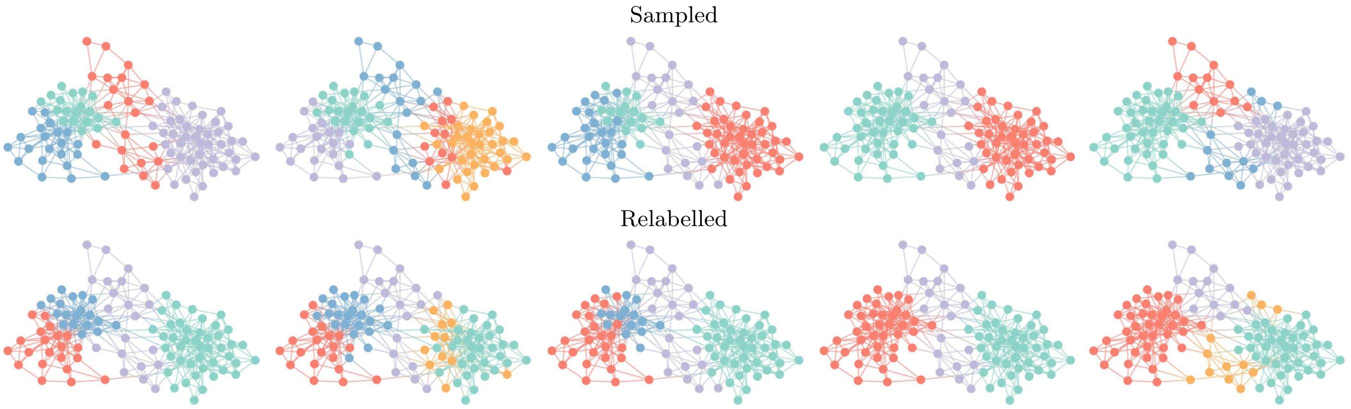


FIG. 3. Five sampled partitions from Fig. 1, in the top panel, with their relabeled counterparts on the bottom panel, using the algorithm described in the text, where it becomes possible to identify groups consistently according to their label (color).

other in a general and meaningful way. Consider the situation where we employ the model above, but we have only $M = 2$ partitions. In this case, without loss of generality, we can set one of them arbitrarily to correspond to the canonical labeling, and we seek to relabel the second one, by maximizing Eq. (20), which, in this case, simplifies to

$$\sum_r m_{r,\mu(r)} \ln 2, \tag{23}$$

where

$$m_{rs} = \sum_i \delta_{b_i^{(1)},r}\delta_{b_i^{(2)},s} \tag{24}$$

is the so-called contingency table between partitions $\boldsymbol{b}^{(1)}$ and $\boldsymbol{b}^{(2)}$, which quantifies how many nodes in group r of $\boldsymbol{b}^{(1)}$ belong to group s of $\boldsymbol{b}^{(2)}$. Therefore, maximizing Eq. (23) is equivalent to finding the bijection $\boldsymbol{\mu}$ so that $\boldsymbol{x}$ with $x_i = \mu(b_i^{(1)})$ and $\boldsymbol{y} = \boldsymbol{b}^{(2)}$ maximize the partition overlap

$$\omega(\boldsymbol{x},\boldsymbol{y}) = \sum_i \delta_{x_i,y_i}, \tag{25}$$

which counts how many nodes share the same label in both partitions. Therefore, incorporating our inference procedure leads to the maximum overlap distance

$$d(\boldsymbol{x},\boldsymbol{y}) = N - \max_{\boldsymbol{\mu}} \sum_i \delta_{\mu(x_i),y_i}. \tag{26}$$

This quantity has a simple interpretation as the minimal classification error, i.e., the smallest possible number of nodes with an incorrect group placement in one partition if the other is assumed to be the correct one. This measure has been considered before in Refs. [31–33], but here, we see its derivation based on a probabilistic generative model. In the Appendix A, we review some of its useful properties.

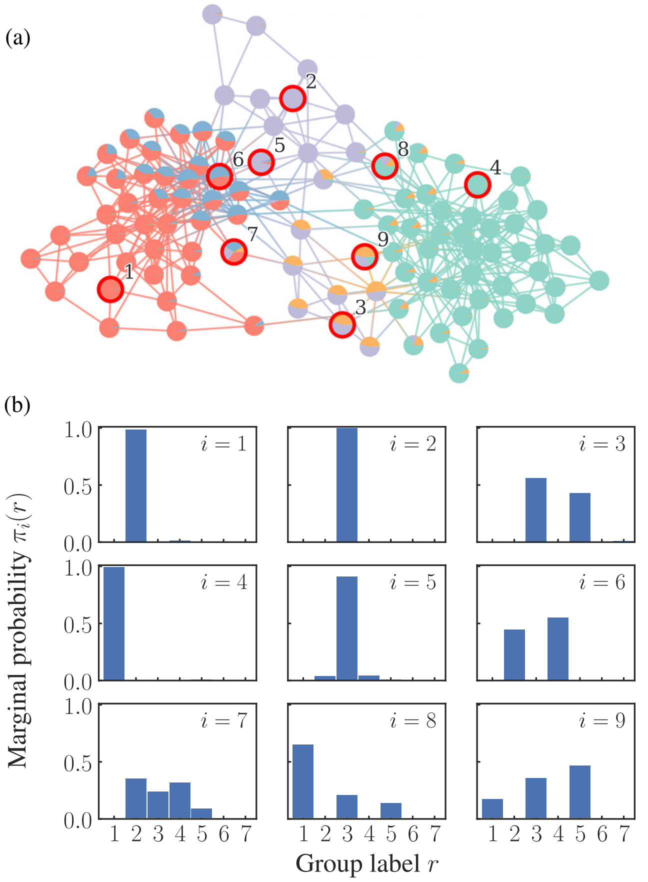


FIG. 4. (a) Marginal posterior group membership distribution on the nodes obtained from relabeled partitions for a network of political books, the same as in Fig. 1, obtained with the algorithm described in the text with $M = 10^5$ samples, represented as pie diagrams on the nodes. (b) Same distributions for the nodes highlighted in red in panel (a).

IV. CONSENSUS AS POINT ESTIMATES

The explicit objective of community detection, like any data clustering method, is to find a partition of the nodes of a network in a manner that captures its structure in a meaningful way. However, instead of a single partition, the inference approach gives us a distribution of partitions, which ascribes to every possible division of the network a plausibility, reflecting both our modeling assumptions and the actual structure of the network. In order to convert this information into a single partition "point estimate," we have to be more specific about what we would consider a successful outcome or, more precisely, how we define the error of our estimate. A consistent scenario is to assume that our observed network is indeed generated from our model $P(\boldsymbol{A}|\boldsymbol{b}^*)$, where $\boldsymbol{b}^*$ is the true partition we are trying to find. In order to quantify the quality of our inference, we need to specify an error function $\epsilon(\boldsymbol{x}, \boldsymbol{y})$ that satisfies

$$\boldsymbol{b}^* = \underset{\boldsymbol{b}}{\operatorname{argmin}}\ \epsilon(\boldsymbol{b}, \boldsymbol{b}^*). \tag{27}$$

Based on a choice for this function, and since we do not really have access to the true partition $\boldsymbol{b}^*$, our best possible estimate $\hat{\boldsymbol{b}}$ from the posterior distribution is the one that minimizes the average error over all possible answers, weighted according to their plausibility, i.e.,

$$\hat{\boldsymbol{b}} = \underset{\boldsymbol{b}}{\operatorname{argmin}} \sum_{\boldsymbol{b}'} \epsilon(\boldsymbol{b}, \boldsymbol{b}') P(\boldsymbol{b}'|\boldsymbol{A}). \tag{28}$$

Therefore, it is clear that our final estimate will depend on our choice of error function $\epsilon(\boldsymbol{x}, \boldsymbol{y})$, and hence, it is not a property of the posterior distribution alone. In statistics and optimization literature, the function $\epsilon(\boldsymbol{x}, \boldsymbol{y})$ is called a "loss function," and it determines the ultimate objective of the inference procedure.

In addition to producing a point estimate $\hat{\boldsymbol{b}}$, it is also useful for our inference procedure to yield an uncertainty value $\sigma_{\hat{\boldsymbol{b}}}$, which quantifies how sure we are about the result, with $\sigma_{\hat{\boldsymbol{b}}} = 0$ indicating perfect certainty. Such choices are not unique, as there are often multiple ways to characterize the uncertainty or how broad a distribution is. But as we will see, the choice of the error function allows us to identify what are arguably the simplest and most direct options.

In the following, we consider simple choices of the error function and investigate how they compare to each other in the inference results they produce.

A. MAP estimation

Arguably the simplest error function we can use is the indicator function (also called the "zero-one" or "all-or-nothing" loss)

$$\epsilon(\boldsymbol{x}, \boldsymbol{y}) = 1 - \prod_i \delta_{x_i, y_i}, \tag{29}$$

which would completely separate the true partition from any other, without differentiating among wrong ones. Inserting this in Eq. (28), we obtain the MAP estimator

$$\hat{\boldsymbol{b}} = \underset{\boldsymbol{b}}{\operatorname{argmax}}\ P(\boldsymbol{b}|\boldsymbol{A}), \tag{30}$$

which is simply the most plausible partition according to the posterior distribution. The corresponding uncertainty for this estimate is simply $\sigma_{\hat{\boldsymbol{b}}} = 1 - P(\boldsymbol{b}|\boldsymbol{A})$, such that if $\sigma_{\hat{\boldsymbol{b}}} = 0$, we are maximally certain about the result. Despite its simplicity, there are several problems with this kind of estimation. Namely, the drastic nature of the error function completely ignores partitions that may be almost correct, with virtually all nodes correctly classified, except very few or, in fact, even one node placed in the incorrect group. We therefore rely on a very strong signal in the data, where the true partition is given a plausibility that is larger than any small perturbation around it in order to be able to make an accurate estimation. This approach puts us in a precarious position in realistic situations where our data are noisy and complex, and it does not perfectly match our modeling assumptions. Furthermore, the uncertainty $\sigma_{\hat{\boldsymbol{b}}}$ is, in most cases, difficult to compute, as it involves determining the intractable sum $P(\boldsymbol{A}) = \sum_{\boldsymbol{b}} P(\boldsymbol{A}, \boldsymbol{b})$, which serves as a normalization constant for $P(\boldsymbol{b}|\boldsymbol{A})$ (although we will consider approximations for this in Sec. VII). Even if computed exactly, typically, $\sigma_{\hat{\boldsymbol{b}}}$ approaches the maximum value of one since very few networks have a single partition with a dominating posterior probability.

B. Maximum overlap consensus (MOC) estimation

As an alternative to the MAP estimation, we may consider a more relaxed error function given by the overlap distance

$$\epsilon(\boldsymbol{x}, \boldsymbol{y}) = N - \sum_i \delta_{x_i, y_i}, \tag{31}$$

which counts the number of nodes correctly classified when compared to the true partition. With this function, from Eq. (28), we obtain the maximum marginal estimator

$$\hat{b}_i = \underset{r}{\operatorname{argmax}}\ \pi_i(r), \tag{32}$$

with

$$\pi_i(r) = \sum_{\boldsymbol{b}} \delta_{b_i, r} P(\boldsymbol{b}|\boldsymbol{A}) \tag{33}$$

being the marginal posterior distribution for node i. The uncertainty, in this case, is then simply the average of the uncertainty for each node, $\sigma_{\hat{\boldsymbol{b}}} = 1 - \sum_i \pi_i(\hat{b}_i)/N$. Since this estimator considers the average over all partitions instead of simply its maximum, it incorporates more information from the posterior distribution. Nevertheless, we again encounter the same problem we described before, namely, that because of label permutation invariance, the marginal distribution will be identical for every node, and this estimator will, in fact, yield useless results. We can fix this problem by instead employing the *maximum* overlap distance of Eq. (26) as an error function $\epsilon(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{x}, \boldsymbol{y})$, leading to the estimator

$$\hat{\boldsymbol{b}} = \underset{\boldsymbol{b}}{\operatorname{argmax}} \sum_{\boldsymbol{b}} \max_{\boldsymbol{\mu}} \sum_i \delta_{\hat{b}_i, \mu(b_i)} P(\boldsymbol{b}|\boldsymbol{A}). \tag{34}$$

Performing the maximization now yields a set of self-consistent equations,

$$\hat{b}_i = \underset{r}{\operatorname{argmax}}\ \pi'_i(r|\{\boldsymbol{\mu}_{\boldsymbol{b}}\}), \tag{35}$$

with the marginal distributions obtained over the relabeled partitions,

$$\pi'_i(r|\{\boldsymbol{\mu}_{\boldsymbol{b}}\}) = \sum_{\boldsymbol{b}} \delta_{\mu_{\boldsymbol{b}}(b_i), r} P(\boldsymbol{b}|\boldsymbol{A}), \tag{36}$$

where the relabeling is done in order to maximize the overlap with $\hat{\boldsymbol{b}}$,

$$\boldsymbol{\mu}_{\boldsymbol{b}} = \underset{\boldsymbol{\mu}}{\operatorname{argmax}} \sum_i \delta_{\hat{b}_i, \mu(b_i)}. \tag{37}$$

Like before, the uncertainty is given by $\sigma_{\hat{\boldsymbol{b}}} = 1 - \sum_i \pi'_i(\hat{b}_i|\{\boldsymbol{\mu}_{\boldsymbol{b}}\})$. In practice, we implement this estimator by sampling a set of M partitions $\{\boldsymbol{b}\}$ from the posterior distribution and then performing the double maximization

$$\hat{b}_i = \underset{r}{\operatorname{argmax}} \sum_m \delta_{\mu_m(b_i^m), r}, \tag{38}$$

$$\boldsymbol{\mu}_{\boldsymbol{m}} = \underset{\boldsymbol{\mu}}{\operatorname{argmax}} \sum_r \hat{m}^{(m)}_{r, \mu(r)}, \tag{39}$$

where, in the last equation, we have that $\hat{m}^{(m)}_{rs} = \sum_i \delta_{b_i^m, r} \delta_{\hat{b}_i, s}$ is the contingency table between $\boldsymbol{b}^{(m)}$ and $\hat{\boldsymbol{b}}$. The solution of Eq. (38) is obtained by simply counting how often each label appears for each node and then extracting the label with the largest count, and Eq. (39) is, once more, an instance of the maximum bipartite weighted matching problem. The overall solution can be obtained by simple iteration, starting from an arbitrary choice of $\hat{\boldsymbol{b}}$, and then alternating between the solution of Eq. (39) and using its result to solve Eq. (38) until $\hat{\boldsymbol{b}}$ no longer changes. This process guarantees a local optimum of the optimization problem but not necessarily a global one; therefore, this algorithm needs to be repeated multiple times with different initial conditions, and the best result is kept. Since it involves relabeling over all M partitions, the overall algorithmic complexity of a single iteration is $O(MNB + MB^3)$.

Note that the marginal distributions obtained via Eq. (36) with the MOC estimator are not necessarily the same as those obtained by inferring the random label model considered previously. This is because, while the MOC calculation attempts to find a single partition with a maximum overlap to all samples, inferring the random label model amounts to finding the most likely marginal distribution compatible with all samples, irrespective of its maximum. Although, in many cases, these two calculations will give similar answers, they are not equivalent.

C. Error functions based on the contingency table

In principle, we can make a variety of other choices for error functions. A particular class of them are those based on the contingency table between partitions, using concepts from information theory. These error functions are not based on an explicit labeling or alignment of partitions but instead focus on the joint probability of labels in both partitions being compared. A popular function of this kind is the variation of information (VI) [34], which is defined as

$$\mathrm{VI}(\boldsymbol{x}, \boldsymbol{y}) = -\frac{1}{N} \sum_{rs} m_{rs} \left[\ln \frac{m_{rs}}{n_r} + \ln \frac{m_{rs}}{n'_s} \right], \tag{40}$$

with $m_{rs} = \sum_i \delta_{x_i, r} \delta_{y_i, s}$ being the contingency table between $\boldsymbol{x}$ and $\boldsymbol{y}$, and $n_r = \sum_s m_{rs}$ and $n'_s = \sum_r m_{rs}$ are the group sizes in both partitions. We can use VI as an error function by setting

$$\epsilon(\boldsymbol{x}, \boldsymbol{y}) = \mathrm{VI}(\boldsymbol{x}, \boldsymbol{y}). \tag{41}$$

As detailed in Ref. [34], VI is a dissimilarity function that fulfills many desirable formal properties, including triangle inequality, making it a proper metric distance (like the maximum overlap distance). Another possible alternative consists of using the reduced mutual information (RMI) [35], as done by Riolo and Newman [9], with

$$\epsilon(\boldsymbol{x}, \boldsymbol{y}) = -\mathrm{RMI}(\boldsymbol{x}, \boldsymbol{y}), \tag{42}$$

where

$$\mathrm{RMI}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{N} \left[\ln \frac{N! \prod_{rs} m_{rs}!}{\prod_r n_r! \prod_s n'_s!} - \ln \Omega(\boldsymbol{n}, \boldsymbol{n}') \right], \tag{43}$$

with $\Omega(\boldsymbol{n}, \boldsymbol{n}')$ being the total number of contingency tables with fixed row and column sums, which we omit here for brevity (see Ref. [35] for asymptotic approximations).

The negative sign used in the definition of $\epsilon(\boldsymbol{x},\boldsymbol{y})$ is because RMI is a similarity function, which takes its maximum value when $\boldsymbol{x}$ and $\boldsymbol{y}$ are identical, unlike VI, which is a dissimilarity that takes its minimum value of zero in the same case. RMI can be seen as a correction to mutual information, which fails as an appropriate similarity function in key cases. It is based on a nonparametric minimum description length (MDL) encoding of both partitions, which quantifies the amount of information required to describe them if the contingency table is known, together with the necessary information required to describe the contingency table itself.

In either of the above cases, our point estimate $\hat{\boldsymbol{b}}$ consists of minimizing the sum of the error function over M samples from the posterior distribution, according to Eq. (28). Unlike the indicator and the maximum overlap distance, the above loss functions are more cumbersome to optimize, with the overall optimization itself amounting to a nonconvex clustering problem of its own. Therefore, we can use some of the same algorithms we use to perform community detection in the first place, with a good choice being the merge-split MCMC of Ref. [22], which we have used in our analysis.

D. Consensus point estimates are inconsistent for heterogeneous distributions

Our aim is not to list or perform an exhaustive comparison between all possible error functions but instead to focus on the fact that they do not always yield the same answer. Although there is only one way with which all partitions in an ensemble can be identical, there are many ways in which they can be different. While the various error functions allow us to extract a form of consensus between differing partitions, they each achieve this based on different features of the population. Therefore, for partition ensembles with sufficiently strong heterogeneity, the different estimators may give conflicting answers. Such a disagreement can signal an arbitrariness in the inference procedure and our inability to summarize the population in a simple manner. We illustrate this problem with a few simple examples.

First, we consider a simple artificial scenario with strong heterogeneity, composed of M independently sampled partitions of N nodes, where, to each node, a group label is sampled uniformly at random from the interval $[1,B]$. Indeed, in this example, there is no real consensus between partitions. Intuitively, we might expect the estimated consensus between such fully random partitions to be a sort of "neutral" partition, in the same way that the average of a fully isotropic set of points in Cartesian space will tend towards the origin. However, all consensus estimators considered previously behave very differently from each other in this example. In Fig. 5, we compare the effective number of groups $B_e(\hat{\boldsymbol{b}}) = \mathrm{e}^S$ obtained for each point estimate, with

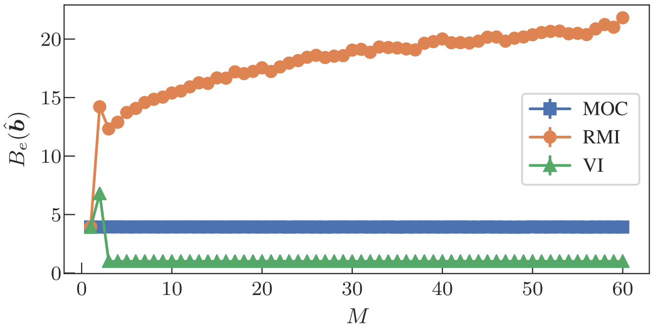


FIG. 5. Effective number of groups $B_e(\hat{\boldsymbol{b}})$ for the consensus estimate $\hat{\boldsymbol{b}}$ obtained for M random partitions of $N = 100$ nodes into $B = 4$ groups, according to the different error functions as indicated in the legend. The results were obtained by averaging over 50 realizations.

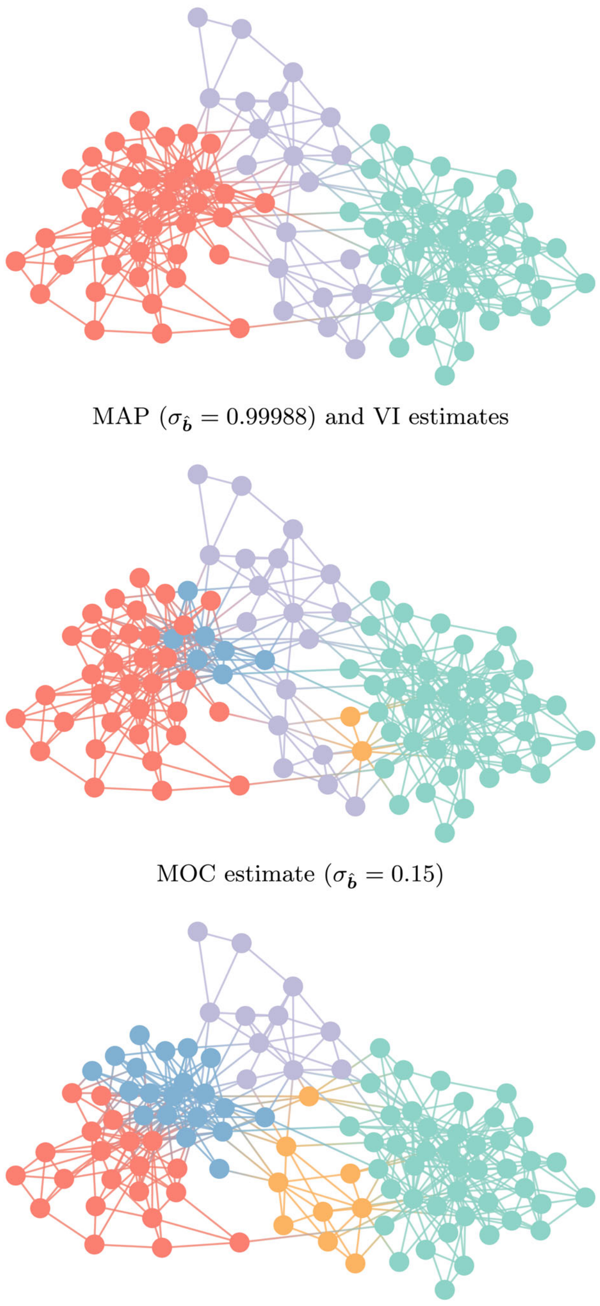
$$S = -\sum_r \frac{n_r}{N} \ln \frac{n_r}{N} \tag{44}$$

being the group label entropy. Arguably, the estimator that behaves the closest to the intuitive expectation just mentioned is VI, which for $M > 2$ yields a consensus partition composed of a single group, $B_e(\hat{\boldsymbol{b}}) = 1$. The MOC estimator yields instead a partition into $B_e(\hat{\boldsymbol{b}}) = 4$ groups, which itself is hard to distinguish from a random partition sampled from the original ensemble. This is because the marginal distributions obtained by Eq. (36) will be close to uniform, even after the label alignments of Eq. (39) are achieved, such that the maximum chosen by Eq. (38) will be determined by small quenched fluctuations in the partition ensemble. Finally, the RMI estimate yields consensus partitions with a number of groups that increases with the number of samples M. This is because the RMI estimate tends to find the overlaps between partitions, i.e., sets of nodes that tend to occur together in the same group across many partitions [9]. In our random case, two nodes belong to the same group because of pure coincidence; therefore, the probability of this happening for a large set of nodes decreases for larger M, thus making the overlapping sets progressively smaller and leading to a larger number of groups in the consensus. Inspecting any of the obtained point estimates in isolation, it would be difficult to get a coherent picture of the underlying ensemble since none of them allows us to distinguish between an ensemble concentrated on the point estimate or the maximally heterogeneous situation we have just considered. If we consider, instead, the uncertainty of the MOC estimate (which yields $\sigma_{\hat{\boldsymbol{b}}} \approx 0.69$ for $M \to \infty$) or even, more explicitly, the marginal distributions of Eq. (36) (or those of the inferred random label model of Sec. III), we would see that they are very broad, closely matching the true random distribution. Nevertheless, none of the point estimates can reveal this information by themselves.

We further illustrate the discrepancy issue with a more realistic example where we can see both agreements and disagreements between the different estimates. In Fig. 6, we show the estimates obtained for the same political books network considered previously, again using the DC-SBM to obtain a posterior distribution of partitions. We observe, rather curiously, that the MAP estimate coincides perfectly with the VI estimate but gives a different result from the MOC and RMI estimates. The MAP/VI estimates separate the network into three groups, which, in this context, can be understood as types of books describing "liberal" and "conservative" politics, and "neutral" books not taking any side. The MOC estimate further divides the "liberal" category into a new subgroup and, somewhat strangely at first, singles out two "neutral" books into their own category. As can be seen in Fig. 4, the reason for this is that the posterior distribution exhibits a possible subdivision of the neutral group into two; however, it is only these two nodes that happen to belong to this subdivision with the highest probability. The RMI estimate also yields a division into five groups, but the two extra groups have a larger size when compared to the MOC result. In view of the behavior seen for the fully random example considered earlier, the discrepancies raise some doubts about what is the most faithful division. Are the MOC and RMI arbitrary divisions due to the randomness of the posterior distribution, or do they point to a meaningful summary? Are the MAP/VI estimates being too conservative about the structure of the posterior distribution?

With some other networks, the discrepancy between estimators can be even stronger, making such questions even harder to answer. In Fig. 7, we show the results for Zachary's karate club network [36], again using the DC-SBM. In this case, the MAP, MOC, and VI estimators yield the same division of the network into a single group, whereas the RMI estimate yields a partition into five groups, following no clear pattern. None of the estimates resembles the putative division for this network in two assortative communities.

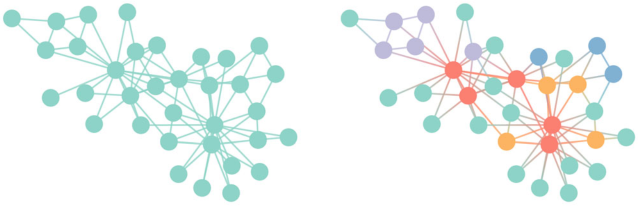
Despite the partial agreement between some of the estimates in the examples above, the disagreements still raise obvious interpretation questions. Here, we argue that this discrepancy cannot be resolved simply by trying alternative ways to form a consensus since trying to summarize a whole distribution with a point estimate is, in general, an impossible task; therefore, we need, instead, a way to also characterize the dissensus between partitions by exposing the existing heterogeneity of the posterior distribution.


MAP ($\sigma_{\hat{b}} = 0.99988$) and VI estimates

MOC estimate ($\sigma_{\hat{b}} = 0.15$)

RMI estimate

FIG. 6. Inference of the community structure of the political books network, according to the DC-SBM and using the different estimators as shown in the legend. For the VI/MAP estimate (top panel), the three groups can be interpreted, from left to right, as "liberal," "neutral," and "conservative."


MAP ($\sigma_{\hat{b}} = 0.51$), MOC ($\sigma_{\hat{b}} = 0.18$) and VI estimates

RMI estimate

FIG. 7. Inference of the community structure of Zachary's karate club network, according to the DC-SBM and using the different estimators as shown in the legend.

To some extent, the characterization of dissensus is already achieved by the random label model of Sec. III since it attempts to describe the posterior distribution via marginal probabilities rather than just a point estimate and therefore can convey how concentrated it is. However, because this model assumes the group membership of each node to be independent, it still hides a significant fraction of the potential heterogeneity in the ensemble, which can come from the correlation between these memberships. In the next section, we generalize this approach to the situation where the posterior distribution is multimodal, so multiple consensuses are simultaneously possible. We see how this allows us to extract a more complete and coherent picture of distributions of partitions.

V. EXTRACTING DISSENSUS BETWEEN PARTITIONS

We aim to characterize the discrepancy between partitions by considering the possibility of several consensuses that only exist between a subset of the partitions. This corresponds to the situation where the inference procedure can yield substantially different explanations for the same network. We achieve this goal by modeling the posterior distribution of partitions with a mixture model, where each partition can belong to one of the K clusters—which we call “modes” to differentiate from the groups of nodes in the network. Inside each mode, the partitions are generated according to the same random label model considered before but with different parameters. More specifically, a partition $\boldsymbol{b}$ is sampled according to

$$P(\boldsymbol{b}|\boldsymbol{p},\boldsymbol{w}) = \sum_k P(\boldsymbol{b}|\boldsymbol{p},k)P(k|\boldsymbol{w}), \tag{45}$$

where

$$P(k|\boldsymbol{w}) = w_k \tag{46}$$

is the relative size of mode k, with $\sum_k w_k = 1$, and inside a mode k, the partitions are sampled according to the random label model,

$$P(\boldsymbol{b}|\boldsymbol{p},k) = \sum_{\boldsymbol{c}} P(\boldsymbol{b}|\boldsymbol{c})P_{\mathrm{MF}}(\boldsymbol{c}|\boldsymbol{p},k), \tag{47}$$

with the hidden labels generated according to

$$P_{\mathrm{MF}}(\boldsymbol{c}|\boldsymbol{p},k) = \prod_i p_i^{(k)}(c_i), \tag{48}$$

where $p_i^{(k)}(r)$ is the probability that a node i has group label r in mode k, and finally a random label permutation chosen uniformly at random,

$$P(\boldsymbol{b}|\boldsymbol{c}) = \frac{[\boldsymbol{b}\sim\boldsymbol{c}]}{q(\boldsymbol{b})!}. \tag{49}$$

Naturally, we recover the original random label model for $K = 1$.

We perform the inference of the above model by considering the mode label k as a latent variable, which yields a joint probability together with the original and relabeled partitions,

$$P(\boldsymbol{b},\boldsymbol{c},k|\boldsymbol{p},\boldsymbol{w}) = P(\boldsymbol{b}|\boldsymbol{c})P(\boldsymbol{c}|\boldsymbol{p},k)P(k|\boldsymbol{w}). \tag{50}$$

If we now observe M partitions $\{\boldsymbol{b}\} = \{\boldsymbol{b}^{(1)},\ldots,\boldsymbol{b}^{(M)}\}$ sampled from the SBM posterior distribution, we assume that each one has been sampled from one of the K modes, so for each observed partition $\boldsymbol{b}_m$, we want to infer its relabeled counterpart, together with its originating mode, i.e., $(\boldsymbol{c}^{(m)},k)$. The joint posterior distribution for these pairs, together with the total number of modes K and the number of groups $\boldsymbol{B} = \{B_k\}$ in each mode, is given by

$$P(\{\boldsymbol{c},k_m\},\boldsymbol{B},K|\{\boldsymbol{b}\}) = \frac{P(\{\boldsymbol{b}\}|\{\boldsymbol{c}\})P(\{\boldsymbol{c}\}|\boldsymbol{k},\boldsymbol{B})P(\boldsymbol{B})P(\boldsymbol{k}|K)P(K)}{P(\{\boldsymbol{b}\})}, \tag{51}$$

where the relabeling probability is given by

$$P(\{\boldsymbol{b}\}|\{\boldsymbol{c}\}) = \prod_m P(\boldsymbol{b}^{(m)}|\boldsymbol{c}^{(m)}), \tag{52}$$

and with the marginal likelihood obtained by integrating over all possible probabilities $\boldsymbol{p}$ for each mode,

$$P(\{\boldsymbol{c}\}|\boldsymbol{k},\boldsymbol{B}) = \prod_k \int\left[\prod_m P(\boldsymbol{c}^{(m)}|\boldsymbol{p},k_m)^{\delta_{k_m,k}}\right]P(\boldsymbol{p})\mathrm{d}\boldsymbol{p} \tag{53}$$

$$= \prod_k\prod_i \frac{(B_k-1)!}{(M_k+B_k-1)!}\prod_r n_i^{(k)}(r)!, \tag{54}$$

with $M_k = \sum_m \delta_{k_m,k}$ being the number of samples that belong to mode k, B_k the total number of group labels in mode k, and $n_i^{(k)}(r) = \sum_m \delta_{c_i^m,r}\delta_{k_m,k}$ the marginal label counts in mode k; finally, the prior mode distribution is obtained by integrating over all possible mode mixtures $\boldsymbol{w}$,

$$P(\boldsymbol{k}|K) = \int\left[\prod_m P(k_m|\boldsymbol{w})\right]P(\boldsymbol{w}|K)\mathrm{d}\boldsymbol{w} \tag{55}$$

$$= \frac{(K-1)!}{(M+K-1)!}\prod_k M_k!. \tag{56}$$

where we used, once more, an uninformative prior

$$P(\boldsymbol{w}|K) = (K-1)!. \tag{57}$$

For the total number of modes K, we use a uniform prior $P(K) \propto 1$, which has no effect on the resulting inference. With this posterior in place, we can find the most likely mode distribution with a clustering algorithm that attempts to maximize it. We do so by starting with an arbitrary initial placement of the M partitions into modes and by implementing a greedy version of the merge-split algorithm of Ref. [22] that chooses at random between the following steps and accepting it only if it increases the posterior probability:

1. A random partition $\boldsymbol{b}^{(m)}$ is moved from its current mode to a randomly chosen one, including a new mode.
2. Two randomly chosen modes are merged into one, reducing the total number of modes.
3. A randomly chosen mode is split into two, increasing the total number of modes. The division itself is chosen by a surrogate greedy algorithm, which tries one of the following strategies at random:
    - (a) Start with a random split of the modes into two; attempt to move each sample in random sequence between the two modes if the move increases the posterior probability, and stop when no improvement is possible.
    - (b) Start with each of the samples in their own modes, with a single sample each, and place them in sequence in two new modes that are initially empty, according to the choice with the largest posterior probability.
    - (c) Start with all samples in a single mode, and proceed like in strategy (b).
4. Two randomly chosen modes are merged into one and then split like in option 3, preserving the total number of modes.

The algorithm stops whenever further improvements to the posterior cannot be made. In the above, whenever a sample m is placed into a mode k, its hidden labeling $\boldsymbol{c}^{(m)}$ is obtained by maximizing the conditional posterior probability,

$$P(\boldsymbol{c}^{(m)}|\{\boldsymbol{b}\}, \{\boldsymbol{c}^{(m'\neq m)}\}, B_k, k) \propto \prod_i \prod_r [n_i'(r|k) + 1]^{\delta_{c_i^m,r}}, \tag{58}$$

where $n_i'(r|k) = \sum_{m'\neq m} \delta_{k_m',k}\delta_{c_i^{m'},r}$ is the label count of node i considering all samples belonging to mode k, excluding $\boldsymbol{c}^{(m)}$. Like in the original random label model, this maximization is performed by solving the corresponding maximum bipartite weighted matching problem with the Kuhn-Munkres algorithm in time $O(N + B^3)$, where B is the number of partition labels involved. Overall, a single "sweep" of the above algorithm, where each sample has been moved once, is achieved in time $O[M(N + B^3)]$. For the choice of M itself, this result will, in general, depend on the structure of the data. The general guideline is that M should be large enough so that if it is increased, the inference results (i.e., number of modes and their composition) no longer change. A good strategy is to make M as large as the initial computational budget allows and then compare the results with a *smaller* choice of M and evaluate if the results are the same. In terms of practical speed, when compared, e.g., to sampling partitions from the SBM posterior via MCMC, we find that performing the overall clustering algorithm is most often substantially faster than generating the partitions in the first place.

After we find the mode memberships $\boldsymbol{k}$, the mode fractions can be estimated as

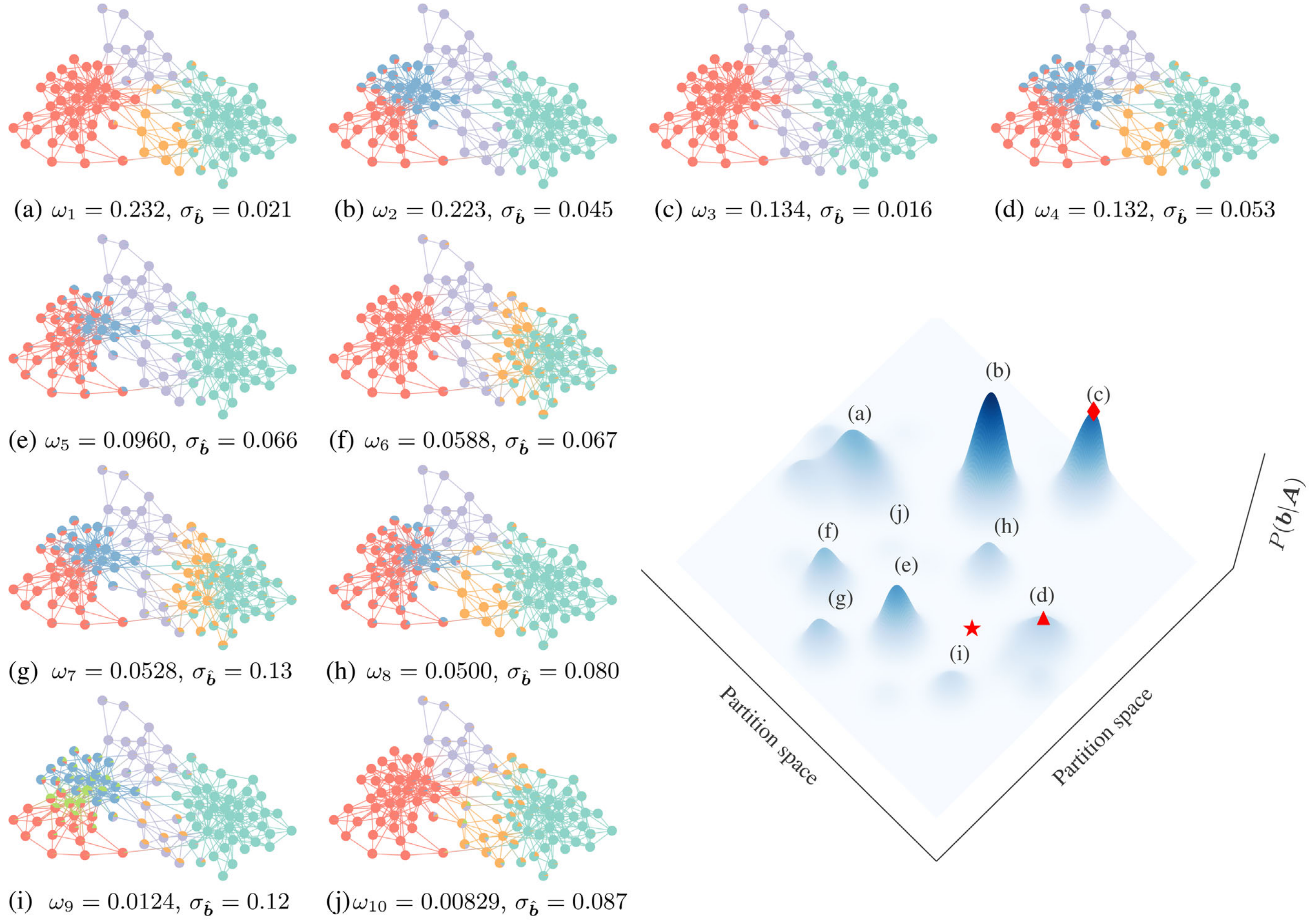
$$w_k = \frac{M_k}{M}, \tag{59}$$

which is interpreted as the relative posterior plausibility of each mode serving as an alternative explanation for the data.

In the following, we consider a simple example that illustrates how the method above can characterize the structure of a distribution of partitions, and we proceed to investigate how the multimodal nature of the posterior distribution can be used to assess the quality of fit of the network model being used.

A. Simple example

In Fig. 8, we show the result of the above algorithm for the posterior distribution obtained for the same political books network considered previously, where, in total, $K = 11$ modes are identified. For each mode, we show the corresponding marginal distribution of the relabeled partitions and the uncertainty $\sigma_{\hat{\boldsymbol{b}}} = 1 - \sum_i p_i(\hat{b}_i)$ of its maximum $\hat{\boldsymbol{b}}$, which serves as a quantification of how broadly distributed the individual modes are. As a means of illustration, in Fig. 8, we also show a two-dimensional projection of the distribution of partitions, obtained using the UMAP dimensionality reduction algorithm [17] with the maximum overlap distance as the dissimilarity metric (similar results can also be found with other dissimilarity functions, as shown in Appendix D). This algorithm attempts to project the distribution of partitions in two dimensions while preserving the relative distances between partitions in the projection. As a result, we see that each mode is clearly discernible as a local concentration of partitions, much like we would expect of a heterogeneous mixture of continuous variables. We note here that we have not informed the UMAP algorithm of the modes we have found with the algorithm above, and therefore, this serves as additional evidence for the existence of the uncovered heterogeneity in the posterior distribution. The most important result of this analysis is that no single mode



(a) $\omega_1 = 0.232$, $\sigma_{\hat{b}} = 0.021$ (b) $\omega_2 = 0.223$, $\sigma_{\hat{b}} = 0.045$ (c) $\omega_3 = 0.134$, $\sigma_{\hat{b}} = 0.016$ (d) $\omega_4 = 0.132$, $\sigma_{\hat{b}} = 0.053$

(e) $\omega_5 = 0.0960$, $\sigma_{\hat{b}} = 0.066$ (f) $\omega_6 = 0.0588$, $\sigma_{\hat{b}} = 0.067$

(g) $\omega_7 = 0.0528$, $\sigma_{\hat{b}} = 0.13$ (h) $\omega_8 = 0.0500$, $\sigma_{\hat{b}} = 0.080$

(i) $\omega_9 = 0.0124$, $\sigma_{\hat{b}} = 0.12$ (j)$\omega_{10} = 0.00829$, $\sigma_{\hat{b}} = 0.087$

FIG. 8. Inferred partition modes from $M = 10^5$ samples of the DC-SBM posterior distribution for the political books network. Panels (a)–(j) show the marginal distributions for each identified mode as pie diagrams on the nodes of the network, with the legend specifying the relative mode fraction w_k and the uncertainty $\sigma_{\hat{b}}$ of the maximum for each mode. The bottom-right panel shows the projection of the partition distribution in two dimensions according to the UMAP dimensionality reduction algorithm [17], where the different modes can be identified as local peaks of the distribution. The star symbol shows the location of the MOC estimate, the diamond symbol the position of the MAP/VI estimate, and the triangle the position of the RMI estimate.

has a dominating fraction of the distribution, with the largest mode corresponding only to around 23% of the posterior distribution and with the second largest mode being very close to it. Thus, there is no single cohesive picture that emerges from the distribution, and therefore, our attempt at summarizing it with a single partition seems particularly ill suited.

In view of this more detailed picture of the ensemble of partitions, it is worth revisiting the consensus results obtained previously with the various error functions. As shown in Fig. 8, the MAP/VI estimates correspond to the most likely partition of mode (c), which is, overall, only the third most plausible mode with $w_3 = 0.134$. From the point of view of the MAP estimator, this serves to illustrate how choosing the most likely partition may, in fact, run counter to intuition: Although the single-most-likely partition belongs to mode (c), collectively, the partitions in modes (a) and (b) have a larger plausibility. Thus, if we are forced to choose a single explanation for the data, it would make more sense to choose mode (a), despite the fact that it does not contain the single-most-likely partition. More concretely, when comparing modes (a)–(c), we see that the network does, in fact, contain more evidence for a division of either the "neutral" or the "liberal" groups into subgroups than the MAP estimate implies; however, it does not contain evidence for both, as mode (d), corresponding to the simultaneous subdivisions, has a smaller plausibility than the other options. The VI estimate also points to mode (c), but it is unclear why. This is indeed a problem with using VI since, despite its strong formal properties, it lacks a clear interpretability.

Differently from MAP and VI, the MOC estimation combines the properties of all modes into a "Frankenstein's monster," where local portions of the final inferred partition correspond to different modes. Thus, the resulting point estimate has a very low posterior probability and hence is a misleading representation of the population—a classic estimation failure of multimodal distributions.

The RMI estimate behaves differently and corresponds to a typical partition of mode (d), which has an overall plausibility of $w_4 = 0.132$. We can understand this choice by inspecting its composition and noticing that the more plausible modes (a)–(c) correspond to partitions where groups of (d) are merged together. Therefore, the RMI similarity sees this partition as the "center" composed of the building blocks required to obtain the other ones via simple operations. But by no means is it the most likely explanation of the data according to the model, and given that it is a division into a larger number of groups, it is more likely to be an overfit, in view of the existence of simpler modes (a)–(c).

B. Evaluating model consistency

The full characterization of the posterior distribution with our approach gives us the opportunity to assess the quality of fit between the model and data. Indeed, if the model is an excellent fit, e.g., if the data are, in fact, generated by the SBM, we should expect a single mode in the posterior distribution that is centered in the true partition [5] (although the broadness of the mode, represented by the variance of the marginal distribution on the nodes, will depend on how easily detectable the true partition is). Therefore, the fact that we observe multiple modes is an indication of some degree of mismatch, with the model offering multiple explanations for the data. Since our analysis allows us to inspect each individual explanation and ascribe to it a plausibility, it can be used to make a more precise evaluation of the fit.

Inspecting the modes observed for the political books network in Fig. 8, we notice that the four largest modes approximately amount to different combinations of the same five groups that appear in the fourth mode [Fig. 8(d)]—although the remaining modes deviate from this pattern. This case is reminiscent of a situation considered by Riolo and Newman [9], who applied RMI estimation to artificial networks where none of the posterior samples matches the true division, which is only uncovered by the RMI consensus. In particular, in their scenario, the consensus exposed "building blocks," i.e., groups of nodes that tend to be clustered together, although the building blocks themselves always appear merged together into bigger groups. The situation where the partitions exhibit clear shared building blocks that always appear merged together, but in different combinations, begs the question as to why the posterior distribution fails to concentrate on the isolated building blocks in the first place. One possibility is that the building blocks do not correspond to the same kind of communities that the inference approach is trying to uncover; e.g., in the case of the SBM, these should be nodes that have the same probability of connection to the rest of the network, which would be a case of model mismatch; hence, it would be difficult to interpret what the building blocks actually mean. Another option, which we can address more directly, is that the model being used underfits the data; i.e., the model formulation fails to recognize the available statistical evidence, resulting in the choice of simpler SBMs with fewer groups, such that some "true" groups are merged together. A common cause of underfitting is the use of noninformative priors that overly penalize larger numbers of groups, as was shown in Ref. [37]. The use of hierarchical priors solves this particular underfitting problem, as discussed in Refs. [25,38]. Another potential cause of underfitting is the use of Poisson formulations for the SBM for networks with heterogeneous density, which assumes that the observed simple graph is a possible realization of a multigraph model that generates simple graphs with a very small probability. Reference [39] introduced an alternative SBM variation based on a simple but consequential modification of the Poisson SBMs, where multigraphs are generated at a first stage and the multiedges are converted into simple edges, resulting in a Bernoulli distribution obtained from the cumulative Poisson distribution. These "latent Poisson" SBMs also prevent underfitting and, in fact, make the posterior distribution concentrate on the correct answer for the examples considered by Riolo and Newman [9], as shown in Ref. [39].

In Fig. 9, we show our method employed on the posterior distribution of the political books network using the latent Poisson DC-SBM with nested priors, which should be able to correct the kinds of underfitting mentioned above. Indeed, the most likely mode shows a more elaborate division of the network into $B = 8$ groups, corresponding to particular subdivisions of the same liberal-neutral-conservative groups seen previously. However, these subdivisions are not quite the same as those seen in Fig. 8 for the Poisson SBM. Therefore, in this example, it would be futile to search for these uncovered groups in the posterior distribution of the Poisson DC-SBM, even if we search for overlaps between partitions. However, despite the more detailed division of the network, the latent Poisson SBM is far from being a perfect fit for this network, as we still observe $K = 11$ modes, corresponding mostly to different divisions of the "conservative" books. When comparing the structure of the different modes, we see that these are not simple combinations of the same subdivisions but rather different rearrangements. This case seems to point to a kind of structure in the network that is not fully captured by the strict division of the nodes in discrete categories, at least not in the manner assumed by the SBM.

In Fig. 10, we also compare the inferences obtained with both SBM models for the karate club network considered previously. The posterior distribution obtained with the Poisson DC-SBM is very heterogeneous, with $K = 30$ modes. It has, as the most plausible mode, one composed of a single partition into a single group (implying that the degree sequence alone is enough to explain the network,

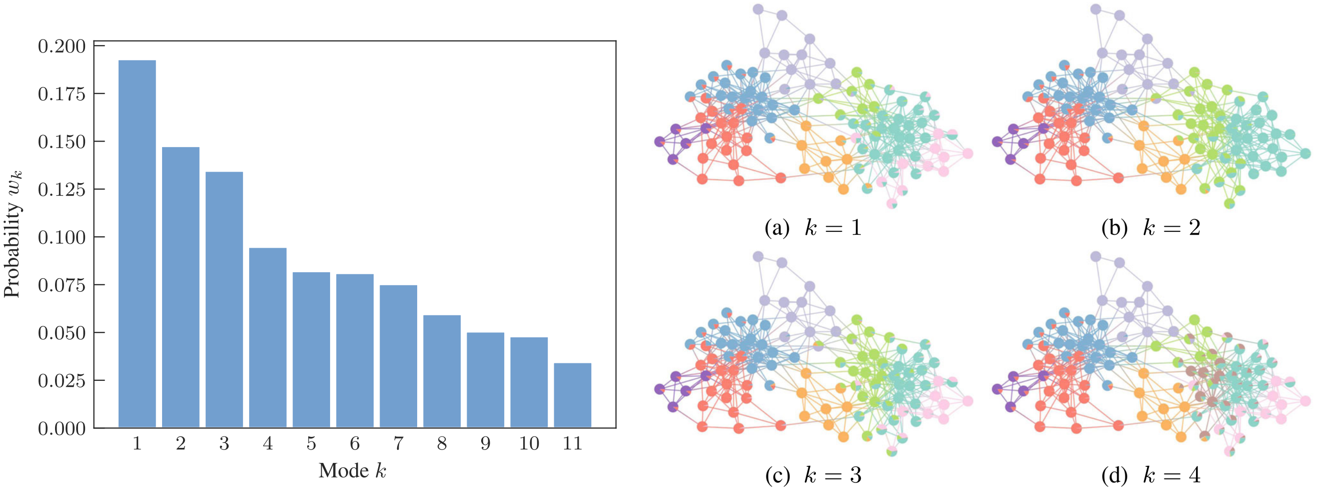


FIG. 9. Inferred partition modes from $M = 10^5$ samples of the latent Poisson DC-SBM posterior distribution for the political books network. The left panel shows the mode fractions w_k, and the right panel the four largest modes (a)–(d), with the marginal distributions shown as pie diagrams on the nodes of the network.

and no community structure is needed). The second most likely mode corresponds to leader-follower partitions, largely dividing the nodes according to degree (despite the degree correction). The putative division of this network into two assortative communities is only the ninth most likely mode. With such an extreme heterogeneity between partitions, finding a consensus between them seems particularly futile, thus explaining the obtained point estimates in Fig. 7, in particular, the odd behavior of the RMI estimate that tries to assemble all diverging modes into a single partition. On the other hand, with the latent Poisson SBM, the posterior distribution changes drastically, as is shown in the right panel of Fig. 10. In this case, the dominating mode corresponds to partitions that, while not fully identical to the accepted division, are more compatible with it, as they only further divide one of the communities into two extra groups. The commonly accepted division itself comes as a typical partition of the second most likely mode. Overall, the posterior distribution becomes more homogeneous, with only $K = 9$ modes identified and with most of the posterior probability assigned to the first few.

It is important to observe that the heterogeneity of the posterior distribution by itself cannot be used as a criterion in the decision of which model is a better fit. Indeed, a typical behavior encountered in statistical inference is the "bias-variance trade-off" [40], where a more accurate representation of the data comes at the cost of increased variance in the set of answers. We illustrate this with a network of American football games [41] shown in Fig. 11. The Poisson DC-SBM yields a very simple posterior distribution, strongly concentrated on a typical partition

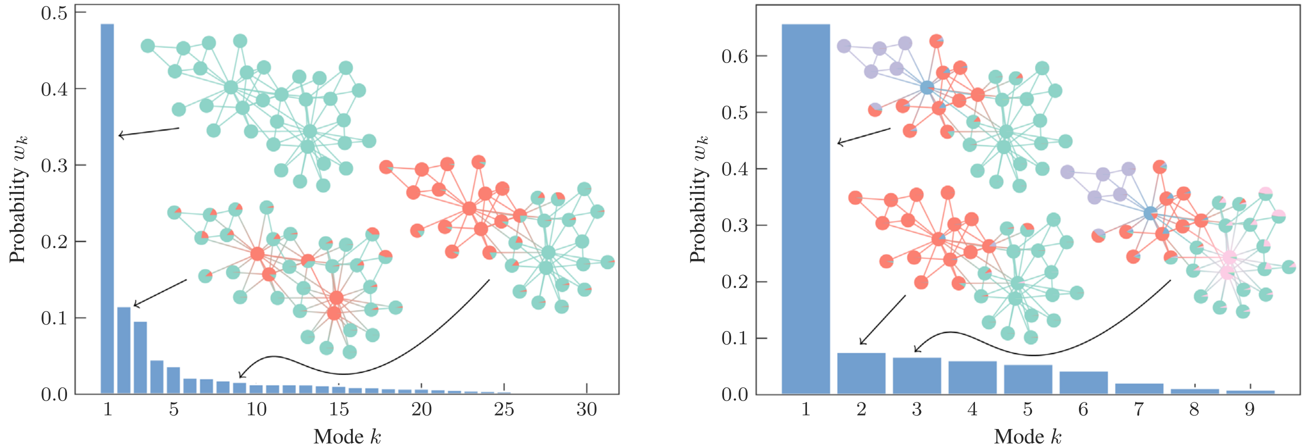


FIG. 10. Inferred partition modes from $M = 10^5$ samples of the posterior distribution obtained with the Poisson DC-SBM (left panel) and latent Poisson DC-SBM (right panel) for the karate club network. The insets show the modes as indicated by the arrows, with the marginal distributions shown as pie diagrams on the nodes of the network.

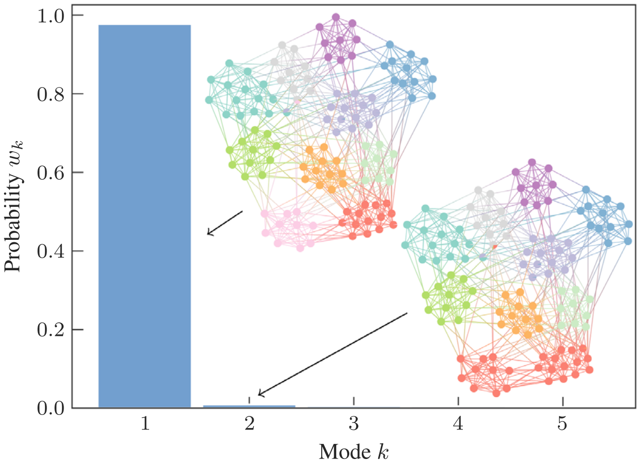


FIG. 11. Inferred partition modes from $M = 10^5$ samples of the posterior distribution obtained with the Poisson DC-SBM for the American college football network. The insets show the modes as indicated by the arrows, with the marginal distributions shown as pie diagrams on the nodes of the network.

into $B = 10$ groups. On the other hand, as seen in Fig. 12, the latent Poisson DC-SBM yields a more heterogeneous posterior distribution with $K = 7$ modes, typically uncovering a larger number of groups. It would be wrong to conclude that the Poisson SBM provides a better fit only because it concentrates on a single answer, if that single answer happens to be underfitting. But from this analysis alone, it is not possible to say if the latent Poisson SBM is not overfitting either. To make the final decision, we need to compute the total evidence for each model, as we consider in Sec. VII. This computation takes the heterogeneity of the posterior distribution into consideration, but it is combined with the model plausibility.

Before we proceed with model selection, we first show how the methods constructed so far can be generalized for hierarchical partitions, which form the basis of generically better-fitting models of community structure in networks [25].

VI. HIERARCHICAL PARTITIONS

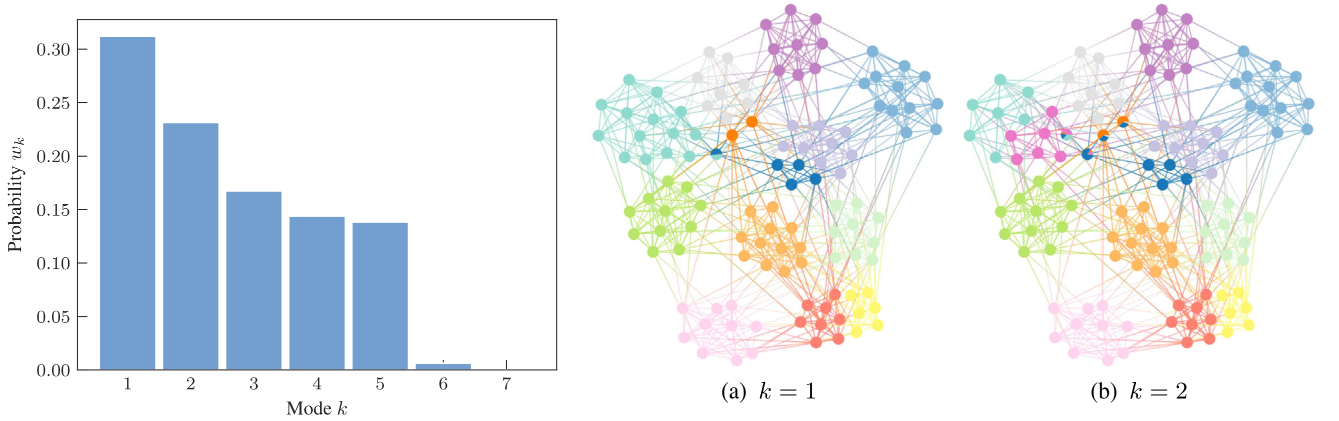
An important extension of SBM formulations is one where the choice of priors is replaced by a nested sequence of priors and hyperpriors, where groups of nodes are also clustered in their own metagroups, associated with a coarse-grained version of the network described via its own smaller SBM, and so on recursively, resulting in a nested version of the model [25,38]. This hierarchical formulation recovers the usual SBMs when the hierarchy has only a single level, and it also introduces many useful properties, including a dramatically reduced tendency to underfit large networks [25,38], as well as a simultaneous description of the network structure at several scales of resolution. This model variant takes as a parameter a hierarchical partition $\bar{\boldsymbol{b}} = \{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_L\}$, where $b_i^{(l)}$ is the group membership of node i in level l, and each group label in level l is a node in the above level $l + 1$, which results in the number of nodes in level l being the number of groups in the level below, $N_l = B_{l-1}$, except for the first level, $N_1 = N$. For this model, we have a posterior distribution over hierarchical partitions given by

$$\pi(\bar{\boldsymbol{b}}) = \frac{P(\boldsymbol{A}|\bar{\boldsymbol{b}})P(\bar{\boldsymbol{b}})}{P(\boldsymbol{A})}. \tag{60}$$

Like in the nonhierarchical case, this posterior distribution is invariant to label permutations, i.e.,

$$\pi(\bar{\boldsymbol{b}}) = \pi(\bar{\boldsymbol{c}}), \tag{61}$$

if $\bar{\boldsymbol{b}}$ and $\bar{\boldsymbol{c}}$ are identical up to a relabeling of the groups. However, in the hierarchical scenario, the group relabelings that keep the posterior distribution invariant must



(a) $k = 1$

(b) $k = 2$

FIG. 12. Inferred partition modes from $M = 10^5$ samples of the latent Poisson DC-SBM posterior distribution for the American college football network. The left panel shows the mode fractions w_k, and the right panel shows the two largest modes (a) and (b), with the marginal distributions shown as pie diagrams on the nodes of the network.

keep the same partitions when projected at the lower levels. In other words, the invariant permutation of the labels in level l affects the nodes in level $l+1$. More specifically, if we consider a bijection $\mu(r)$ for labels at level l, such that $b_i^l(r) = \mu(c_i^l(r))$, then we must change the membership in level $l+1$ to $b_{\mu(i)}^{l+1} = c_i^{l+1}$. If two hierarchical partitions $\bar{\boldsymbol{b}}$ and $\bar{\boldsymbol{c}}$ are identical up to this kind of transformation, we denote this with the indicator function

$$[\bar{\boldsymbol{b}} \sim \bar{\boldsymbol{c}}] = 1, \tag{62}$$

or $[\bar{\boldsymbol{b}} \sim \bar{\boldsymbol{c}}] = 0$ otherwise. Based on this approach, we can generalize the random label model considered before to model hierarchical partitions sampled from the posterior distribution. We first assume that the labels at all levels are sampled independently as

$$P_{\mathrm{MF}}(\bar{\boldsymbol{c}}|\bar{\boldsymbol{p}}) = \prod_{l=1}^{L} P_{\mathrm{MF}}(\boldsymbol{c}_l|\boldsymbol{p}_l), \tag{63}$$

with

$$P_{\mathrm{MF}}(\boldsymbol{c}_l|\boldsymbol{p}_l) = \prod_i p_i^l(c_i^l), \tag{64}$$

where $p_i^l(r)$ is the probability that node i in level l belongs to group r. After sampling a partition $\bar{\boldsymbol{c}}$, we then obtain a final partition $\bar{\boldsymbol{b}}$ by choosing uniformly among all label permutations, yielding

$$P(\bar{\boldsymbol{b}}|\bar{\boldsymbol{p}}) = \sum_{\bar{\boldsymbol{c}}} P(\bar{\boldsymbol{b}}|\bar{\boldsymbol{c}}) P_{\mathrm{MF}}(\bar{\boldsymbol{c}}|\bar{\boldsymbol{p}}), \tag{65}$$

where

$$P(\bar{\boldsymbol{b}}|\bar{\boldsymbol{c}}) = \frac{[\bar{\boldsymbol{b}} \sim \bar{\boldsymbol{c}}]}{\prod_l q(\boldsymbol{c}_l)!}. \tag{66}$$

If we now consider M sampled hierarchical partitions $\{\bar{\boldsymbol{b}}\} = \{\bar{\boldsymbol{b}}^{(1)}, \ldots, \bar{\boldsymbol{b}}^{(M)}\}$, the posterior distribution of the hidden relabeled hierarchical partitions $\{\bar{\boldsymbol{c}}\}$ is given by

$$P(\{\bar{\boldsymbol{c}}\}|\{\bar{\boldsymbol{b}}\}, B_l) \propto \left(\prod_{m=1}^{M} [\bar{\boldsymbol{b}}^{(m)} \sim \bar{\boldsymbol{c}}^{(m)}]\right) \times \prod_l \prod_i \frac{(B_l - 1)!}{(M + B_l - 1)!} \prod_r n_i^{(l)}(r)!, \tag{67}$$

where $n_i^{(l)}(r) = \sum_{m=1}^{M} \delta_{b_i^l, r}$ is how often node i in level l has group label r in all samples. Similarly to before, if we consider the conditional probability of a single partition relabeling $\boldsymbol{c}_l^{(m)}$, but marginalized over the upper levels $l' > l$, we obtain

$$\begin{aligned} P(\boldsymbol{c}_l^{(m)}|\{\bar{\boldsymbol{b}}\}, \{\bar{\boldsymbol{c}}^{(m' \neq m)}\}, \{\boldsymbol{c}_{l'<l}^{(m)}\}) &\propto \sum_{\boldsymbol{c}_{l+1}^{(m)}, \ldots, \boldsymbol{c}_L^{(m)}} P(\{\bar{\boldsymbol{c}}\}|\{\bar{\boldsymbol{b}}\}) \\ &\propto \prod_i \prod_r [n'^l_i(r) + 1]^{\delta_{c_i^{l,m}, r}}, \end{aligned} \tag{68}$$

where $n'^l_i(r)$ are the label counts excluding $\boldsymbol{c}_l^{(m)}$. Just like in the nonhierarchical case, we can write

$$\ln P(\boldsymbol{c}_l^{(m)}|\{\bar{\boldsymbol{b}}\}, \{\bar{\boldsymbol{c}}^{(m' \neq m)}\}, \{\boldsymbol{c}_{l'<l}^{(m)}\}) = \sum_r w_{r,\mu(r)}, \tag{69}$$

up to an unimportant additive constant, where

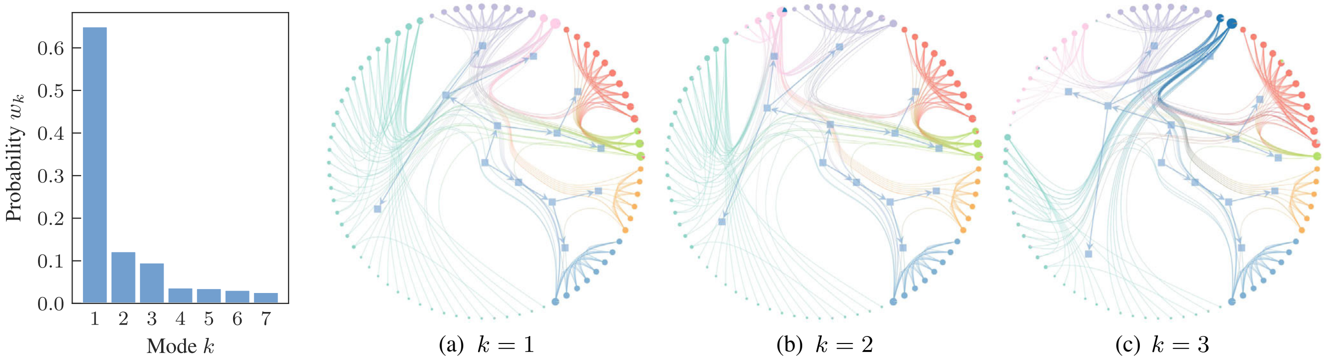
$$w_{rs} = \sum_i \delta_{b_i^l, r} \ln [n'^l_i(s) + 1], \tag{70}$$

and $\mu(r)$ is the bijection that matches the group labels between $\boldsymbol{c}_l^{(m)}$ and $\boldsymbol{b}_l^{(m)}$. Therefore, we can find the maximum of Eq. (69) once more by solving the maximum-weight bipartite matching problem, with weights given by w_{rs}. This method leads to an overall algorithm entirely analogous to the nonhierarchical case, where, starting from some configuration, we remove a sample m from the ensemble and add it again, choosing its labels according to the maximization of Eq. (69), starting from level $l = 1$ and going up to $l = L$, and stopping if such moves no longer increase the posterior probability. Relabeling every sample once takes time $O[M \sum_l (N_l + B_l^3)]$, where N_l and B_l are the typical number of nodes and groups at level l. Typically, the number of groups decreases exponentially with the hierarchical level, $N_l = O(N/\sigma^{l-1})$ with $\sigma > 1$, so we have $L = O(\log N)$ and thus $\sum_l N_l = O(N)$; the entire running time for a single "sweep" over all samples is then simply $O[M(N + B^3)]$, where B is the number of labels in the first hierarchical level.

The mixed random label model of Sec. V can also be generalized in a straightforward manner for hierarchical partitions, i.e.,

$$P(\bar{\boldsymbol{b}}|\bar{\boldsymbol{p}}, \boldsymbol{w}) = \sum_k P(\bar{\boldsymbol{b}}|\bar{\boldsymbol{p}}, k) P(k|\boldsymbol{w}), \tag{71}$$

where, inside a mode k, the partitions are sampled according to the hierarchical random label model given by Eq. (65). The inference algorithm from this point onward is exactly the same as in the nonhierarchical case, where we only need to relabel the hierarchical partitions according to Eq. (69) when we move them between modes.



(a) $k = 1$ (b) $k = 2$ (c) $k = 3$

FIG. 13. Inferred hierarchical partition modes from $M = 10^5$ samples of the hierarchical latent Poisson DC-SBM posterior distribution for the co-occurrence network of characters in the *Les Misérables* novel. The left panel shows the mode fractions w_k, and the right panel shows the three largest modes (a)–(c), with the marginal distributions shown as pie diagrams on the nodes of the network.

In Fig. 13, we show the inferred modes for hierarchical partitions sampled from the posterior distribution using the nested latent Poisson DC-SBM for a co-occurrence network of characters from the *Les Misérables* novel [42]. As this example shows, this algorithm allows us to summarize a multimodal distribution of hierarchical partitions in a rather compact manner. In this particular example, we see that the distribution is fairly dominated by one of the modes [shown in Fig. 13(a)], followed by less probable alternatives.

A. Comparing and finding consensus between hierarchical partitions

If we infer the hierarchical random label model above for two hierarchical partitions $\bar{\boldsymbol{x}}$ and $\bar{\boldsymbol{y}}$, it amounts to solving a recursive maximum bipartite weighted matching problem on every level, starting from $l = 1$ to $l = L$, using as weights the contingency table at each level l,

$$m_{rs}^{(l)} = \sum_{i \in \mathcal{N}_{\boldsymbol{x}^l} \cap \mathcal{N}_{\boldsymbol{y}^l}} \delta_{x_i^l, r} \delta_{y_i^l, s}, \tag{72}$$

where $\mathcal{N}_{\boldsymbol{x}}$ is the set of nodes in partition $\boldsymbol{x}$ (as upper-level partitions might have a disjoint set of nodes), and propagating the matched labels to the upper levels. This approach is equivalent to maximizing the recursive overlap across all levels,

$$w(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}}) = \sum_l \sum_i \delta_{x_i^l, \mu_l(\hat{y}_i^l)}, \tag{73}$$

where, at each level, we need to incorporate the relabeling at the lower levels via

$$\hat{y}_i^l = y_{\mu_{l-1}(i)}^l, \tag{74}$$

where $\boldsymbol{\mu_l}$ is a label bijection at level l, with the boundary condition $\mu_0(i) = i$. This process leads us to the hierarchical maximum overlap distance, defined as

$$d(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}}) = \sum_l N_l - \underset{\boldsymbol{\mu_l}}{\operatorname{argmax}} \sum_i \delta_{x_i^l, \mu_l(\hat{y}_i^l)}, \tag{75}$$

where $N_l = \max(|\mathcal{N}_{\boldsymbol{x}^l}|, |\mathcal{N}_{\boldsymbol{y}^l}|)$. A version of this distance that is normalized in the range [0, 1] can be obtained by dividing it by the largest possible value,

$$\frac{d(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}})}{\sum_l N_l - 1}. \tag{76}$$

It is important to note here that hierarchy levels with a single node, $N_l = 1$, always have a contribution of zero to the distance; therefore, this measure can be applied to infinite hierarchies with $L \to \infty$, as long as any level is eventually grouped into a single group. For hierarchies with a single level, $L = 1$, we recover the maximum overlap distance considered previously, except for the normalized version, which is slightly different, with $d(\boldsymbol{x}, \boldsymbol{y})/(N - 1)$. This normalization is also valid for the nonhierarchical distance since we must always have $d(\boldsymbol{x}, \boldsymbol{y}) < N$. The label matching at level l of the hierarchy can be done in time $O[(q(\boldsymbol{x}_l) + q(\boldsymbol{y}_l))E_m^l + N_l]$, using the sparse version of the Kuhn-Munkres algorithm [26–28], where $E_m^l \le q(\boldsymbol{x}_l)q(\boldsymbol{y}_l)$ is the number of nonzero entries in the contingency matrix m_{rs}. If we assume, once more, the typical case with $N_l = O(N/\sigma^{l-1})$ and $L = O(\log N)$, so that $\sum_l N_l = O(N)$, the overall computation can then be done in time $O[(q(\boldsymbol{x}^1) + q(\boldsymbol{y}^1))E_m^1 + N]$.

Following the same steps as before, we can use the hierarchical maximum overlap distance as an error function $\epsilon(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}}) = d(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}})$ to define a MOC estimator over hierarchical partitions based on the minimization of the mean posterior loss,

$$\hat{\bar{\boldsymbol{b}}} = \underset{\bar{\boldsymbol{b}}}{\operatorname{argmin}} \sum_{\bar{\boldsymbol{b}}'} \epsilon(\bar{\boldsymbol{b}}, \bar{\boldsymbol{b}}') P(\bar{\boldsymbol{b}}' | \boldsymbol{A}). \tag{77}$$

Substituting its definition leads us to a set of self-consistent equations at each level l,

$$\hat{b}_i^l = \underset{r}{\operatorname{argmax}}\ \hat{\pi}_i^l(r|\{\mu_{\boldsymbol{b}}^l\}), \tag{78}$$

with the marginal distributions obtained over the relabeled partitions,

$$\hat{\pi}_i^l(r|\{\mu_{\boldsymbol{b}}^l\}) = \sum_{\bar{\boldsymbol{b}}} \delta_{\mu_{\boldsymbol{b}}^l(\tilde{b}_i^l),r} P(\bar{\boldsymbol{b}}|\boldsymbol{A}), \tag{79}$$

where the relabeling is done in order to maximize the overlap with $\hat{\bar{\boldsymbol{b}}}$,

$$\boldsymbol{\mu}_{\boldsymbol{b}}^l = \underset{\boldsymbol{\mu}}{\operatorname{argmax}} \sum_i \delta_{\hat{b}_i,\mu(\tilde{b}_i)} \tag{80}$$

and where, once again, we need to recursively incorporate the relabelings at the lower levels,

$$\tilde{b}_i^l = b_{\mu_{l-1}(i)}^l. \tag{81}$$

We can define an uncertainty $\sigma_{\hat{\boldsymbol{b}}} \in [0,1]$ for this estimator by inspecting the marginal distributions computed along the way,

$$\sigma_{\hat{\boldsymbol{b}}} = 1 - \frac{1}{N-L} \sum_l \frac{N_l - 1}{N_l} \sum_i \hat{\pi}_i(\hat{b}_i^l|\{\mu_{\boldsymbol{b}}^l\}). \tag{82}$$

In the above sum, we omit levels with $N_l = 1$ since those always have a trivial marginal distribution concentrated on a single group. In practice, we implement this estimator by sampling a set of M hierarchical partitions $\{\bar{\boldsymbol{b}}\}$ from the posterior distribution and then performing the sequential maximizations starting from $l = 1$ to $l = L$,

$$\hat{b}_i^l = \underset{r}{\operatorname{argmax}} \sum_m \delta_{\mu_m(\tilde{b}_i^{l,m}),r}, \tag{83}$$

$$\boldsymbol{\mu}_m^l = \underset{\boldsymbol{\mu}}{\operatorname{argmax}} \sum_r \hat{m}_{r,\mu(r)}^{(l,m)}, \tag{84}$$

where $m_{r,s}^{(l,m)}$ is the contingency table of level l of sample m with $\hat{\boldsymbol{b}}^l$. The final solution is obtained when repeating the above maximization no longer changes the result. Like in the nonhierarchical case, this algorithm yields a local optimum of the optimization problem but not necessarily a global one; therefore, it needs to be repeated multiple times with different initial conditions, and the best result is kept. Since it involves the relabeling over all M hierarchical partitions, the overall algorithmic complexity of a single iteration is $O(MNB + MB^3)$, assuming, once more, the typical case with $N_l = O(N/\sigma^{l-1})$ and $L = O(\log N)$.

VII. MODEL SELECTION AND EVIDENCE APPROXIMATION

If we are interested in comparing two models $\mathcal{M}_1$ and $\mathcal{M}_2$ in their plausibility for generating some network $\boldsymbol{A}$, we can do so by computing the ratio of their posterior probability given the data,

$$\frac{P(\mathcal{M}_1|\boldsymbol{A})}{P(\mathcal{M}_2|\boldsymbol{A})} = \frac{P(\boldsymbol{A}|\mathcal{M}_1)P(\mathcal{M}_1)}{P(\boldsymbol{A}|\mathcal{M}_2)P(\mathcal{M}_2)}. \tag{85}$$

Therefore, if we are *a priori* agnostic about either model with $P(\mathcal{M}_1) = P(\mathcal{M}_2)$, this ratio will be determined by the total probability of the data $P(\boldsymbol{A}|\mathcal{M})$ according to that model. This quantity is called the evidence, and it appears as a normalization constant in the posterior distribution of Eq. (1). For any particular choice of model, it is obtained by summing the joint probability of data and partitions over all possible partitions (we drop the explicit dependence on $\mathcal{M}$ from now on, to unclutter the expressions),

$$P(\boldsymbol{A}) = \sum_{\boldsymbol{b}} P(\boldsymbol{A},\boldsymbol{b}). \tag{86}$$

Unfortunately, the exact computation of this sum is intractable since the number of partitions is too large in most cases of interest. It also cannot be obtained directly from samples of the posterior distribution, which makes its estimation from MCMC also very challenging. To illustrate this, it is useful to write the logarithm of the evidence in the following manner,

$$\ln P(\boldsymbol{A}) = \sum_{\boldsymbol{b}} \pi(\boldsymbol{b}) \ln P(\boldsymbol{A},\boldsymbol{b}) - \sum_{\boldsymbol{b}} \pi(\boldsymbol{b}) \ln \pi(\boldsymbol{b}) \tag{87}$$

$$= \langle \ln P(\boldsymbol{A},\boldsymbol{b}) \rangle + H(b), \tag{88}$$

where

$$\pi(\boldsymbol{b}) = \frac{P(\boldsymbol{A},\boldsymbol{b})}{\sum_{\boldsymbol{b}'} P(\boldsymbol{A},\boldsymbol{b}')} = \frac{P(\boldsymbol{A},\boldsymbol{b})}{P(\boldsymbol{A})} \tag{89}$$

is the posterior distribution of Eq. (1), and

$$\langle \ln P(\boldsymbol{A},\boldsymbol{b}) \rangle = \sum_{\boldsymbol{b}} \pi(\boldsymbol{b}) \ln P(\boldsymbol{A},\boldsymbol{b}) \tag{90}$$

is the mean joint log-probability computed over the posterior distribution; finally,

$$H(b) = -\sum_{\boldsymbol{b}} \pi(\boldsymbol{b}) \ln \pi(\boldsymbol{b}) \tag{91}$$

is the entropy of the posterior distribution. Equation (87) has the shape of a negative Gibbs free energy of a physical ensemble, if we interpret $\langle \ln P(\boldsymbol{A},\boldsymbol{b}) \rangle$ as the mean negative "energy" over the ensemble of partitions. It tells us that what contributes to the evidence is not only the mean joint probability but also the multiplicity of solutions with similar probabilities, which is captured by the posterior entropy. In this formulation, we see that while it is possible to estimate $\langle \ln P(\boldsymbol{A},\boldsymbol{b})\rangle$ from MCMC simply be averaging $\ln P(\boldsymbol{A},\boldsymbol{b})$ for sufficiently many samples, the same approach does not work for the entropy term $H(b)$ since it would require the computation of the log-posterior $\ln \pi(\boldsymbol{b})$ for every sample, something that cannot be done without knowing the normalization constant $P(\boldsymbol{A})$, which is what we want to find in the first place. However, the mixed random label model of Sec. V can be used to fit the posterior distribution, allowing us to compute the entropy term via the inferred model, and we use the rich information gained on its structure to perform model selection. Let us recall that the mixed random label model, when inferred from partitions sampled from $\pi(\boldsymbol{b})$, amounts to an approximation given by

$$\pi(\boldsymbol{b}) \approx \sum_{k,\boldsymbol{c}} P(\boldsymbol{b}|\boldsymbol{c})P(\boldsymbol{c}|k)P(k), \tag{92}$$

where $P(k) = w_k$ determines the mode mixture and

$$P(\boldsymbol{c}|k) = \prod_i p_i^{(k)}(c_i) \tag{93}$$

are the independent marginal distributions of mode k; finally,

$$P(\boldsymbol{b}|\boldsymbol{c}) = \frac{[\boldsymbol{b} \sim \boldsymbol{c}]}{q(\boldsymbol{b})!} \tag{94}$$

is the random relabeling of groups. In most cases we have investigated, the inferred modes tend to be very well separated (otherwise, they would get merged together into a larger mode), such that we can assume

$$\pi(\boldsymbol{b}) \approx \max_{k,\boldsymbol{c}} \; P(\boldsymbol{b}|\boldsymbol{c})P(\boldsymbol{c}|k)P(k). \tag{95}$$

This means we can write the entropy as

$$H(b) \approx H(b,c,k) = H(b|c) + H(c|k) + H(k), \tag{96}$$

where

$$H(k) = -\sum_k w_k \ln w_k \tag{97}$$

is the entropy of the mode mixture distribution,

$$H(c|k) = -\sum_k w_k \sum_{\boldsymbol{c}} P(\boldsymbol{c}|k) \ln P(\boldsymbol{c}|k) \tag{98}$$

$$= -\sum_k w_k \sum_i \sum_r p_i^{(k)}(r) \ln p_i^{(k)}(r) \tag{99}$$

is the entropy of mode k, and

$$H(b|c) = -\sum_{\boldsymbol{c}} P(\boldsymbol{c}) \sum_{\boldsymbol{b}} P(\boldsymbol{b}|\boldsymbol{c}) \ln P(\boldsymbol{b}|\boldsymbol{c}) \tag{100}$$

$$= \sum_{\boldsymbol{c}} P(\boldsymbol{c}) \ln q(\boldsymbol{c})! = \sum_{\boldsymbol{b}} P(\boldsymbol{b}) \ln q(\boldsymbol{b})! \tag{101}$$

is the relabeling entropy. Putting it all together, we have the following approximation for the evidence according to the mixed random label model,

$$\ln P(\boldsymbol{A}) \approx \langle \ln P(\boldsymbol{A},\boldsymbol{b})\rangle + \langle \ln q(\boldsymbol{b})!\rangle - \sum_k w_k \ln w_k - \sum_k w_k \sum_i \sum_r p_i^{(k)}(r) \ln p_i^{(k)}(r). \tag{102}$$

We can extend this for hierarchical partitions in an entirely analogous way, which leads to

$$\ln P(\boldsymbol{A}) \approx \langle \ln P(\boldsymbol{A},\bar{\boldsymbol{b}})\rangle + \sum_l \langle \ln q(\boldsymbol{b}^l)!\rangle - \sum_k w_k \ln w_k - \sum_k w_k \sum_l \sum_i \sum_r p_i^{(l,k)}(r) \ln p_i^{(l,k)}(r). \tag{103}$$

The above quantities are then computed by sampling M partitions from the posterior distribution, using them (or a superset thereof) to compute the first two means $\langle \ln P(\boldsymbol{A},\boldsymbol{b})\rangle$ and $\langle \ln q(\boldsymbol{b})!\rangle$, and then fitting the mixed random label model, from which the parameters $\boldsymbol{w}$ and $\boldsymbol{p}$ are obtained; we then compute the remaining terms.

In Table I, we show the evidence obtained for several SBM variants and data sets, including latent Poisson versions (which require special considerations; see Appendix E). Overall, we find that when considering the Poisson SBMs, degree correction is only favored for larger networks, corroborating a similar, previous analysis based on a less-accurate calculation [25]. This case changes for latent Poisson models, where, for some networks, the balance tips in favor of degree correction. Overall, we find more evidence for the latent Poisson models for all networks considered, which is unsurprising given that they are all simple graphs. Likewise, we always find more evidence for the hierarchical SBMs, which further demonstrates their more flexible nature.

A. Bayesian evidence and the MDL criterion

In this section, we briefly explore some direct connections between Bayesian model selection and the MDL criterion based on information theory [49]. We begin by pointing out the simple fact that the MAP point estimate given by the single-most-likely partition yields a lower bound for the evidence, i.e.,

TABLE I. Description length (negative log-evidence) $\Sigma = -\ln P(A)$ for several networks and SBM variations, with DC and NDC indicating degree correction and no degree correction, respectively. The italic fonts indicate the smallest value for each model class, with the bold fonts indicating the best-fitting model overall. The "single partition" columns correspond to the two-part description length $\Sigma = -\ln P(\boldsymbol{A}, \boldsymbol{b})$ obtained with the best-fitting partition of the Poisson model.

	Poisson				Latent Poisson				Single partition			
	Non-nested		Nested		Non-nested		Nested		Non-nested		Nested	
Data	NDC	DC	NDC	DC	NDC	DC	NDC	DC	NDC	DC	NDC	DC
Karate club [36]	213.1	220.3	*212.6*	221.7	174.0	172.4	**170.6**	171.6	*215.3*	222.7	*215.3*	222.7
Dolphins [43]	522.4	539.3	*522.1*	540.1	480.9	483.6	**477.6**	478.7	*529.6*	544.1	*529.6*	544.1
Les Misérables [42]	674.1	680.1	*667.5*	672.4	513.7	471.0	454.6	402.7	*688.7*	697.6	*688.7*	697.6
Political books [23]	1305.2	1334.4	*1288.8*	1330.8	1188.2	1178.6	**1136.7**	1137.4	1321.9	1343.4	*1317.4*	1343.4
American football [41]	1722.4	1769.2	*1709.7*	1755.7	1427.7	1505.8	**1319.8**	1373.1	1738.9	1785.9	*1733.5*	1780.6
Network scientist [44]	3871.5	3869.5	*3592.6*	3645.1	3728.4	3611.5	3059.9	3043.6	4007.8	3982.2	*3813.4*	3826.2
High school [45]	4530.5	4620.6	*4482.8*	4592.3	4378.1	4421.7	**4257.4**	4307.6	4599.9	4676.8	*4585.9*	4668.2
C. elegans neurons [46]	6968.2	7040.3	*6812.7*	6943.0	6492.3	6485.7	**6048.3**	6411.3	7043.7	7144.4	*6959.5*	7091.3
E-mail [47]	25 020.5	24 845.5	*24 145.3*	24 264.8	24 577.1	24 047.4	23 544.7	**23 002.0**	25 617.1	25 311.2	25 163.8	*25 094.7*
Political blogs [48]	51 389.1	50 638.2	50 528.9	*50 138.0*	47 787.8	46 380.7	46 065.2	**45 006.4**	51 639.1	51 084.1	51 195.2	*50 892.7*

$$P(\boldsymbol{A}) = \sum_{\boldsymbol{b}} P(\boldsymbol{A}, \boldsymbol{b}) \geq \max_{\boldsymbol{b}} P(\boldsymbol{A}, \boldsymbol{b}). \tag{104}$$

Thus, taking into account the full posterior distribution, rather than only its maximum, almost always can be used to compress the data, as we now show. We can see this by first inspecting the usual "two-part" description length,

$$\Sigma_1(\boldsymbol{A}, \boldsymbol{b}) = -\ln P(\boldsymbol{A}, \boldsymbol{b}) \tag{105}$$

$$= -\ln P(\boldsymbol{A}|\boldsymbol{b}) - \ln P(\boldsymbol{b}), \tag{106}$$

which corresponds to the amount of information necessary to describe the data if one first describes the partition $\boldsymbol{b}$ and then, conditioned on it, the network $\boldsymbol{A}$. Therefore, finding the most likely partition $\boldsymbol{b}$ means finding the one that most compresses the network, according to this particular two-part encoding. However, the full posterior distribution gives us a more efficient "one-part" encoding, where no explicit description of the partition is necessary. Simply defining the joint distribution $P(\boldsymbol{A}, \boldsymbol{b})$ means we can compute the marginal probability $P(\boldsymbol{A}) = \sum_{\boldsymbol{b}} P(\boldsymbol{A}, \boldsymbol{b})$, which directly yields a description length

$$\Sigma_2(\boldsymbol{A}) = -\ln P(\boldsymbol{A}). \tag{107}$$

According to Eq. (104), we have

$$\Sigma_2(\boldsymbol{A}) \leq \min_{\boldsymbol{b}} \Sigma_1(\boldsymbol{A}, \boldsymbol{b}), \tag{108}$$

which means that considering all possible partitions can only increase the overall compression achievable. In Table I, we can verify that this holds for all results obtained.

In a slightly more concrete setting, let us consider a transmitter that wants to convey the network $\boldsymbol{A}$ to a receiver, which both know the joint distribution $P(\boldsymbol{A}, \boldsymbol{b})$. According to the two-part code, the transmitter first sends the partition $\boldsymbol{b}$, for that using $-\log_2 P(\boldsymbol{b})$ bits, and then sends the final network using $-\log_2 P(\boldsymbol{A}|\boldsymbol{b})$ bits, using in total $\Sigma_1(\boldsymbol{A}, \boldsymbol{b})/\ln 2$ bits. In practice, this process is achieved, for example, by both the sender and receiver sharing the same two tables of optimal prefix codes derived from $P(\boldsymbol{b})$ and $P(\boldsymbol{A}|\boldsymbol{b})$. On the other hand, using the second one-part code, both the transmitter and receiver share only a single table of optimal prefix codes derived directly from the marginal distribution $P(\boldsymbol{A})$, which means that only $\Sigma_2(\boldsymbol{A})/\ln 2 = -\log_2 P(\boldsymbol{A})$ bits need to be transmitted. In practice, it will be more difficult to construct the one-part code since it involves marginalizing over a high-dimensional distribution, which is intractable via brute force—although our mixed random label model can be used as the basis of an analytical approximation. However, what is important in our model selection context is only that such a code exists; we are not concerned with its computational tractability.

VIII. CONCLUSION

We have shown how the random label model can be used to solve the group identification problem in community detection, allowing us to compute marginal distributions of group membership on the nodes in an unambiguous way. This process led us to the notion of maximum overlap distance as a general way of comparing two network partitions, which we then used as a loss function to obtain the consensus of a population of network partitions.

By investigating the behavior of different loss functions on artificial and empirical ensembles of heterogeneous partitions, we have demonstrated that they can yield inconsistent results due precisely to a lack of uniformity between divisions. We then developed a more comprehensive characterization of the posterior distribution, based on a mixed version of the random label model that is capable of describing multimodal populations of partitions, where multiple consensuses exist at the same time. This kind of structure corresponds to a "multiple truths" phenomenon, where a model can yield diverging hypotheses for the same data. We showed how our method provides a compact representation for structured populations of network partitions and allows us to assess the quality of fit and perform model selection. The latter was achieved by using the multimodal fit of the posterior distribution as a proxy for the computation of its entropy, which is a key, but often elusive ingredient in Bayesian model selection.

Although we have focused on community detection, the methods developed here are applicable for any kind of clustering problem from which a population of answers can be produced. They allowed us to be more detailed in our assessment of the consistency of results when applied to real or artificial data. In particular, we no longer need to rely on "point estimates" that can give a very misleading picture of high-dimensional and structured populations of partitions, even if they attempt to assemble a consensus among them. We achieve this without losing interpretability, as our method yields groupings of partitions that share a local consensus, each telling a different version of how the data might have been generated and weighted according to the statistical evidence available.

APPENDIX A: PROPERTIES OF MAXIMUM OVERLAP DISTANCE

In Sec. III A of the main text, we considered the maximum overlap distance, which corresponds to the minimal classification error, i.e., the smallest possible number of nodes with an incorrect group placement in a partition $\boldsymbol{y}$ if another partition $\boldsymbol{x}$ is assumed to be the correct one. It is defined as

$$d(\boldsymbol{x},\boldsymbol{y}) = N - \max_{\boldsymbol{\mu}} \sum_i \delta_{\mu(x_i),y_i}. \tag{A1}$$

This measure has been considered before in Refs. [31–33], and here we review some of its useful properties.

(1) *Simple interpretation.* Since it quantifies the classification error, it is easy to intuitively understand what the distance is conveying. In particular, its normalized version $d(\boldsymbol{x},\boldsymbol{y})/N$ yields values in the range [0, 1], which can be interpreted as fractions of differing nodes and hence allows the direct comparison between results obtained for partitions of different sizes and numbers of groups.

(2) *Behaves well for unbalanced partitions.* The distance $d(\boldsymbol{x},\boldsymbol{y})$ behaves as one would expect even when the partitions have very different numbers of groups or the number of groups approaches N for either $\boldsymbol{x}$ or $\boldsymbol{y}$, unlike alternatives such as mutual information [50]. More specifically, if we simply increase the number of groups of either partition being compared, this does not spuriously introduce small values of $d(\boldsymbol{x},\boldsymbol{y})$. We see this by noticing that if $q(\boldsymbol{x}) = B$ and $q(\boldsymbol{y}) = N$, the maximum overlap is always $\omega(\boldsymbol{x},\boldsymbol{y}) = B$ since each group in $\boldsymbol{x}$ can be trivially matched with any of the single-node groups in $\boldsymbol{y}$, yielding

$$d(\boldsymbol{x},\boldsymbol{y}) = N - B, \tag{A2}$$

which leads to the maximum normalized distance $d(\boldsymbol{x},\boldsymbol{y})/N \to 1$ as $N \gg B$.

(3) *Simple asymptotic behavior for uncorrelated partitions.* Suppose partitions $\boldsymbol{x}$ and $\boldsymbol{y}$ are sampled independently and uniformly from the set of all possible partitions into $q(\boldsymbol{x})$ and $q(\boldsymbol{y})$ labeled groups, respectively. In this case, as $N \gg 1$, the contingency table will tend to the uniform one with $m_{rs} = N/[q(\boldsymbol{x})q(\boldsymbol{y})]$, which results in the asymptotic normalized distance given by

$$\lim_{N\to\infty} \frac{d(\boldsymbol{x},\boldsymbol{y})}{N} = \frac{1}{\max(q(\boldsymbol{x}), q(\boldsymbol{y}))}. \tag{A3}$$

Although it is not a substitute for a proper hypothesis test (which would need to account for finite values of N), this asymptotic value gives a rule of thumb of how to interpret the distance between two partitions as a strength of statistical correlation.

(4) *Defines a metric space.* The distance $d(\boldsymbol{x},\boldsymbol{y})$ is a proper metric since it fulfills the properties of identity $d(\boldsymbol{x},\boldsymbol{x}) = 0$, non-negativity $d(\boldsymbol{x},\boldsymbol{y}) \geq 0$, symmetry $d(\boldsymbol{x},\boldsymbol{y}) = d(\boldsymbol{y},\boldsymbol{x})$, and most notably, triangle inequality $d(\boldsymbol{x},\boldsymbol{z}) \leq d(\boldsymbol{x},\boldsymbol{y}) + d(\boldsymbol{y},\boldsymbol{z})$ (we offer a simple proof of this in Appendix B). Thus, this notion of distance is well defined and unambiguous, and conforms to intuition.

(5) *Information-theoretic interpretation.* The maximum overlap has a direct information-theoretic interpretation, due to its connection to the random label generative model exposed earlier. According to the model of Eq. (12), the joint probability of observing two partitions $\{\boldsymbol{c}\} = \{\boldsymbol{x},\boldsymbol{y}\}$, up to an arbitrary relabeling of the groups, is given by

$$P(\boldsymbol{x},\boldsymbol{y}) = \frac{2^{\omega(\boldsymbol{x},\boldsymbol{y})}}{[B(B+1)]^N}, \tag{A4}$$

which means that any two partitions have a joint description length

$$\Sigma(\boldsymbol{x},\boldsymbol{y}) = -\log_2 P(\boldsymbol{x},\boldsymbol{y}) \tag{A5}$$

$$= N\log_2[B(B+1)] - \omega(\boldsymbol{x},\boldsymbol{y}), \tag{A6}$$

which measures the amount of information (in bits) necessary to describe both partitions. The above quantity is proportional to the negative value of the maximum overlap $\omega(\boldsymbol{x},\boldsymbol{y})$ and hence is proportional to $d(\boldsymbol{x},\boldsymbol{y})$. (Note that this is not the most efficient encoding scheme based on the maximum overlap; we consider an alternative in Appendix C.)

(6) *Efficient computation.* As discussed previously, computing the maximum overlap involves solving an instance of the maximum bipartite weighted matching problem, with weights given by the contingency table, $w_{rs} = m_{rs}$ (see Fig. 2), which can be done using the Kuhn-Munkres algorithm [27,28]. In its sparse version, the running time is bound by $O[(q(\boldsymbol{x}) + q(\boldsymbol{y}))E_m]$, with $E_m \le q(\boldsymbol{x})q(\boldsymbol{y})$ being the number of nonzero entries in the contingency matrix m_{rs} [26]. Combining this with the work required to build the contingency table itself, the computation of $d(\boldsymbol{x},\boldsymbol{y})$ is bound by $O[(q(\boldsymbol{x}) + q(\boldsymbol{y}))E_m + N]$. Therefore, the running time will depend on whether we expect the number of labels and the density of the contingency table to be much smaller than or comparable to N. In the former case, the maximum matching algorithm takes a comparatively negligible time, and the linear term dominates, yielding a running time $O(N)$. Otherwise, if we have $q(\boldsymbol{x}) = O(N)$ or $q(\boldsymbol{y}) = O(N)$, then $E_m = O(N)$, and hence the running time will be quadratic, $O(N^2)$. However, the latter scenario is atypical when N is very large; therefore, we most often encounter the linear regime, allowing for very fast computations (see Fig. 14).

The maximum overlap distance has been used before in situations where the labeling is unambiguous or the number of labels is so small that exhaustive iteration over label permutations is feasible (e.g., Refs. [5,51]), but, to the best of our knowledge, it is rarely in combination with the maximum bipartite weighted matching algorithm as outlined above (with an exception being Ref. [52], which employed it when comparing with other metrics), which makes it usable in general settings. Instead, more focus has been given to measures such as mutual information (and its several variants) [53] or variation of information (VI) [34], which are based on the contingency table without requiring us to obtain a label matching. As pointed out by Meilă [34], it is not meaningful to talk about the "best" way of comparing partitions without any context since such a task must be unavoidably tied with our ultimate objective. Therefore, a different set of axiomatic conditions might prefer another dissimilarity function, and indeed, it can be proven that no single function can simultaneously fulfil some elementary set of axioms [32]. In particular, since the maximum overlap distance is based only on the number of nodes correctly classified, it ignores the nodes that do not match and hence does not exploit any potential regularity with which the labels are *mismatched*. Other functions, such as variation of information, might provide alternatives that can be used to highlight different properties of partition ensembles. Nevertheless, few other dissimilarity functions share the same ease of interpretation with the maximum overlap distance while possessing its other useful formal properties, such as natural normalization, information-theoretical interpretation, and the fact that it defines a metric space.

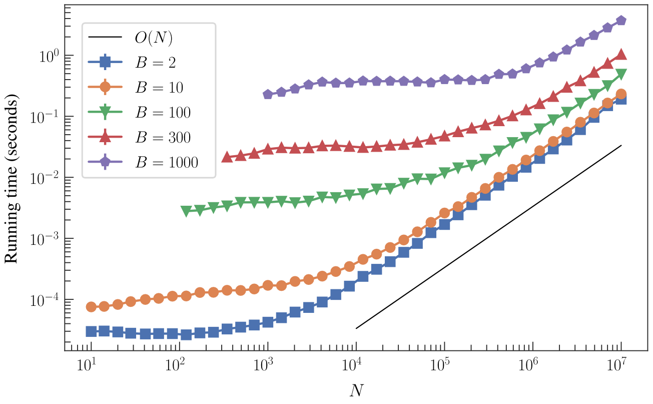


FIG. 14. Time required to compute $d(\boldsymbol{x},\boldsymbol{y})$ for $\boldsymbol{x}$ and $\boldsymbol{y}$ both randomly sampled with $q(\boldsymbol{x}) = q(\boldsymbol{y}) = B$ groups, as shown in the legend, as a function of N, averaged over 100 samples, using an Intel i9-9980HK CPU. The solid line shows an $O(N)$ slope.

Among the alternative partition similarities and dissimilarities, the recently introduced reduced mutual information (RMI) [35] deserves particular mention because, like the maximum overlap distance, it is related to a joint description length of two partitions, which, in the case of RMI, involves encoding the full contingency table. Thus, both similarities can be compared to each other in their own terms, and the most appropriate measure must yield the shortest description length. We perform a succinct comparison between RMI and an overlap-based encoding in Appendix C. We also consider both RMI and VI more closely in the following Appendix.

APPENDIX B: MAXIMUM OVERLAP DISTANCE OBEYS TRIANGLE INEQUALITY

Here, we show that the maximum overlap distance of Eq. (26) obeys triangle inequality, i.e.,

$$d(\boldsymbol{x},\boldsymbol{z}) \le d(\boldsymbol{x},\boldsymbol{y}) + d(\boldsymbol{y},\boldsymbol{z}), \tag{B1}$$

for any set of labeled partitions $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$. Let us consider the maximum overlap

$$\omega(\boldsymbol{x},\boldsymbol{y}) = N - d(\boldsymbol{x},\boldsymbol{y}) = \max_{\boldsymbol{\mu}} \sum_i \delta_{\mu(x_i),y_i}. \tag{B2}$$

Now, for an arbitrary choice of $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$, let us consider the sum

$$\omega(\boldsymbol{x},\boldsymbol{y}) + \omega(\boldsymbol{y},\boldsymbol{z}). \tag{B3}$$

The maximum value either term in the above sum can take is N, corresponding to partitions that are identical up to relabeling, i.e., $[\boldsymbol{x} \sim \boldsymbol{y}] = 1$ or $[\boldsymbol{y} \sim \boldsymbol{z}] = 1$. If we condition on one of the terms taking its maximum value N, the remaining term can take a value of at most $\omega(\boldsymbol{x},\boldsymbol{z})$, either via the first term with $\omega(\boldsymbol{x},\boldsymbol{y}) = \omega(\boldsymbol{x},\boldsymbol{z})$ if $[\boldsymbol{y} \sim \boldsymbol{z}] = 1$ or via the second term with $\omega(\boldsymbol{y},\boldsymbol{z}) = \omega(\boldsymbol{x},\boldsymbol{z})$ if $[\boldsymbol{x} \sim \boldsymbol{y}] = 1$. Thus, we can write

$$\omega(\boldsymbol{x},\boldsymbol{y}) + \omega(\boldsymbol{y},\boldsymbol{z}) \leq N + \omega(\boldsymbol{x},\boldsymbol{z}). \tag{B4}$$

Substituting $\omega(\boldsymbol{x},\boldsymbol{y}) = N - d(\boldsymbol{x},\boldsymbol{y})$ and rearranging gives us Eq. (B1).

APPENDIX C: ENCODING PARTITIONS BASED ON OVERLAP

As described in the main text, the random label model yields a description length for a pair of partitions given by

$$\Sigma(\boldsymbol{x},\boldsymbol{y}) = -\ln P(\boldsymbol{x},\boldsymbol{y}) \tag{C1}$$

$$= N\ln[B(B+1)] - \omega(\boldsymbol{x},\boldsymbol{y})\ln 2. \tag{C2}$$

Likewise, if we observe $\boldsymbol{y}$ and use it to describe partition $\boldsymbol{x}$, the additional amount of information we need to convey is

$$\Sigma(\boldsymbol{x}|\boldsymbol{y}) = -\ln P(\boldsymbol{x}|\boldsymbol{y}) \tag{C3}$$

$$= -\ln P(\boldsymbol{x},\boldsymbol{y})/P(\boldsymbol{y}) \tag{C4}$$

$$= N\ln(B+1) - \omega(\boldsymbol{x},\boldsymbol{y})\ln 2, \tag{C5}$$

where we have used $P(\boldsymbol{y}) = 1/B^N$ from Eq. (12). From this information, we note that this encoding is suboptimal in the sense that, even when the overlapping is maximal with $\omega(\boldsymbol{x},\boldsymbol{y}) = N$, the additional information needed to encode $\boldsymbol{x}$ is $\Sigma(\boldsymbol{x}|\boldsymbol{y}) = N\ln[(B+1)/2]$, which scales as $O(N)$ when $B > 1$.

Nevertheless, we can develop a different encoding that is more efficient at using the overlap information. We do so by incorporating it as an explicit parameter as follows:

(1) We sample an overlap value ω uniformly in the range $[1,N]$, such that

$$P(\omega) = \frac{1}{N}. \tag{C6}$$

(2) We choose a subset V_ω of the N nodes of size ω, uniformly with probability

$$P(V_\omega|\omega) = \binom{N}{\omega}^{-1}. \tag{C7}$$

(3) For the nodes in V_ω, we sample a partition $\boldsymbol{z}$ with probability

$$P(\boldsymbol{z}|V_\omega,\boldsymbol{\gamma}) = \prod_{i\in V_\omega}\gamma_{z_i}, \tag{C8}$$

which leads to a marginal distribution

$$P(\boldsymbol{z}|V_\omega) = \int P(\boldsymbol{z}|V_\omega,\boldsymbol{\gamma})P(\boldsymbol{\gamma})\mathrm{d}\boldsymbol{\gamma} \tag{C9}$$

$$= \binom{\omega+B-1}{\omega}^{-1}\frac{\omega!}{\prod_r n_z(r)!}, \tag{C10}$$

where $n_z(r) = \sum_{i\in V_\omega}\delta_{z_i,r}$, assuming a uniform prior $P(\boldsymbol{\gamma}) = (B-1)!$.

(4) For the remaining $N-\omega$ nodes not in V_ω, we sample the values of partitions $\boldsymbol{x}$ and $\boldsymbol{y}$ analogously, i.e.,

$$P(\boldsymbol{x}|V_\omega) = \binom{N-\omega+B-1}{N-\omega}^{-1}\frac{(N-\omega)!}{\prod_r n_x(r)!},$$
$$P(\boldsymbol{y}|V_\omega) = \binom{N-\omega+B-1}{N-\omega}^{-1}\frac{(N-\omega)!}{\prod_r n_y(r)!},$$

with $n_x(r) = \sum_{i\notin V_\omega}\delta_{x_i,r}$ and $n_y(r) = \sum_{i\notin V_\omega}\delta_{y_i,r}$.

(5) For the nodes $i \in V_\omega$, we set $x_i = y_i = z_i$, and we choose a label bijection $\boldsymbol{\mu}$ uniformly at random from the set of size $B!$ and use it to relabel either $\boldsymbol{x}$ or $\boldsymbol{y}$ arbitrarily.

In the end, this model generates partitions $\boldsymbol{x}$ and $\boldsymbol{y}$ that have an overlap of at least ω, although the actual overlap can be larger by chance. The scheme above allows groups to be unpopulated in the final partition, which is suboptimal, but this can be neglected for our current purpose. The final joint probability of this scheme is

$$P(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z},V_\omega,\omega,\boldsymbol{\mu}) = P(\boldsymbol{x}|V_\omega)P(\boldsymbol{y}|V_\omega)P(\boldsymbol{z}|V_\omega) \times P(V_\omega|\omega)P(\omega)P(\boldsymbol{\mu}), \tag{C11}$$

which leads to a description length

$$\begin{aligned}\Sigma(\boldsymbol{x},\boldsymbol{y},z,V_\omega,\omega,\boldsymbol{\mu})&\\ &= -\ln P(\boldsymbol{x},\boldsymbol{y},z,V_\omega,\omega,\boldsymbol{\mu})\\ &= 2\ln\binom{N-\omega+B-1}{N-\omega} + \ln\binom{\omega+B-1}{\omega}\\ &\quad + \ln\frac{(N-\omega)!}{\prod_r n_x(r)!} + \ln\frac{(N-\omega)!}{\prod_r n_y(r)!} + \ln\frac{\omega!}{\prod_r n_z(r)!}\\ &\quad + \ln\binom{N}{\omega} + \ln N + \ln B!. \end{aligned} \tag{C12}$$

The minimum description length for $\boldsymbol{x}$ and $\boldsymbol{y}$ is given by

$$\Sigma(\boldsymbol{x},\boldsymbol{y}) = \min_{z,V_\omega,\omega,\boldsymbol{\mu}} \Sigma(\boldsymbol{x},\boldsymbol{y},z,V_\omega,\omega,\boldsymbol{\mu}), \tag{C13}$$

which corresponds simply to finding the maximum overlap $\omega(\boldsymbol{x},\boldsymbol{y})$ and the corresponding label matching between $\boldsymbol{x}$ and $\boldsymbol{y}$ from which V_ω, z, and μ can be derived. It is easy to see now that if the overlap is maximal with $\omega = N$, the description length amounts to

$$\begin{aligned}\Sigma(\boldsymbol{x},\boldsymbol{y}) = \ln\binom{N+B-1}{N} + \ln N! - \sum_r \ln n_y(r)!&\\ + \ln N + \ln B!,&\end{aligned} \tag{C14}$$

where we have arbitrarily chosen $\boldsymbol{y}$ as the reference partition but without loss of generality. Hence, if we subtract the necessary information required to describe $\boldsymbol{y}$, given by

$$-\ln P(\boldsymbol{y}) = \ln\binom{N+B-1}{N} + \ln N! - \sum_r \ln n_y(r)!, \tag{C15}$$

we are left with negligible logarithmic terms

$$\Sigma(\boldsymbol{x}|\boldsymbol{y}) = \ln N + \ln B!, \tag{C16}$$

meaning that the additional information needed to describe $\boldsymbol{x}$ given $\boldsymbol{y}$ is vanishingly small with respect to N, and hence the code is efficient in this case.

It is instructive to compare the above scheme with the RMI encoding recently proposed in Ref. [35]. It corresponds to a three-part scheme, where one first encodes partition $\boldsymbol{y}$, then the full contingency table between both partitions m_{rs}, and finally the remaining partition $\boldsymbol{x}$, leading to a description length

$$\begin{aligned}\Sigma'_{\mathrm{RMI}}(\boldsymbol{x},\boldsymbol{y}) = \ln\binom{N-1}{B_y+1} + \ln\binom{N-1}{B_x+1} + \ln\frac{N!}{\prod_r n_y(r)!}&\\ + \sum_r \ln\frac{n_x(r)!}{\prod_s m_{rs}} + \ln\Omega(\boldsymbol{n}_x,\boldsymbol{n}_y),&\end{aligned} \tag{C17}$$

where B_x and B_y are the number of labels in partitions $\boldsymbol{x}$ and $\boldsymbol{y}$ and $\Omega(\boldsymbol{n}_x,\boldsymbol{n}_y)$ is the number of possible contingency tables, with row and column sums given by $\boldsymbol{n}_x$ and $\boldsymbol{n}_y$, which cannot be computed in closed form but for which approximations are available (see Ref. [35]). Note that the encoding above is not symmetric; i.e., in general,

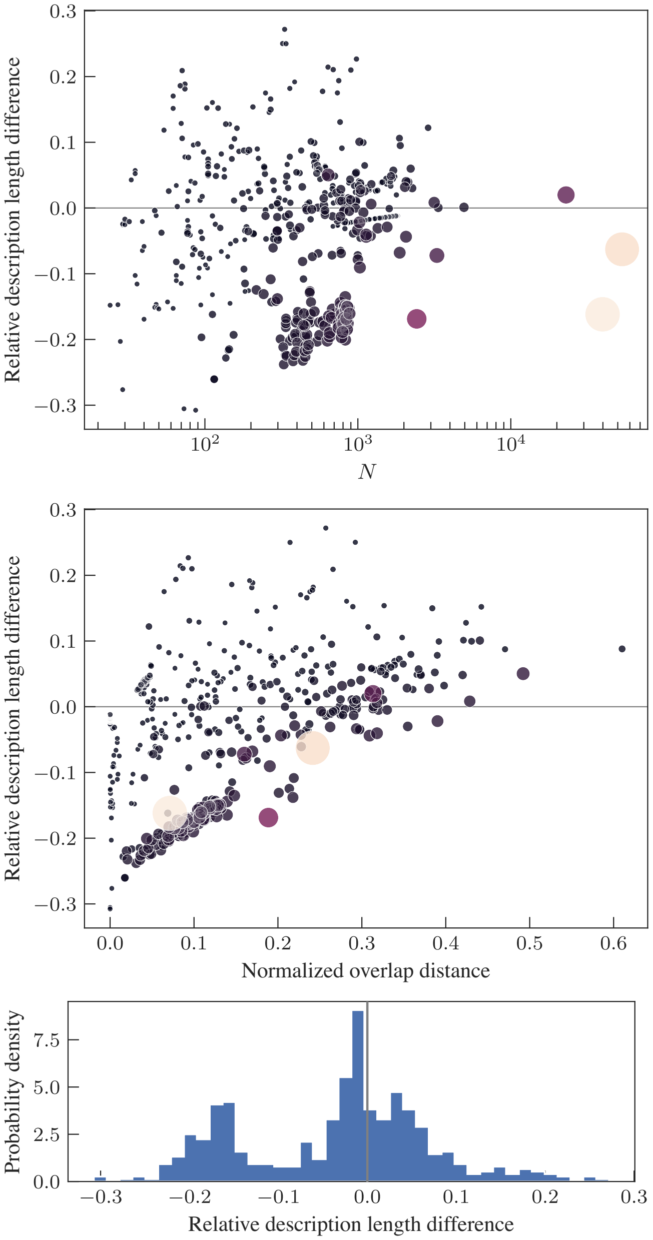


FIG. 15. (Top panel) Average relative description length difference $(\Sigma - \Sigma_{\mathrm{RMI}})/\max(\Sigma,\Sigma_{\mathrm{RMI}})$ between maximum overlap and RMI encodings for empirical networks with N nodes, averaged over pairs of partitions independently sampled from the Poisson DC-SBM posterior distribution. The point size and color indicate the size of the network. (Middle panel) Like the top panel, but with the mean normalized overlap distance computed for each network. (Bottom panel) Histogram of average relative description length differences over all empirical networks.

$\Sigma_{\rm RMI}(\boldsymbol{x},\boldsymbol{y}) \neq \Sigma_{\rm RMI}(\boldsymbol{y},\boldsymbol{x})$, as the overall description length will depend on which partition is encoded first [although the relative description length $\Sigma_{\rm RMI}(\boldsymbol{x}) - \Sigma_{\rm RMI}(\boldsymbol{x},\boldsymbol{y})$ is always symmetric]. Therefore, the minimum description length amounts to choosing the optimal partition to first encode

$$\Sigma_{\rm RMI}(\boldsymbol{x},\boldsymbol{y}) = \min\,[\Sigma'_{\rm RMI}(\boldsymbol{x},\boldsymbol{y}), \Sigma'_{\rm RMI}(\boldsymbol{y},\boldsymbol{x})]. \tag{C18}$$

In Fig. 15, we compare the compression of two partitions sampled independently from the DC-SBM posterior distribution of 571 empirical networks selected from the Konect [54] and CommunityFitNet [55] repositories. Overall, we observe somewhat mixed results, with the overlap encoding providing a better compression for around 61% of the networks. As we might expect, the overlap encoding tends to provide a better description if the overlap between partitions is very high, such that a full description of the nonmatching nodes becomes superfluous. Otherwise, for highly differing partitions, the RMI encoding is able to capture similarities more efficiently.

APPENDIX D: COMPARISON WITH DIMENSIONALITY REDUCTION

The clustering algorithm presented in Sec. V of the main text is based on a particular definition of what a mode is, according to the random label model presented in Sec. III. As has been shown in Fig. 9, there is an intimate relationship between the clusters founds and the metric space of partitions as defined by the maximum overlap distance, such that dimensionality reduction algorithms like UMAP tend to identify the same clusters. One may wonder, however, if this picture changes if we consider another underlying metric space defined by a different distance function. To give a glimpse into this question, in Fig. 16, we show the results of dimensionality reduction using both the variation and information and reduced mutual information [56] functions, both of which make use of the entire contingency table when comparing partitions. As we can see, not only is the overall multimodal structure preserved, but also the composition of the modes is compatible with what was obtained in Fig. 9, showing that the existence of the clusters is not intrinsically tied to the modeling choices made but is, in fact, a property of the data that can be probed in different ways. Naturally, the local shapes and relative positions of the modes vary according to the distance used—and, in fact, even across different runs of the UMAP algorithm since it is nondeterministic.

We stress that the approach we present in the main text offers some advantages over dimensionality reduction: (1) From the beginning, we know what the identified modes mean, and this is not something that needs to be interpreted *a posteriori*. (2) Clustering is performed in a nonparametric manner, without having to decide on an embedding dimension or even the number of clusters that need to be found. Dimensionality reduction, on the other hand, comprises only an intermediary step that yields an input to a surrogate clustering algorithm, like k-means, which is often parametric.

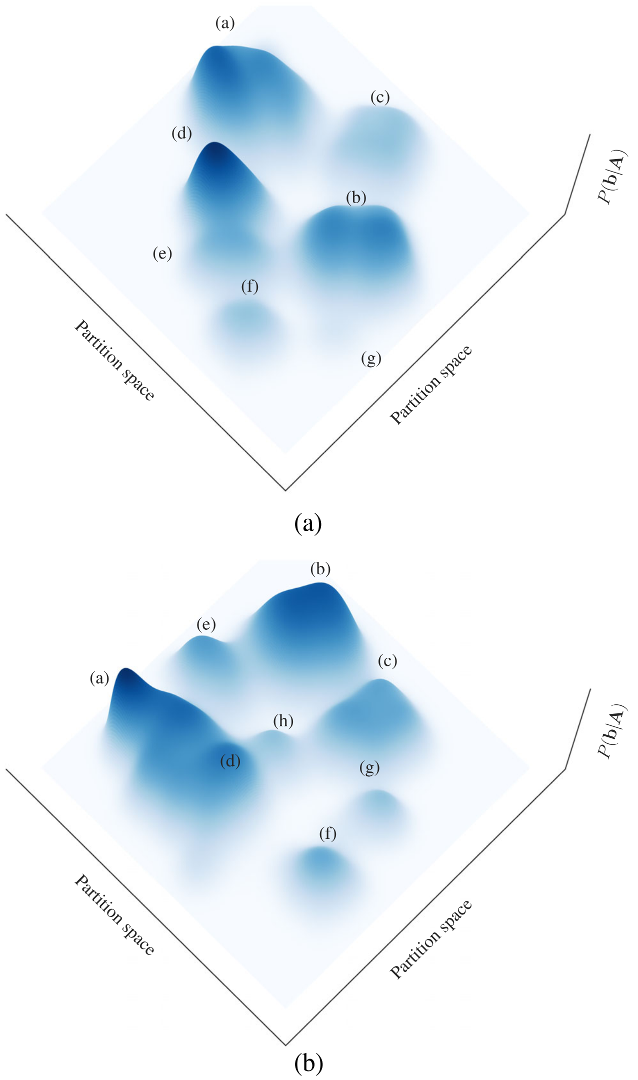


FIG. 16. Projection of the partition distribution in two dimensions according to the UMAP dimensionality reduction algorithm [17], for the same data of Fig. 9, using (a) the variation of information and (b) the (negative) reduced mutual information as dissimilarity functions. The labels indicate a correspondence of the modes with those found in Fig. 9 according to the majority of partitions.

APPENDIX E: EVIDENCE FOR LATENT POISSON SBMs

The latent Poisson SBMs of Ref. [39] are generative models for simple graphs, where, at first, a multigraph $\boldsymbol{G}$ is generated with probability

$$P(\boldsymbol{G}|\boldsymbol{b}) \tag{E1}$$

from a Poisson SBM, and then a simple graph is obtained by collapsing the multiedges to simple edges with

$$P(A_{ij}|\boldsymbol{G}) = \begin{cases} 1 & \text{if } i \neq j \text{ and } G_{ij} > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{E2}$$

The joint posterior distribution of partitions and latent multiedges is then

$$P(\boldsymbol{b}, \boldsymbol{G}|\boldsymbol{A}) = \frac{P(\boldsymbol{A}|\boldsymbol{G})P(\boldsymbol{G}|\boldsymbol{b})P(\boldsymbol{b})}{P(\boldsymbol{A})}, \tag{E3}$$

with evidence given by

$$P(\boldsymbol{A}) = \sum_{\boldsymbol{b},\boldsymbol{G}} P(\boldsymbol{A}, \boldsymbol{G}, \boldsymbol{b}). \tag{E4}$$

Because of the latent multiedges, we need to approximate the evidence in a similar, but different manner. We write the log evidence as

$$\ln P(\boldsymbol{A}) = \sum_{\boldsymbol{b},\boldsymbol{G}} \pi(\boldsymbol{b}, \boldsymbol{G}) \ln P(\boldsymbol{A}, \boldsymbol{G}, \boldsymbol{b}) - \sum_{\boldsymbol{b},\boldsymbol{G}} \pi(\boldsymbol{b}, \boldsymbol{G}) \ln \pi(\boldsymbol{G}, \boldsymbol{b}) \tag{E5}$$

$$= \langle \ln P(\boldsymbol{A}, \boldsymbol{G}, \boldsymbol{b}) \rangle + H(b, G), \tag{E6}$$

where

$$\pi(\boldsymbol{G}, \boldsymbol{b}) = \frac{P(\boldsymbol{A}, \boldsymbol{G}, \boldsymbol{b})}{\sum_{\boldsymbol{G}',\boldsymbol{b}'} P(\boldsymbol{A}, \boldsymbol{G}', \boldsymbol{b}')} \tag{E7}$$

is the joint posterior distribution. For our approximation, we assume the factorization

$$\pi(\boldsymbol{G}, \boldsymbol{b}) \approx \pi(\boldsymbol{G})\pi(\boldsymbol{b}), \tag{E8}$$

together with the "mean field" over the latent multiedges,

$$\pi(\boldsymbol{G}) = \prod_{i \leq j} q_{ij}(G_{ij}), \tag{E9}$$

with the marginals estimated via MCMC,

$$q_{ij}(x) = \sum_{\boldsymbol{G},\boldsymbol{b}} \delta_{G_{ij},x} \pi(\boldsymbol{G}, \boldsymbol{b}), \tag{E10}$$

so that the latent edge entropy can be computed as

$$H(G) = -\sum_{i \leq j} \sum_{x} q_{ij}(x) \ln q_{ij}(x). \tag{E11}$$

From this calculation, we obtain the final approximation,

$$\ln P(\boldsymbol{A}) = \langle \ln P(\boldsymbol{A}, \boldsymbol{G}, \boldsymbol{b}) \rangle + H(b) + H(G), \tag{E12}$$

where $H(b)$ is computed using the mixed random label models as done in the main text. The approximation for the hierarchical model follows analogously.

[1] S. Fortunato, *Community Detection in Graphs*, Phys. Rep. **486**, 75 (2010).

[2] S. Fortunato and D. Hric, *Community Detection in Networks: A User Guide*, Phys. Rep. **659**, 1 (2016).

[3] B. H. Good, Y.-A. de Montjoye, and A. Clauset, *Performance of Modularity Maximization in Practical Contexts*, Phys. Rev. E **81**, 046106 (2010).

[4] U. Brandes, D. Delling, M. Gaertler, R. Görke, M. Hoefer, Z. Nikoloski, and D. Wagner, *On Modularity-NP-Completeness and Beyond*, Universität Karlsruhe (TH), Tech. Rep. **19**, 2006 (2006).

[5] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, *Asymptotic Analysis of the Stochastic Block Model for Modular Networks and Its Algorithmic Applications*, Phys. Rev. E **84**, 066106 (2011).

[6] R. Guimerà and M. Sales-Pardo, *Missing and Spurious Interactions and the Reconstruction of Complex Networks*, Proc. Natl. Acad. Sci. U.S.A. **106**, 22073 (2009).

[7] A. Clauset, C. Moore, and M. E. J. Newman, *Hierarchical Structure and the Prediction of Missing Links in Networks*, Nature (London) **453**, 98 (2008).

[8] J. Calatayud, R. Bernardo-Madrid, M. Neuman, A. Rojas, and M. Rosvall, *Exploring the Solution Landscape Enables More Reliable Network Community Detection*, Phys. Rev. E **100**, 052308 (2019).

[9] M. A. Riolo and M. E. J. Newman, *Consistency of Community Structure in Complex Networks*, Phys. Rev. E **101**, 052306 (2020).

[10] A. Strehl and J. Ghosh, *Cluster Ensembles—A Knowledge Reuse Framework for Combining Multiple Partitions*, J. Mach. Learn. Res. **3**, 583 (2002).

[11] A. Topchy, A. K. Jain, and W. Punch, *Clustering Ensembles: Models of Consensus and Weak Partitions*, IEEE Trans. Pattern Analysis Machine Intelligence **27**, 1866 (2005).

[12] A. Goder and V. Filkov, *Consensus Clustering Algorithms: Comparison and Refinement*, in *2008 Proceedings of the Workshop on Algorithm Engineering and Experiments (ALENEX)* (Society for Industrial and Applied Mathematics, Philadelphia, PA, 2008), pp. 109–117, https://doi.org/10.1137/1.9781611972887.11.

[13] A. Lancichinetti and S. Fortunato, *Consensus Clustering in Complex Networks*, Sci. Rep. **2**, 336 (2012).

[14] P. Zhang and C. Moore, *Scalable Detection of Statistically Significant Communities and Hierarchies, Using Message Passing for Modularity*, Proc. Natl. Acad. Sci. U.S.A. **111**, 18144 (2014).

[15] A. Tandon, A. Albeshri, V. Thayananthan, W. Alhalabi, and S. Fortunato, *Fast Consensus Clustering in Complex Networks*, Phys. Rev. E **99**, 042301 (2019).

[16] L. van der Maaten and G. Hinton, *Visualizing Data Using t-SNE*, J. Machine Learning Research **9**, 2579 (2008).

[17] L. McInnes, J. Healy, and J. Melville, *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*, arXiv:1802.03426.

[18] T. P. Peixoto, *Bayesian Stochastic Blockmodeling*, in *Advances in Network Clustering and Blockmodeling* (John Wiley & Sons, New York, 2019), pp. 289–332.

[19] P. W. Holland, K. Blackmond Laskey, and S. Leinhardt, *Stochastic Blockmodels: First Steps*, Soc. Networks **5**, 109 (1983).

[20] T. P. Peixoto, *Efficient Monte Carlo and Greedy Heuristic for the Inference of Stochastic Block Models*, Phys. Rev. E **89**, 012804 (2014).

[21] M. A. Riolo, G. T. Cantwell, G. Reinert, and M. E. J. Newman, *Efficient Method for Estimating the Number of Communities in a Network*, Phys. Rev. E **96**, 032310 (2017).

[22] T. P. Peixoto, *Merge-Split Markov Chain Monte Carlo for Community Detection*, Phys. Rev. E **102**, 012305 (2020).

[23] V. Krebs, *Political Books Network*, http://www-personal.umich.edu/~mejn/netdata/.

[24] B. Karrer and M. E. J. Newman, *Stochastic Blockmodels and Community Structure in Networks*, Phys. Rev. E **83**, 016107 (2011).

[25] T. P. Peixoto, *Nonparametric Bayesian Inference of the Microcanonical Stochastic Block Model*, Phys. Rev. E **95**, 012317 (2017).

[26] L. Ramshaw and R. E. Tarjan, *On Minimum-Cost Assignments in Unbalanced Bipartite Graphs*, HP Labs, Palo Alto, CA, Tech. Rep. HPL-2012-40R1 (2012).

[27] H. W. Kuhn, *The Hungarian Method for the Assignment Problem*, Naval research logistics quarterly **2**, 83 (1955).

[28] J. Munkres, *Algorithms for the Assignment and Transportation Problems*, J. Soc. Indust. Appl. Math. **5**, 32 (1957).

[29] We offer a freely available reference C++ implementation of every algorithm described in this work as part of the `graph-tool` PYTHON library [30].

[30] T. P. Peixoto, The `Graph-Tool` Python Library, https://graph-tool.skewed.de.

[31] M. Meilă and D. Heckerman, *An Experimental Comparison of Model-Based Clustering Methods*, Mach. Learn. **42**, 9 (2001).

[32] M. Meilă, *Comparing Clusterings: An Axiomatic View*, in *Proceedings of the 22nd International Conference on Machine Learning* (2005), pp. 577–584, https://doi.org/10.1145/1102351.1102424.

[33] M. Meilă, *Comparing Clusterings—An Information Based Distance*, J. Multivariate Anal. **98**, 873 (2007).

[34] M. Meilă, *Comparing Clusterings by the Variation of Information*, in *Learning Theory and Kernel Machines*, Lecture Notes in Computer Science No. 2777, edited by B. Schölkopf and M. K. Warmuth (Springer, Berlin, Heidelberg, 2003), pp. 173–187.

[35] M. E. J. Newman, G. T. Cantwell, and J.-G. Young, *Improved Mutual Information Measure for Clustering, Classification, and Community Detection*, Phys. Rev. E **101**, 042304 (2020).

[36] W. W. Zachary, *An Information Flow Model for Conflict and Fission in Small Groups*, J. Anthropol. Res. **33**, 452 (1977).

[37] T. P. Peixoto, *Parsimonious Module Inference in Large Networks*, Phys. Rev. Lett. **110**, 148701 (2013).

[38] T. P. Peixoto, *Hierarchical Block Structures and High-Resolution Model Selection in Large Networks*, Phys. Rev. X **4**, 011047 (2014).

[39] T. P. Peixoto, *Latent Poisson Models for Networks with Heterogeneous Density*, Phys. Rev. E **102**, 012309 (2020).

[40] S. Geman, E. Bienenstock, and R. Doursat, *Neural Networks and the Bias/Variance Dilemma*, Neural Comput. 1992) 1 ,**4**).

[41] M. Girvan and M. E. J. Newman, *Community Structure in Social and Biological Networks*, Proc. Natl. Acad. Sci. U.S.A. **99**, 7821 (2002).

[42] D. E. Knuth, *The Stanford GraphBase: A Platform for Combinatorial Computing*, 1st ed. (Addison-Wesley Professional, Reading, Mass, 1993).

[43] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, *The Bottlenose Dolphin Community of Doubtful Sound Features a Large Proportion of Long-Lasting Associations*, Behav. Ecol. Sociobiol. **54**, 396 (2003).

[44] M. E. J. Newman, *Finding Community Structure in Networks Using the Eigenvectors of Matrices*, Phys. Rev. E **74**, 036104 (2006).

[45] K. M. Harris, C. T. Halpern, E. Whitsel, J. Hussey, J. Tabor, P. Entzel, and J. R. Udry, *The National Longitudinal Study of Adolescent to Adult Health: Research Design*, see http://www.cpc.unc.edu/projects/addhealth/design (accessed 9 April 2015).

[46] J. G. White, E. Southgate, J. N. Thomson, and S. Brenner, *The Structure of the Nervous System of the Nematode Caenorhabditis elegans*, Phil. Trans. R. Soc. B **314**, 1 (1986).

[47] R. Guimerà, L. Danon, A. Díaz-Guilera, F. Giralt, and A. Arenas, *Self-Similar Community Structure in a Network of Human Interactions*, Phys. Rev. E **68**, 065103(R) (2003).

[48] L. A. Adamic and N. Glance, *The Political Blogosphere and the 2004 U.S. Election: Divided They Blog*, in *Proceedings of the 3rd International Workshop on Link Discovery*, LinkKDD '05 (ACM, New York, NY, 2005), pp. 36–43.

[49] P. D. Grünwald, *The Minimum Description Length Principle* (MIT Press, Cambridge, MA, 2007).

[50] A. J. Gates, I. B. Wood, W. P. Hetrick, and Y.-Y. Ahn, *Element-Centric Clustering Comparison Unifies Overlaps and Hierarchy*, Sci. Rep. **9**, 8574 (2019).

[51] P. Zhang, C. Moore, and M. E. J. Newman, *Community Detection in Networks with Unequal Groups*, Phys. Rev. E **93**, 012303 (2016).

[52] M. Rezaei and P. Fränti, *Set Matching Measures for External Cluster Validity*, IEEE Trans. Knowledge Data Eng. **28**, 2173 (2016).

[53] E. Amigó, J. Gonzalo, J. Artiles, and F. Verdejo, *A Comparison of Extrinsic Clustering Evaluation Metrics*

Based on Formal Constraints, Information storage and retrieval **12**, 461 (2009).

[54] J. Kunegis, *KONECT: The Koblenz Network Collection*, in *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13 Companion (ACM, New York, NY, 2013), pp. 1343–1350.

[55] A. Ghasemian, H. Hosseinmardi, and A. Clauset, *Evaluating Overfit and Underfit in Models of Network Community Structure*, IEEE Trans. Knowledge Data Eng. **32**, 1722 (2020).

[56] The reduced mutual information is not a metric distance, since it does not obey triangle inequality, hence it is not really suitable for use with UMAP, which requires a true metric. Nevertheless, the results obtained are robust even to this inconsistency.