Nonparametric Bayesian inference of the microcanonical stochastic block model

Tiago P. Peixoto*
Department of Mathematical Sciences and Centre for Networks and Collective Behaviour, University of Bath, Claverton Down, Bath BA2 7AY, United Kingdom and ISI Foundation, Via Alassio 11/c, 10126 Torino, Italy

A principled approach to characterize the hidden modular structure of networks is to formulate generative models, and then infer their parameters from data. When the desired structure is composed of modules or "communities", a suitable choice for this task is the stochastic block model (SBM), where nodes are divided into groups, and the placement of edges is conditioned on the group memberships. Here, we present a nonparametric Bayesian method to infer the modular structure of empirical networks, including the number of modules and their hierarchical organization. We focus on a microcanonical variant of the SBM, where the structure is imposed via hard constraints. We show how this simple model variation allows simultaneously for two important improvements over more traditional inference approaches: 1. Deeper Bayesian hierarchies, with noninformative priors replaced by sequences of priors and hyperpriors, that not only remove limitations that seriously degrade the inference on large networks, but also reveal structures at multiple scales; 2. A very efficient inference algorithm that scales well not only for networks with a large number of nodes and edges, but also with an *unlimited number of modules*. We show also how this approach can be used to sample modular hierarchies from the posterior distribution, as well as to perform model selection. Furthermore, we expose a direct equivalence between our microcanonical approach and alternative derivations based on the canonical SBM.

I. INTRODUCTION

One of the most basic goals in the study of social, biological and technological networks is the characterization of their structural patterns. As these systems become large, this quickly becomes a nontrivial problem, as naive methods of inspection are no longer useful, and simple statistics often hide crucial information. A popular approach to this problem is the development of methods that divide the network by grouping together nodes that share similar features, thereby reducing it to a more manageable size, and in the process revealing any latent modular organization. This is the core idea behind a very large number of heuristic methods proposed in the last decade and a half [1, 2], which despite sharing the same motivation differ substantially from each other, due mainly to the various ways this intuitive idea can be implemented concretely. Over time it has become clear that most of these methods are marred by serious limitations, such as the incapacity of distinguishing structure from noise [3] and to find small structures in large systems [4], as well as the fact that the same method often yields multiple diverging results for the same network [5], and that the outcomes of most methods agree neither with each other [2] nor with side information [6].

Like some more recent works in this area, here we follow a different and arguably more principled path, designed to overcome some of these limitations. Namely, instead of formulating heuristics, we construct probabilistic generative models of networks, that include the aforementioned idea of modular structure as parameters to the model. The modular organization is then determined by inferring these parameters from data, using well-founded methods from Bayesian inference and statistical physics. In this context, the problem of separating structure from noise is dealt with by employing nonparametric inference, where generative processes for the model parameters are also formulated via prior likelihoods. Additionally, the comparison of different modular partitions — obtained either from the same or from different models incorporating potentially different ideas about modular organization — can be performed probabilistically, and amount to a comparison of alternative generative hypotheses according to statistical evidence.

In this work, we focus on a specific family of generative models based on the stochastic block model (SBM) [7], where nodes are divided into groups, and the edges are placed randomly between nodes, with probabilities that depend on their group memberships. In particular, we consider a *microcanonical* variation of this family, where the structural constraints are imposed strictly across the ensemble, as opposed to only on average, as is more typically done. We show how this approach makes it easier to incorporate more elaborate generative models, where parameters are sampled from conditioned prior likelihoods, which themselves are sampled from hyperprior distributions. This yields a more powerful method that reveals the hierarchical organization of networks in multiple scales, and has a much increased capacity of finding statistically significant structures in large data. Furthermore, we show how this particular formulation allows for a very efficient inference algorithm that scales well not only for networks with a large number of nodes and edges, but also with an unlimited number of modules — in contrast to the majority of other similar inference algorithms that become increasingly slower as the number of groups becomes large.

* t.peixoto@bath.ac.uk

The approach taken here builds upon ideas from previous work [8–10], but here we focus on obtaining hierarchical network partitions that are *sampled* from the posterior distribution, instead of finding only the most likely partition, which requires a different ansatz. We also show how model selection can be used to choose between different model variants according to the statistical evidence available in the data, and how the method fares for a variety of empirical networks. Furthermore we show that the microcanonical formulation used here is — in its most basic form — equivalent to a specific Bayesian formulation of the "canonical" SBM, and thus we establish a bridge between both approaches.

The paper is divided as follows. We begin in Sec. II with the microcanonical SBM, and follow in Sec. III with the outline of the nonparametric inference approach, by describing in turn the priors and hyperpriors of the different set of parameters. In Sec. IV we show how the microcanonical formulation is related to the more usual canonical approach, and in Sec. V we analyze the limitations of the inference procedure, and we show how the hierarchical approach is capable of finding a much larger number of groups in large networks. In Sec. VI we present an efficient MCMC algorithm to sample hierarchical partitions from the posterior distribution. In Sec. VII we show how different model variations can be compared, and in Sec. VIII we show how the same variations behave for empirical networks. We finalize in Sec. IX with a discussion.

II. THE MICROCANONICAL DEGREE-CORRECTED SBM

We begin with a "degree-corrected" version of the SBM [11] (DC-SBM), where in addition to the modular structure, the networks generated possess a prescribed degree sequence. However, differently from its original definition, here we assume that the degree sequence is fixed exactly, instead of only in expectation. We will see later that the non-degree-corrected version of the model (NDC-SBM) can be obtained from this more general formulation as a special case.

The parameters of the model are the partition $\boldsymbol{b} = \{b_i\}$ of N nodes into B groups, where $b_i \in [1, B]$ is the group membership of node i, the degree sequence $\boldsymbol{k} = \{k_i\}$, and the matrix of edge counts between groups $\boldsymbol{e} = \{e_{rs}\}$, where e_{rs} is the number of edges between groups r and s (for convenience of notation, e_{rr} is *twice* the number of edges inside group r). Given these parameters, networks are generated like in the configuration model [12, 13]: To each vertex i is attributed k_i half-edges (or "stubs"), which are paired randomly to each other — allowing for multiple pairings between the same pair of nodes as well as self-loops — respecting the constraint that between groups r and s there are exactly e_{rs} pairings. Assuming momentarily that the half-edges are distinguishable, the number of possible pairings that satisfy this constraint is given by

$$\Omega(\boldsymbol{e}) = \frac{\prod_r e_r!}{\prod_{r<s} e_{rs}! \prod_r e_{rr}!!}, \tag{1}$$

where $e_r = \sum_s e_{rs}$ and $(2m)!! = 2^m m!$. However, many different pairings correspond to the same graph. Given an adjacency matrix $\boldsymbol{A}$, the number of different half-edge pairings to which it corresponds is analogously given by

$$\Xi(\boldsymbol{A}) = \frac{\prod_i k_i!}{\prod_{i<j} A_{ij}! \prod_i A_{ii}!!}. \tag{2}$$

Hence, the likelihood of observing a particular network given the model parameters is simply the ratio between these two numbers,

$$P(\boldsymbol{A}|\boldsymbol{k}, \boldsymbol{e}, \boldsymbol{b}) = \frac{\Xi(\boldsymbol{A})}{\Omega(\boldsymbol{e})}. \tag{3}$$

(Naturally, the above likelihood only holds if the network $\boldsymbol{A}$ matches exactly the hard constraints imposed by the parameters, i.e. $e_{rs} = \sum_{ij} A_{ij}\delta_{b_i r}\delta_{b_j s}$ and $k_i = \sum_j A_{ij}$, otherwise the likelihood is zero. In order to leave the expressions uncluttered, we will always implicitly assume that the hard constraints must hold for the likelihoods to be nonzero.)

The model above generates graphs with multiple edges between nodes, which may not be strictly appropriate for many types of networks where this cannot occur. However — as is true with the traditional configuration model — the likelihood of multiple edges will decrease with $1/N$ for sparse networks with $E \propto N$ edges, and hence their occurrence can be neglected as N becomes large.

III. NONPARAMETRIC BAYESIAN INFERENCE

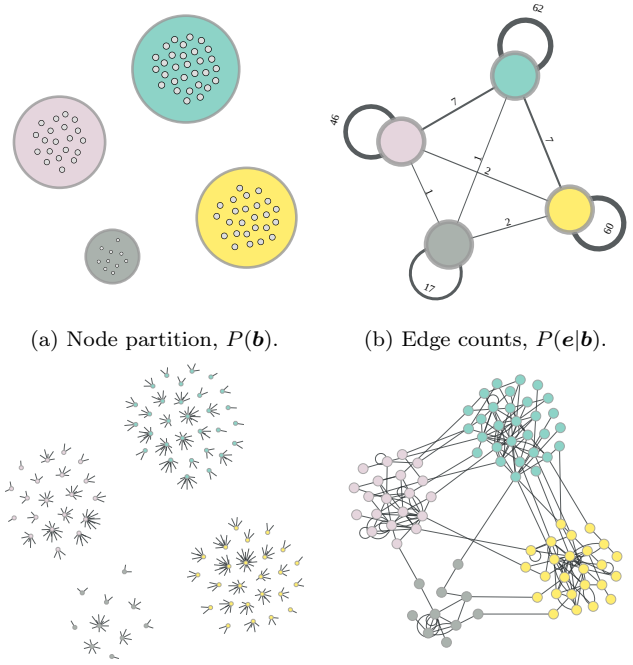
Although one could find the best divisions of the network by maximizing, or sampling from Eq. 3 directly, this requires the number of groups B to be known in advance, i.e. it is a *parametric* inference procedure that requires certain properties of the model to be determined *a priori*. Instead, here we wish to formulate a *nonparametric* framework, where the number of groups as well as any other model parameter is determined from the data itself. In order to do this, we need to write the full joint distribution for the data and the parameters,

$$P(\boldsymbol{A}, \boldsymbol{k}, \boldsymbol{e}, \boldsymbol{b}) = P(\boldsymbol{A}|\boldsymbol{k}, \boldsymbol{e}, \boldsymbol{b})P(\boldsymbol{k}|\boldsymbol{e}, \boldsymbol{b})P(\boldsymbol{e}|\boldsymbol{b})P(\boldsymbol{b}), \tag{4}$$

where $P(\boldsymbol{k}|\boldsymbol{e}, \boldsymbol{b})$, $P(\boldsymbol{e}|\boldsymbol{b})$, and $P(\boldsymbol{b})$ are prior probabilities. The above defines a complete generative model for the data and parameters, as illustrated in Fig. 1.

Based on this, we can obtain the *posterior* distribution of network partitions,

$$P(\boldsymbol{b}|\boldsymbol{A}) = \frac{P(\boldsymbol{A}, \boldsymbol{b})}{P(\boldsymbol{A})}, \tag{5}$$



(a) Node partition, $P(\boldsymbol{b})$. (b) Edge counts, $P(\boldsymbol{e}|\boldsymbol{b})$.

(c) Degrees, $P(\boldsymbol{k}|\boldsymbol{e},\boldsymbol{b})$. (d) Network, $P(\boldsymbol{A}|\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})$.

Figure 1. Illustration of the complete nonparametric generative process for the DC-SBM considered in this work. First the partition of the nodes is sampled (a), followed by the edge counts between groups (b), the degrees of the nodes (c) and finally the network itself (d).

where the normalization constant

$$P(\boldsymbol{A}) = \sum_{\boldsymbol{b}} P(\boldsymbol{A},\boldsymbol{b}) \tag{6}$$

is called the *model evidence*, and $P(\boldsymbol{A},\boldsymbol{b})$ is the marginal distribution corresponding to the joint likelihood summed over the remaining parameters,

$$P(\boldsymbol{A},\boldsymbol{b}) = \sum_{\boldsymbol{k},\boldsymbol{e}} P(\boldsymbol{A},\boldsymbol{k},\boldsymbol{e},\boldsymbol{b}) \tag{7}$$

$$= P(\boldsymbol{A},\hat{\boldsymbol{e}},\hat{\boldsymbol{k}},\boldsymbol{b}), \tag{8}$$

where $\hat{\boldsymbol{e}}$ and $\hat{\boldsymbol{k}}$ above are the only parameter choices that fulfill the model constraints compatible with the data $\boldsymbol{A}$ and the partition $\boldsymbol{b}$. Hence, here we already observe a useful property of the microcanonical formulation: Because of the hard constraints, there is no difference between the joint and marginal likelihoods. This means that we encounter no additional computational difficulty after we have determined our priors probabilities. This is in general different from "canonical" model formulations with continuous parameters, where the marginal likelihood needs to be obtained via integration, which sometimes cannot be done exactly, even if the choice of prior happens to be well motivated. In the particular case of the SBM, there are in fact typical canonical formulations where the marginal likelihood can be computed exactly [10, 14–17], but this has been done only for simple non-informative or conjugate priors, which leads to serious problems for large networks, as we discuss further in Sec. V. Here, instead, we can focus on priors that are chosen according to more fundamental principles, without having to worry about the computation of the marginal likelihood, provided the priors themselves can be computed. As we will show below, this will allow deeper Bayesian hierarchies to be developed, which make fewer assumptions about the data generating process, and lifts important practical limitations present in shallower approaches.

A. Sampling vs. optimization and the minimum description length principle (MDL).

The Bayesian formulation outlined above has an alternative — but entirely equivalent — information-theoretic interpretation. We can re-write the joint likelihood of Eq. 4 as

$$P(\boldsymbol{A},\boldsymbol{k},\boldsymbol{e},\boldsymbol{b}) = 2^{-\Sigma} \tag{9}$$

where

$$\Sigma = -\log_2 P(\boldsymbol{A},\boldsymbol{k},\boldsymbol{e},\boldsymbol{b}) = \mathcal{S} + \mathcal{L} \tag{10}$$

is called the description length of the data [18, 19], with

$$\mathcal{S} = -\log_2 P(\boldsymbol{A}|\boldsymbol{k},\boldsymbol{e},\boldsymbol{b}) \tag{11}$$

being the number of bits necessary to precisely describe the network, if the model parameters are known, and

$$\mathcal{L} = -\log_2 P(\boldsymbol{k},\boldsymbol{e},\boldsymbol{b}) \tag{12}$$

being the number of bits necessary to describe the model parameters. Hence, if we find the network partition that maximizes the posterior of Eq. 5, we are automatically finding the choice of parameters that *most compresses* the data, i.e. yields the shortest description length. This equivalence between Bayesian inference and MDL holds much more generally [19], but with the microcanonical formulation used here it is more directly evident.

The MDL interpretation also provides an intuitive explanation to why this nonparametric approach is robust against overfitting: If the number of groups becomes large, it will decrease $\mathcal{S}$ but increase $\mathcal{L}$, with the latter functioning as a "penalty" that disfavors overly complex models. For the same reason, the description length can also be used as an application-independent criterion to select between models of different classes, i.e. with a different internal structure and set of parameters. This type of comparison amounts to a formal implementation of Occam's razor, where the simplest model that can explain the data according to its statistical significance should be selected.

This equivalence means that other Bayesian approaches such as Refs. [14–17, 20–22], and those based

on MDL, e.g. Refs. [8, 10, 23], correspond in fact to the same underlying criterion. The main differences between those lie only in the actual models used, the choice of priors, as well as more practical aspects such as algorithmic complexity and approximations used.

However, it is important to emphasize that using either the Bayesian or the MDL interpretation, we need to be open to the possibility that different models — or different parametrizations of the same model — may yield the same or very similar values for the description length or posterior likelihood. In such situations, one should accept these alternative explanations for the data on equal footing. The Bayesian interpretation offers a more natural approach in these circumstances, where instead of attempting to find the maximum of the posterior distribution, we consider all possibilities, weighted according to their posterior likelihood. This can be achieved by *sampling* from the posterior distribution using Monte Carlo techniques, as explained in Sec. VI.

When deciding which route to take — to maximize or sample from the posterior — we need to acknowledge that therein lies the typical trade-off between bias and variance: When maximizing the posterior, we make a very specific statement about the data-generating process, but which can include errors from many sources, such as lack of sufficient statistics, degeneracy in the parameters or model misspecification. On the other hand, when sampling from the posterior, we obtain results which tend to be *on average* less susceptible to those errors, but which to the same degree are also more uncertain. Thus, we lose the ability to make more specific assertions. Due to its nature, the latter approach tends to incorporate more noise, and so the individual samples run the risk of overfitting the data. Conversely, the maximization approach tends to yield more conservative results, and thus runs the risk of *underfitting* the data, by omitting meaningful features. Although in the ideal scenario where the model is well specified and the data is plentiful both approaches must yield the same result, in more realistic settings one source of error can only be reduced at the expense of increasing the other. Hence, the final decision must involve the ultimate objective of the inference task. In general, we should expect sampling to be more suitable when the goal is to generalize from the observed data and make predictions about new measurements, wheres maximization tends to produce more accurate representations of the observed data.

In Secs. VIII and VII we compare results obtained via strict MDL (i.e. maximization) and the Bayesian (i.e. sampling) approaches on empirical data. In the following, we proceed with defining the prior likelihoods for the model parameters. When discussing various possibilities, we will make use of the MDL interpretation to decide which alternative yields the shortest description for data that is more likely to be encountered.

B. Prior for the node partition

We begin with the prior for the partitions. Here we outline two general approaches that will also be used for the remaining parameters. Firstly, the simplest choice we could make is to be completely agnostic about the partitions, and choose among all of them with equal probability,

$$P(\boldsymbol{b}|B) = B^{-N}. \tag{13}$$

However, this is not a good choice. The reason for this is that it inherently assumes that the group sizes will be approximately the same, since this is a typical property of completely random partitions. Not only is this unrealistic, but from a MDL perspective, whenever this is not the case, we would miss an opportunity to further compress the data. Therefore, we are better off instead replacing this by a parametric distribution, that is conditioned on the group sizes $\boldsymbol{n} = \{n_r\}$, where n_r is the number of nodes in group r,

$$P(\boldsymbol{b}|\boldsymbol{n}) = \frac{\prod_r n_r!}{N!}, \tag{14}$$

which is a maximum entropy distribution (all allowed configurations are equally likely), constrained on the fixed group sizes. In order to remain nonparametric, we need a noninformative *hyperprior* on the node counts,

$$P(\boldsymbol{n}|B) = \left(\!\!\binom{B}{N}\!\!\right)^{-1}, \tag{15}$$

where $\left(\!\binom{n}{m}\!\right) = \binom{n+m-1}{m}$ counts the number of m-combinations from a set of size n, or equivalently, the number of possible histograms with n bins with counts that sum to m. One may argue, however, that the same principle should be applied again, with the noninformative hyperprior above replaced by a parametric distribution, with parameters sampled from a hyper-hyperprior, and so on, indefinitely. However, proceeding like this yields increasingly diminishing returns, and as we now show, there are good reasons to stop at this point. If we take the logarithm of the joint likelihood $P(\boldsymbol{b}, \boldsymbol{n})$ and assume that the groups are sufficiently large so that Stirling's factorial approximation can be used, as well as $B \ll N$, we obtain

$$\ln P(\boldsymbol{b}, \boldsymbol{n}|B) \approx -NH(\boldsymbol{n}) - B \ln N \tag{16}$$

where $H(\boldsymbol{n}) = -\sum_r (n_r/N) \ln(n_r/N)$ is the entropy of the group size distribution. The first term in the equation above represents an optimal limit, i.e. for sufficient data the negative log-likelihood (the description length) approaches the entropy of the generating distribution. Hence, if we were to replace the noninformative hyperprior of Eq. 15 with an even deeper Bayesian hierarchy, we would gain at most a fairly marginal improvement proportional to $\ln N$, which is unlikely to significantly alter the inference outcome.

The joint likelihood $P(\boldsymbol{b},\boldsymbol{n}|B)$ above has been used in Refs. [16, 17, 20, 24], but in some of these works it was equivalently derived as the marginal distribution of the canonical model,

$$P(\boldsymbol{b}|B) = \int P(\boldsymbol{b}|\boldsymbol{p})P(\boldsymbol{p}|B)\,\mathrm{d}\boldsymbol{p} \tag{17}$$

with

$$P(\boldsymbol{b}|\boldsymbol{p}) = \prod_i p_{b_i} = \prod_r p_r^{n_r} \tag{18}$$

where p_r is the probability of a node belonging to group r, and

$$P(\boldsymbol{p}|B) = (B-1)!\,\delta\left(1-\textstyle\sum_r p_r\right) \tag{19}$$

is a uniform prior. Computing Eq. 17 yields an expression identical to $P(\boldsymbol{b}|B) = P(\boldsymbol{b},\boldsymbol{n}|B) = P(\boldsymbol{b}|\boldsymbol{n})P(\boldsymbol{n}|B)$ using Eqs. 14 and 15 above. However, there is an apparently small detail that needs to be addressed. Namely, the maximum entropy model of Eq. 15 also generates groups with size zero. This means that if we use it, we need to consider in our posterior distributions partitions of the network that contain empty groups, which would force us to treat the number of groups as a free variable that is not necessarily equal to the number of observed (nonempty) groups, as done in Ref. [17] [25]. However, empty groups possess no real value when interpreting the network structure; we could simply discard them and consider instead a fully equivalent model with fewer groups. Hence, in order to avoid dealing with such empty groups, we exclude them from our prior distribution, by using instead

$$P(\boldsymbol{n}|B) = \binom{N-1}{B-1}^{-1} \tag{20}$$

which is a uniform distribution over all histograms with B bins and counts that sum to N, where no bin is allowed to be empty. With this simple modification, the number of groups becomes a hard constraint as well, and is always tied to the partition, thus obviating the need to treat it as a free variable. We note that while this modification is easy in the microcanonical model, it is not as straightforward in the canonical model of Eq. 17, since for every value of $p_r < 1$, the probability that group r will end up empty is strictly nonzero.

Lastly, we need a prior for the number of non-empty groups itself, which we can choose as $P(B) = 1/N$, for $B \in [1, N]$. (We could argue that, since this amounts to a trivial multiplicative constant to the overall likelihood, we could omit it completely. However, as it will be seen further below, this term will not be a constant once we consider hierarchical partitions.) With this, we have a nonparametric prior for the partition that reads

$$\begin{aligned} P(\boldsymbol{b}) &= P(\boldsymbol{b}|\boldsymbol{n})P(\boldsymbol{n}|B)P(B) \\ &= \frac{\prod_r n_r!}{N!}\binom{N-1}{B-1}^{-1}\frac{1}{N}. \end{aligned} \tag{21}$$

Since we are forbidding empty groups *a priori*, from this point onward the value of B will refer strictly to the number of nonempty groups.

C. Prior for the degrees

1. Non-degree-corrected model (NDC-SBM)

We can recover a non-degree-corrected version of the microcanonical SBM as a special case of the model above, by assuming that the half-edges are randomly distributed among nodes of the same group, which yields a particular likelihood for the degree sequence.

If at first we assume that all $e_r = \sum_s e_{rs}$ half-edges incident on group r are distinguishable, they can be distributed among n_r nodes in $\Omega_r = n_r^{e_r}$ different ways. A particular degree sequence inside group r corresponds to exactly $\Xi_r(\boldsymbol{k}) = e_r!/\prod_{i\in r} k_i!$ such combinations, where the numerator accounts for the number of permutations of half-edges, while the denominator discounts the fraction of such permutations involving half-edges that are incident on the same node, and hence amount to the same half-edge partition. The likelihood of a particular degree sequence inside group r is given by the ratio $\Xi_r(\boldsymbol{k})/\Omega_r$, and thus the overall degree sequence likelihood becomes

$$P(\boldsymbol{k}|\boldsymbol{e},\boldsymbol{b}) = \prod_r \frac{e_r!}{n_r^{e_r}\prod_{i\in r} k_i!}, \tag{22}$$

which multiplied with Eq. 3 yields the model likelihood

$$P(\boldsymbol{A}|\boldsymbol{e},\boldsymbol{b}) = \frac{\prod_{r<s} e_{rs}!\prod_r e_{rr}!!}{\prod_r n_r^{e_r}} \times \frac{1}{\prod_{i<j} A_{ij}!\prod_i A_{ii}!!}, \tag{23}$$

which no longer depends explicitly on the degree sequence.

Like its canonical counterpart [11], the NDC-SBM will generate networks where nodes that belong to the same group will have similar degrees, with a degree distribution inside each group approaching asymptotically a Poisson. This means that standard deviation of the degrees inside group r will be $\sigma_k = \sqrt{\langle k\rangle_r}$, with $\langle k\rangle_r = e_r/n_r$ being the average degree. As argued in Ref. [11], this is an unrealistic assumption for many empirical networks, most of which possess very heterogeneous degree distributions. As a result, attempts to infer the SBM on such networks can amount largely to a division of the nodes into degree classes. It is therefore useful to postulate prior likelihoods that can account for arbitrary degree sequences, as we do in the following.

2. Arbitrary degree sequences

Similarly to the partition of the nodes, the simplest choice we can make is to sample the degrees inside each

group from a uniform distribution,

$$P(\boldsymbol{k}|\boldsymbol{e},\boldsymbol{b}) = \prod_r \left(\!\!\binom{n_r}{e_r}\!\!\right)^{-1} \tag{24}$$

where $\left(\!\binom{n_r}{e_r}\!\right)$ counts the number of possible degree sequences on n_r nodes, constrained such that their total sum equals e_r. But again, such a uniform assumption is not the best choice: If we sample from this prior, we still obtain degree sequences where most nodes have very similar degrees. Indeed, if the number of nodes is sufficiently large, it can be shown that the expected degree distribution inside each group with the above prior will approach an exponential $p_k = p(1-p)^k$, with an average $\langle k\rangle = (1-p)/p$ (see Appendix A). The expected standard deviation is therefore $\sigma_k = \sqrt{1-p}/p = O(\langle k\rangle)$, which, although larger than what is obtained with the NDC-SBM, is still significantly smaller that expected for many empirical networks [26].

In view of this, and following the same logic employed for the node partition, a better prior for $\boldsymbol{k}$ should be conditioned on an arbitrary degree distribution $\boldsymbol{\eta} = \{\eta_k^r\}$, with η_k^r being the number of nodes with degree k that belong to group r,

$$P(\boldsymbol{k}|\boldsymbol{e},\boldsymbol{b}) = P(\boldsymbol{k}|\boldsymbol{\eta})P(\boldsymbol{\eta}|\boldsymbol{e},\boldsymbol{b}) \tag{25}$$

and where

$$P(\boldsymbol{k}|\boldsymbol{\eta}) = \prod_r \frac{\prod_k \eta_k^r!}{n_r!} \tag{26}$$

is a uniform likelihood of for degree sequences constrained by the overall degree counts, and

$$P(\boldsymbol{\eta}|\boldsymbol{e},\boldsymbol{b}) = \prod_r q(e_r,n_r)^{-1} \tag{27}$$

is the likelihood of the overall degree counts. The quantity $q(m,n)$ is the number of different degree counts with the sum of degrees being exactly m and that have at most n non-zero counts. This is also known as the number of *restricted partitions* of the integer m into at most n parts [27]. The function $q(m,n)$ can be computed exactly via the recurrence

$$q(m,n) = q(m,n-1) + q(m-n,n), \tag{28}$$

and the boundary conditions $q(m,1) = 1$ for $m > 0$, and $q(m,n) = 0$ for $m \leq 0$ or $n \leq 0$. With this, the full table of values for $m \leq M$ and $n \leq m$ can be computed in time $O(M^2)$. Hence, if the number of edges and nodes is not too large, we can pre-compute these values as a setup to the inference procedure. However, this can still become computationally expensive for very large systems. Unfortunately, no closed-form expression for $q(m,n)$ is known which would allow us to compute it in constant time. Fortunately, however, accurate asymptotic expressions are known, which permit efficient computation for large arguments. Namely, for large values of m the number of partitions approaches asymptotically the following value [28–30]

$$q(m,n) \approx \frac{f(u)}{m}\exp(\sqrt{m}g(u)), \tag{29}$$

where $u = n/\sqrt{m}$ and the functions $f(u)$ and $g(u)$ are given by

$$f(u) = \frac{v(u)}{2^{3/2}\pi u}\left[1-(1+u^2/2)e^{-v(u)}\right]^{-1/2}, \tag{30}$$

$$g(u) = \frac{2v(u)}{u} - u\ln(1-e^{-v(u)}), \tag{31}$$

and $v(u)$ is given implicitly by solving

$$v = u\sqrt{-v^2/2 - \mathrm{Li}_2(1-e^v)}, \tag{32}$$

where $\mathrm{Li}_2(z) = -\int_0^z [\ln(1-t)/t]\mathrm{d}t$ is the dilogarithm function. (Eq. 32 can be easily solved numerically via Newton's method, or simply via repeated iteration, which converges within machine precision usually after only very few steps). This approximation holds for values of $n \geq m^{1/6}$. For smaller values $n \ll m^{1/3}$ we have instead [31]

$$q(m,n) \approx \frac{\binom{m-1}{n-1}}{m!}. \tag{33}$$

With Eqs. 29 to 33 we have an approximation for $q(m,n)$ for the entire range of parameters m and n that is remarkably accurate, as shown in Fig. 2: For arguments of the order 10^3, the largest log ratio between the approximate and exact values is only around 0.1, which has a negligible effect on the outcome of hypothesis testing, and is below the accuracy usually required for MCMC sampling. In our implementation, we pre-compute $q(m,n)$ using the exact Eq. 28 for $m < 10^4$, and resort to Eqs. 29 to 33 only for larger arguments, thus guaranteeing a computation of $q(m,n)$ in time $O(1)$, and hence incurring a negligible impact in the overall algorithmic complexity of the inference procedure.

As seen in Fig. 3, the expected degree distribution sampled from Eq. 27 is typically significantly broader than the exponential distribution obtained with Eq. 24. As show in Appendix A, this will approach a Bose-Einstein distribution, with a variance $\sigma_k^2 \propto \sqrt{N}$ that will diverge for a large system size. In particular, the distribution will asymptotically approach a scale-free form $p_k \sim 1/k$ for $k \ll \sqrt{E}$, followed by an exponential decay for larger arguments.

Although this prior assumption clearly favors broader degree distributions, it could be argued that it still does not properly capture the structure of real networks, most of which also do not possess a Bose-Einstein degree distribution. Indeed, it may seem that by changing between the priors considered above, we have simply switched between Poisson, geometric and Bose-Einstein distributions, which are just three of an infinite range of possibilities. However, in reality, the conditioned prior of Eq. 25

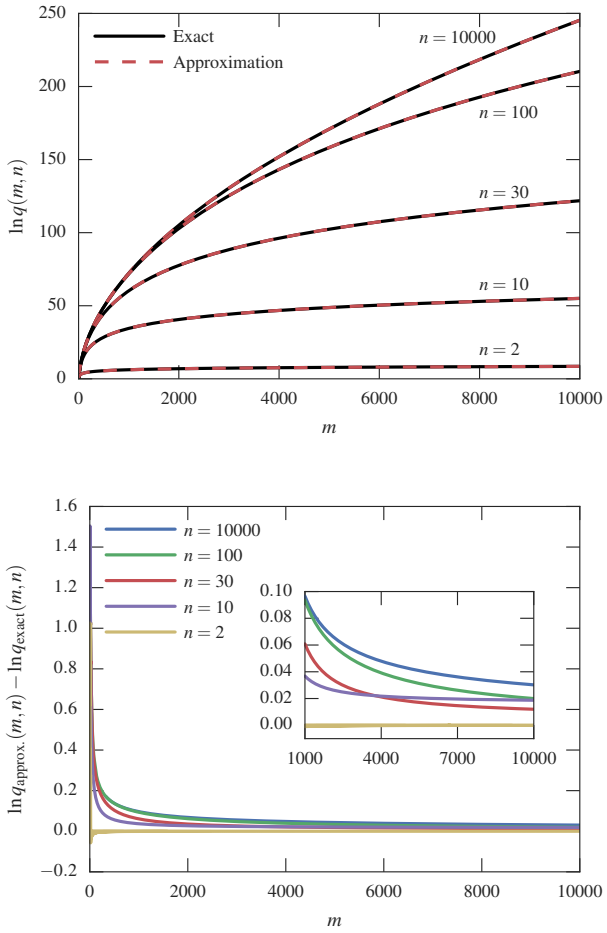


Figure 2. Comparisons between the exact and approximated values of the number of restricted partitions $q(m,n)$, using Eqs. 28 and 29 to 33. The top panel shows both values computed for different values of m and n, and the bottom panel shows the absolute difference of their logarithms, with the inset displaying a zoom into the large m region.

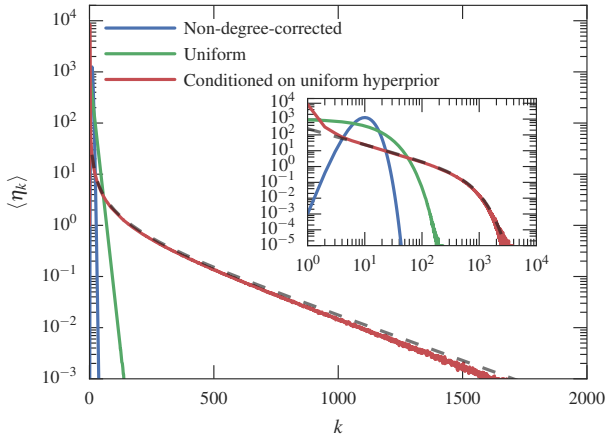


Figure 3. Expected degree distributions for the three different priors considered in the text for the degree sequence inside each group — the NDC-SBM, the uniform prior of Eq. 24 and the prior of Eq. 27 conditioned on a degree distribution sampled randomly — for $N = 10^4$ nodes and average degree $\langle k \rangle = 10$. In all cases, the distributions were sampled from their respective microcanonical distributions using rejection sampling. The dashed line shows the Bose-Einstein distribution of Eq. A12.

will not concentrate as strongly on the expected distribution as the other two, and thus will not significantly penalize distributions that deviate from it, even if the deviation is very large, as will now be shown.

In order to assess the improvement brought on by the conditioned prior, it is instructive to obtain the asymptotic behavior of $q(m,n)$ in the limit of "sufficient data" with $m \gg 1$ and $n \gg 1$, which is given by [31]

$$q(m,n) \approx p(m) \exp\left(-\frac{\sqrt{6m}}{\pi} e^{-\pi n/\sqrt{6m}}\right), \tag{34}$$

as long as $n \gg \sqrt{m}$ and where $p(m) = q(m,m)$ is the number of unconstrained partitions of m, which itself is given exactly by the recursion

$$p(m) = \sum_{k>0} (-1)^{k-1} p\left(m - k(3k-1)/2\right) \tag{35}$$

and for large values of m by the Hardy-Ramanujan formula [32, 33]

$$p(m) \approx \frac{1}{4\sqrt{3}m} \exp(\pi\sqrt{2m/3}). \tag{36}$$

With these results, we see immediately that for "sparse" groups with $e_r \propto n_r$ and $n_r \gg 1$ we have $\ln q(e_r, n_r) \sim O(\sqrt{n_r})$, and hence

$$\ln P(\boldsymbol{k}|\boldsymbol{e},\boldsymbol{b}) \approx -\sum_r n_r H(\boldsymbol{\eta}_r) + O(\sqrt{n_r}), \tag{37}$$

where $H(\boldsymbol{\eta}_r) = -\sum_k (\eta_k^r/n_r) \ln(\eta_k^r/n_r)$ is the entropy of the empirical degree distribution in group r. Therefore, for sufficiently many nodes in each group, the hyperprior of Eq. 27 will "wash out" and the likelihood of Eq. 25 will approach that of the the *actual* degree sequence, whatever its form may be, even if it deviates from the typical form of Fig. 3. This is not the case of the uniform prior of Eq. 24, which is not able to "learn" the underlying distribution in the same manner. Eq. 37 also means that an exact prior knowledge of the *true* degree distribution in each group would improve the log-likelihood (and the description length) only by a factor $O(\sqrt{n_r})$, which will be dwarfed asymptotically by the remaining terms that scale linearly as $O(n_r)$. Therefore, any further improvement in the choice of prior for the degree sequence is confined to a relatively narrow range, similarly to what happens with the prior for the partition of the nodes into groups.

D. Prior for the edge counts and nested SBM hierarchies

The remaining piece is the prior for the edge counts between groups, $\boldsymbol{e}$. We can start again with a uniform prior

$$P(\boldsymbol{e}) = \left(\!\!\binom{\left(\!\binom{B}{2}\!\right)}{E}\!\!\right)^{-1}, \tag{38}$$

where $\left(\!\binom{\left(\!\binom{B}{2}\!\right)}{E}\!\right)$ counts the number of symmetric e_{rs} matrices with a constrained sum $\sum_{rs} e_{rs} = 2E$.

Perhaps unsurprisingly at this point, this also not a good choice. This time, however, the negative effects are somewhat more dramatic than the previous choices of uniform priors. Namely, this assumption will limit our capacity to detect small groups in very large networks: It introduces a "resolution limit", where the largest number of groups that can be inferred scales as $B_{\max} \sim \sqrt{N}$ [8], similar to what is observed with the modularity maximization heuristic [4]. We revisit this issue in more detail in Sec. V.

As was shown in Ref. [10], this problem can be solved again by deepening the Bayesian hierarchy. It is useful now to notice that the matrix $\boldsymbol{e}$ can be interpreted as the adjacency matrix of a multigraph with B nodes and E edges. Hence, an appropriate choice seems to be to use the SBM again to generate it, where each group r belongs to one of another set of groups, and so on recursively, a L number of times,

$$P(\{\boldsymbol{e}_l\}|\{\boldsymbol{b}_l\}) = \prod_{l=1}^{L} P(\boldsymbol{e}_l|\boldsymbol{e}_{l+1}, \boldsymbol{b}_l), \tag{39}$$

where $\boldsymbol{b}_l$ is the partition of the groups in level l, $\boldsymbol{e}_l$ is the (weighted) adjacency matrix at level l, and we enforce always that $B_L = 1$. Note that we can no longer use Eq. 3 to model the upper levels via a configuration process, we need instead a maximum entropy NDC-SBM for multigraphs [34], i.e.

$$P(\boldsymbol{e}_l|\boldsymbol{e}_{l+1}, \boldsymbol{b}_l) = \prod_{r<s} \left(\!\!\binom{n^l_r n^l_s}{e^{l+1}_{rs}}\!\!\right)^{-1} \prod_{r} \left(\!\!\binom{n^l_r(n^l_r+1)/2}{e^{l+1}_{rr}/2}\!\!\right)^{-1}. \tag{40}$$

Note that if we make $L = 1$, we recover the uniform prior of Eq. 38. To complete the model, we need also the prior for the partitions in all levels,

$$P(\{\boldsymbol{b}_l\}) = \prod_{l=1}^{L} P(\boldsymbol{b}_l), \tag{41}$$

where for each level we use again Eq. 21, but replacing $B \to B_l$ and $N \to B_{l-1}$, with the boundary condition $B_1 = N$.

The depth L of the hierarchy itself is something that we want to infer from the data as well. One approach, for instance, is to put a noninformative prior on it $P(L) = 1/L_{\max}$, with some maximum possible value $L_{\max}$ that is sufficiently large, e.g. $L_{\max} = N$. But since this contributes to nothing but an overall multiplicative constant in the likelihood, it can be omitted altogether.

E. Model summary

Putting together the model likelihood with all the priors, we have a joint likelihood for the hierarchical microcanonical DC-SBM that reads

$$\begin{aligned} P(\boldsymbol{A}, \boldsymbol{k}, \{\boldsymbol{e}_l\}, \{\boldsymbol{b}_l\}) &= P(\boldsymbol{A}|\boldsymbol{k}, \boldsymbol{e}, \boldsymbol{b}_1) \times P(\boldsymbol{k}|\boldsymbol{e}_1, \boldsymbol{b}_1) \times P(\{\boldsymbol{e}\}_l) \times P(\{\boldsymbol{b}\}_l) && (42)\\ &= \frac{\prod_i k_i! \prod_{r<s} e_{rs}! \prod_r e_{rr}!!}{\prod_r e_r! \prod_{i<j} A_{ij}! \prod_i A_{ii}!!} \times \prod_r \frac{\prod_k \eta^r_k!}{n_r!} q(e_r, n_r)^{-1} \times && (43)\\ &\quad \prod_{l=1}^{L} \prod_{r<s} \left(\!\!\binom{n^l_r n^l_s}{e^{l+1}_{rs}}\!\!\right)^{-1} \prod_r \left(\!\!\binom{n^l_r(n^l_r+1)/2}{e^{l+1}_{rr}/2}\!\!\right)^{-1} \times \frac{\prod_r n^l_r!}{B_{l-1}!} \binom{B_{l-1}-1}{B_l-1}^{-1} \frac{1}{B_{l-1}}. && (44) \end{aligned}$$

It is important to emphasize that this likelihood has the following useful property: When considering the difference in the log-likelihood after moving a single node i from a group to another, it is necessary only to consider a number of terms that is proportional to the number of groups that are involved in the change, i.e. those of the node that is being moved and its neighbors. Therefore, in the worse case, we need to update $O(k_i)$ terms, a number which is *independent* of the total number of groups in the bottom of the hierarchy, B_1. This contrasts with other formulations that require the computation of a number of terms that is linearly proportional to the total number of groups (e.g. [14–17]), or even quadratic (e.g. [35]). This property will permit the inference on large networks, for

which the appropriate number of groups is likely to be large as well, as we describe in Sec. VI.

In addition to this model, the NDC-SBM and the alternative version of the DC-SBM with uniform priors on the degrees can be obtained simply by replacing the prior $P(\boldsymbol{k}|\boldsymbol{e},\boldsymbol{b}_0)$ in Eq. 42 with the appropriate one. This does not change the efficiency of the likelihood computation described above. Furthermore, as mentioned previously, the non-hierarchical version of each model can be recovered by simply enforcing a hierarchy with just one level, i.e. $L = 1$.

IV. ENSEMBLE EQUIVALENCE

The microcanonical model above differs from the most common "canonical" formulation of the SBM, where the modular network structure is imposed via "soft" constraints, that are obeyed only on average. For example, the original canonical Poisson formulation of the DC-SBM [11] is

$$P(\boldsymbol{A}|\boldsymbol{\lambda},\boldsymbol{\theta}) = \prod_{i<j}\frac{(\theta_i\theta_j\lambda_{b_ib_j})^{A_{ij}}e^{-\theta_i\theta_j\lambda_{b_ib_j}}}{A_{ij}!}\prod_i\frac{(\theta_i^2\lambda_{b_ib_i}/2)^{A_{ij}/2}e^{-\theta_i^2\lambda_{b_ib_i}/2}}{(A_{ii}/2)!} \tag{45}$$

$$= \prod_{r<s}\lambda_{rs}^{e_{rs}}e^{-\lambda_{rs}\hat{\theta}_r\hat{\theta}_s}\prod_r\lambda_{rr}^{e_{rr}/2}e^{-\lambda_{rr}\hat{\theta}_r^2/2}\times\frac{\prod_i\theta_i^{k_i}}{\prod_{i<j}A_{ij}!\prod_i A_{ii}/2!}. \tag{46}$$

where θ_i determines the propensity of node i to receive edges, whereas λ_{rs} controls the distribution of edges between groups, and with

$$\hat{\theta}_r = \sum_i \theta_i\delta_{b_i,r}. \tag{47}$$

In this model, the degrees of the nodes and the number of edges between groups are fixed only in expectation, but otherwise can fluctuate between samples. If one applies Stirling's factorial approximation $\ln m! \approx m\ln m - m$ to the terms of Eqs. 1 and 2 that depend on e_{rs} and k_i, it is easily seen that the microcanonical likelihood of Eq. 3 approaches Eq. 45, which means both models generate the same networks with the same probability asymptotically, if the parameters are chosen in a compatible manner, e.g. $\theta_i = k_i/e_{b_i}$ and $\lambda_{rs} = e_{rs}$. However, this only holds if the edge counts between groups *as well as* the degrees of the nodes become sufficiently large. For smaller or sparser networks, on the other hand, the differences can be important, and it is well understood that the microcanonical and canonical ensembles are not equivalent in these cases [34, 36–38]. However, an exact equivalence between these ensembles can in fact be obtained in a Bayesian setting, via the computation of the marginal likelihood that involves integrating over the canonical parameters, $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$, weighted with a prior likelihood, as will now be shown.

Before we can proceed with the computation of the marginal likelihood, we must notice that the model parameters are determined only up to an arbitrary multiplicative constant, since the likelihood of Eq. 45 depends only on their products $\theta_i\theta_j\lambda_{b_ib_j}$. Although their absolute values are in principle arbitrary, the exact parametrization we choose will affect the choice of priors we can make, an ultimately the marginal likelihood. Here we will contrast two possible choices. We begin with the assumption made in Refs. [17, 22]

$$\hat{\theta}_r = n_r. \tag{48}$$

If we make this choice, the value of λ_{rs} corresponds to the average probability of two nodes in groups r and s being connected. We can then choose a noninformative prior for $\boldsymbol{\lambda}$, conditioned only on the expected density of the network, $p = 2E/N^2$,

$$P(\lambda_{rs}) = e^{-\lambda_{rs}/p}/p. \tag{49}$$

For $\boldsymbol{\theta}$, we use also a noninformative distribution,

$$P(\boldsymbol{\theta}|\boldsymbol{b}) = \prod_r\frac{(n_r-1)!}{n_r^{n_r}}\,\delta(\hat{\theta}_r - n_r), \tag{50}$$

subject only to the scaled simplex constraint of Eq. 48. As computed in Ref. [17], the marginal likelihood is therefore,

$$P_1(\boldsymbol{A}|\boldsymbol{b}) = \int P(\boldsymbol{A}|\boldsymbol{\lambda},\boldsymbol{\theta})P(\boldsymbol{\lambda})P(\boldsymbol{\theta}|\boldsymbol{b})\,\mathrm{d}\boldsymbol{\lambda}\,\mathrm{d}\boldsymbol{\theta} \tag{51}$$

$$= p^E \prod_{r<s} \frac{e_{rs}!}{(pn_rn_s+1)^{e_{rs}+1}} \prod_r \frac{(e_{rr}/2)!}{(pn_r^2/2+1)^{e_{rs}/2+1}} \prod_r \frac{n_r^{e_r}(n_r-1)!}{(e_r+n_r-1)!} \frac{\prod_i k_i!}{\prod_{i<j} A_{ij}! \prod_i A_{ii}/2!}. \tag{52}$$

This marginal likelihood is not equivalent to he microcanonical model presented previously, and hence corresponds to a different overall generative process. However, things are different if we assume another parametrization, namely

$$\hat{\theta}_r = 1. \tag{53}$$

In this case, the value of λ_{rs} represents the average number of edges between groups r and s (or twice that for $r = s$). Similar to the previous case, we can choose a noninformative prior for $\boldsymbol{\lambda}$, conditioned only on the expected total number of edges,

$$P(\lambda_{rs}) = \begin{cases} e^{-\lambda_{rs}/\bar{\lambda}}/\bar{\lambda} & \text{if } r \neq s, \\ e^{-\lambda_{rs}/2\bar{\lambda}}/2\bar{\lambda} & \text{if } r = s, \end{cases} \tag{54}$$

with $\bar{\lambda} = 2E/B(B+1)$. Like before, for $\boldsymbol{\theta}$ we use noninformative distribution,

$$P(\boldsymbol{\theta}|\boldsymbol{b}) = \prod_r (n_r-1)!\,\delta(\hat{\theta}_r - 1), \tag{55}$$

but subject now to the simplex constraint of Eq. 53 instead. Performing the same integral, the marginal likelihood then becomes

$$P_2(\boldsymbol{A}|\boldsymbol{b}) = \frac{\bar{\lambda}^E}{(\bar{\lambda}+1)^{E+B(B+1)/2}} \times \frac{\prod_{r<s} e_{rs}! \prod_r e_{rr}!!}{\prod_{i<j} A_{ij}! \prod_i A_{ii}!!} \prod_r \frac{(n_r-1)!}{(e_r+n_r-1)!} \prod_i k_i!, \tag{56}$$

from which we can immediately recognize the microcanonical model by re-writing the likelihood as

$$P_2(\boldsymbol{A}|\boldsymbol{b}) = P(\boldsymbol{A}|\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})P(\boldsymbol{k}|\boldsymbol{e},\boldsymbol{b})P(\boldsymbol{e}), \tag{57}$$

where $P(\boldsymbol{A}|\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})$ is the microcanonical likelihood of Eq. 3, $P(\boldsymbol{k}|\boldsymbol{e},\boldsymbol{b})$ is the noninformative degree-sequence likelihood of Eq. 24 and $P(\boldsymbol{e})$ is the likelihood of the degree counts as $B(B+1)/2$ independent exponential variables with average $\bar{\lambda}$,

$$P(\boldsymbol{e}) = \prod_{r<s} (1-\mu)^{e_{rs}}\mu \prod_r (1-\mu)^{e_{rr}/2}\mu \tag{58}$$

$$= \bar{\lambda}^E/(\bar{\lambda}+1)^{E+B(B+1)/2}, \tag{59}$$

where $\mu = 1/(\bar{\lambda}+1)$. This last prior $P(\boldsymbol{e})$ is different from the microcanonical one used in Eq. 38 simply in that here the total number of edges is allowed to fluctuate, being constrained only in expectation. Otherwise, the likelihoods of the canonical and microcanonical models are identical. This means that although both formulations involve distinct generative processes, these are not in fact distinguishable from data. This is fortunate, since it eliminates at least one arbitrary choice we have to make prior to inferring the modular structure of networks, and shows that the choice of ensemble can be largely subjective.

However, we are still left with a seemingly arbitrary choice of parametrization, having to decide between Eq. 48 (option 1) and Eq. 53 (option 2). As the results above show, these choices correspond to different assumptions about the data-generating process. In the first case, the expected number of edges between groups r and s (according to the prior for $\boldsymbol{\lambda}$) is assumed to depend on the sizes of the groups, i.e. $\langle e_{rs}\rangle = n_r n_s p$. This is the same expected value for the same partition of a completely random network with density p. In the second case, however, this value is independent of the group sizes $\langle e_{rs}\rangle = \bar{\lambda}$, and deviates from the expected fully random value whenever the groups sizes are not the same. Hence, the ensembles generated in each case are indeed different, and to decide which one should be used is a model selection problem. As will be discussed in more detail in Sec. VII, this can be performed by inspecting the marginal likelihood ratio between both models, assuming the same node partition,

$$\Lambda = \frac{P_2(\boldsymbol{A}|\boldsymbol{b})}{P_1(\boldsymbol{A}|\boldsymbol{b})} \tag{60}$$

where $P_1(\boldsymbol{A}|\boldsymbol{b})$ and $P_2(\boldsymbol{A}|\boldsymbol{b})$ correspond to Eqs. 52 and 56, respectively. If we assume $N \gg B^2$, this ratio

amounts to a simple expression

$$\ln\Lambda \approx \sum_{r\geq s}\left[\frac{e_{rs}}{pn_rn_s} - \ln\frac{(1+\delta_{rs})N^2}{B(B+1)n_rn_s} - 1\right]. \quad (61)$$

From this, and if we further assume groups of equal sizes $n_r = N/B$ as well as $B \gg 1$, we see that as the network approaches a fully random structure with $e_{rs} = pn_rn_s$, we have $\ln\Lambda \to -B\ln 2$ and hence a situation that favors option 1. However, as the data become more structured, this is more often not the case. This is better seen by considering a special case known as the planted partition model [39], composed of B equal-sized groups and edge counts given by

$$e_{rs} = 2E\left[\frac{c}{B}\delta_{rs} + \frac{(1-c)}{B(B-1)}(1-\delta_{rs})\right], \quad (62)$$

with $c \in [0,1]$ controlling the degree of assortativity. Substituting this in the above, we have

$$\ln\Lambda \approx \frac{B^2(c+1)}{2} - \frac{B(B+1)}{2}\ln\left(\frac{eB}{B+1}\right) - B\ln 2, \quad (63)$$

which is independent of the size of the network, and grows only with the number of groups and assortativity. For $B \gg 1$, we have $\ln\Lambda > 0$ if $c > (2\ln 2)/B \approx 1.4/B$. The ensemble is equivalent to a fully random network at a slightly smaller value $c = 1/B$ [but is already undetectable at $c = 1/B \pm (B-1)/(B\sqrt{\langle k\rangle})$ [40]]. Hence, as the number of groups increases, for the vast majority of parameter choices $c \in [(2\ln 2)/B, 1]$ we have that option 2 is favored with a confidence that grows as $\ln\Lambda = O(B^2)$.

Beside these arguments, there are other more important reasons to prefer option 2. If we adopt its microcanonical interpretation, we can address the issues with the noninformative priors discussed in the previous sections, and replace both $P(\boldsymbol{k}|\boldsymbol{e},\boldsymbol{b})$ and $P(\boldsymbol{e})$ by distributions conditioned on hyperparameters. Furthermore, as already mentioned, changes to the likelihood of Eq. 56 can be computed more efficiently than Eq. 52: If we move a node i to a new group, we need to update $O(B)$ terms in Eq. 52, whereas in Eq. 56 at most only $O(k_i)$ terms need to be recomputed (independent of B). This leads to a substantial improvement in the performance of inference algorithms, as discussed further in Sec. VI.

V. HOW MANY GROUPS CAN BE INFERRED?

One of the main strengths of the nonparametric approach presented here is that it can be used to determine the number of groups B, in addition to the other model parameters. One natural question that arises is whether there are intrinsic limitations associated with the inference of this parameter. As shown before in Ref. [8] with a simplified version of the model presented here, the choice of a noninformative prior for the edge counts $P(\boldsymbol{e})$ leads to a limitation where at most only $O(\sqrt{N})$ groups can be identified. Replacing this noninformative prior by a series of nested SBMs was shown in Ref. [10] to significantly alleviate this limitation, increasing the maximum number of groups to $O(N/\ln N)$. Here we revisit this issue, considering the more elaborate models presented in this work.

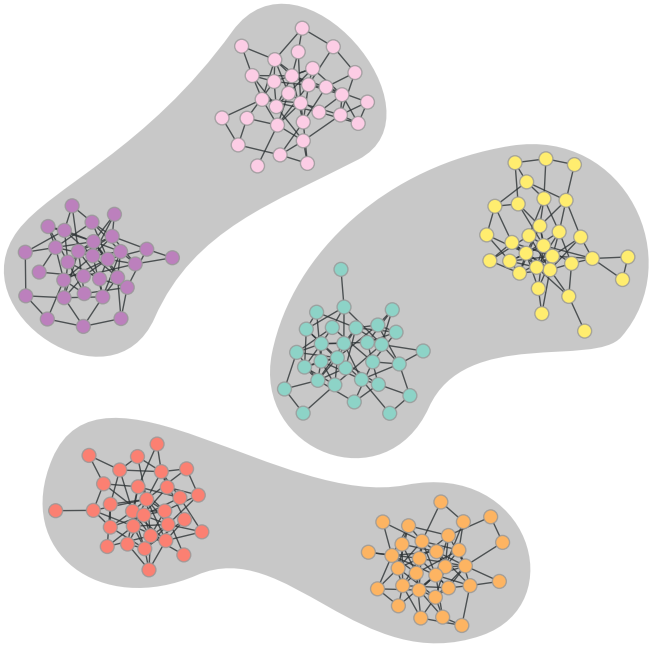


Figure 4. Planted partition of $B = 6$ equal sized groups, being fitted as a $B' = 3$ model by merging groups in pairs.

We perform our analysis on a degree-corrected planted partition model, with B groups of equal size, each containing exactly E/B edges connecting their nodes randomly, and no connections at all between nodes of different groups, i.e. $e_{rs} = 2E\delta_{rs}/B$. The likelihood of any particular network sampled from this model is

$$P(\boldsymbol{A}|\boldsymbol{k},\boldsymbol{e},\boldsymbol{b}) = \frac{(2E/B)!!^B}{(2E/B)!^B} \times \frac{\prod_i k_i!}{\prod_{i<j}A_{ij}!\prod_i A_{ii}!!}, \quad (64)$$

and with prior likelihoods

$$P(\boldsymbol{b}) = \frac{(N/B)!^B}{N!} \times \binom{N-1}{B-1}^{-1}\frac{1}{N}, \quad (65)$$

$$P(\boldsymbol{e}|\boldsymbol{b}) = \left(\!\!\binom{B(B+1)/2}{E}\!\!\right), \quad (66)$$

$$P(\boldsymbol{k}|\boldsymbol{e},\boldsymbol{b}) = \left(\!\!\binom{N/B}{2E/B}\!\!\right)^{-B}, \quad (67)$$

where we have used the noninformative priors for the edge counts and degrees.

We now consider the following scenario, illustrated in Fig. 4: We misclassify the group memberships of the nodes with a smaller number of groups B' by merging the planted groups together, such that the resulting groups

also have equal sizes. Because of the symmetric and self-similar nature of this model, the likelihood of the wrong labelling has a form identical to the equations above, but with the number of groups replaced by the alternative value, $B \to B'$.

Hence we may find the value of B that optimizes the joint likelihood $P(\boldsymbol{A}, \boldsymbol{k}, \boldsymbol{e}, \boldsymbol{b})$, and if it does not coincide with the planted value, this means that the planted value is not the most likely *a posteriori*. In the following, we assume that $N \gg 1$, $E \propto N$, $B \gg 1$, as well as $N \gg B$ (although we make no assumption between B^2 and N). Considering this scaling scenario, and if we keep only the leading terms of the likelihoods above, and also omit those that do not depend on B, we have

$$\ln P(\boldsymbol{A}, \boldsymbol{k}, \boldsymbol{e}, \boldsymbol{b}) \approx (E - N)\ln B - (E + B^2/2)h\left(\frac{E}{E + B^2/2}\right), \quad (68)$$

where $h(x) = -x\ln x - (1-x)\ln(1-x)$. If we maximize the above equation with respect to B, we obtain

$$B^* = x(\langle k\rangle)\sqrt{N}, \quad (69)$$

with $x(\langle k\rangle)$ being the solution of

$$\langle k\rangle - 2 = 2x^2 f'(1 + x^2/\langle k\rangle) \quad (70)$$

with $f(x) = xh(1/x)$, and $\langle k\rangle = 2E/N$. Hence we obtain the same result of Ref. [8] that the maximum number of groups scales as $B^* \propto \sqrt{N}$. This property is robust with respect to details of the model, and is simply a direct result of a noninformative prior used for $P(\boldsymbol{e})$, which is responsible for the dependence on B^2 in the last term of Eq. 68: A lack of prior information on the large-scale structure incurs a cost in the description length that scales roughly as $-\ln P(\boldsymbol{e}) \sim (B^2/2)\ln E$ (for $B^2 \gg E$). This means that we obtain very similar results when considering the other model variants considered in this work. In particular, using either Eq. 52 or 56 we obtain asymptotic expressions for the joint likelihood that are very similar to Eq. 68, and yield only a slightly worse scaling for the maximum number of groups, $B^* \propto \sqrt{N/\log N}$, with the $\sqrt{\log N}$ difference due to the priors of Eqs. 49 and 54, that allow the total number of edges to fluctuate. Using the uniform hyperpriors for the degree sequences also has no effect on this limitation.

On the other hand, as shown in Ref. [10], this issue is significantly improved by using the hierarchical prior for $\boldsymbol{e}$. Here we show this by considering a uniform hierarchical division where at each level the number of groups decrease by a factor σ, $B_l = B/\sigma^l$. Using Eq. 41, we have

$$P(\boldsymbol{e}) = \prod_{l=1}^{\log_\sigma B}\left(\!\!\binom{\sigma(\sigma+1)/2}{2E\sigma^l/B}\!\!\right)^{-B/\sigma^l} \times \frac{\sigma!^{B/\sigma^l}}{(B/\sigma^{l-1})!}\binom{B/\sigma^{l-1}-1}{B/\sigma^l - 1}^{-1}. \quad (71)$$

Assuming $B \gg \sigma$, and keeping only the leading terms, we have $\ln P(\boldsymbol{e}) \approx -[B\sigma(\sigma+1)\ln E]/[2(\sigma-1)]$, and hence

$$\ln P(\boldsymbol{A}, \boldsymbol{k}, \boldsymbol{e}, \boldsymbol{b}) \approx (E - N)\ln B - \frac{\sigma(\sigma+1)}{2(\sigma-1)}B\ln E, \quad (72)$$

from which we obtain the upper bound

$$B^* = \frac{(\sigma-1)(\langle k\rangle - 2)}{\sigma(\sigma+1)} \times \frac{N}{\ln N}. \quad (73)$$

Hence, this choice of priors enables the identification of a number of groups that is far larger that what is possible with the noninformative choice. This comes with no drawbacks, since this prior includes the noninformative one as a special case, and we are still protected against overfitting; becoming only less susceptible to *underfitting*.

VI. INFERENCE ALGORITHM

The inference task we have is to sample from (or maximize) the posterior distribution of the hierarchical partition,

$$P(\{\boldsymbol{b}_l\}|\boldsymbol{A}) = \frac{P(\boldsymbol{A}, \{\boldsymbol{b}_l\})}{P(\boldsymbol{A})}. \quad (74)$$

The approach we will take is based on a Markov chain Monte Carlo importance sampling for the partitions at all hierarchy levels. The algorithm will revolve around moving the membership of nodes in different hierarchical levels at random, and accepting or rejecting those moves, so that after a sufficiently long equilibration time, the hierarchical partitions are sampled according to Eq. 74. We note that this posterior can be factorized as

$$P(\{\boldsymbol{b}_l\}|\boldsymbol{A}) = \frac{\prod_l P(\boldsymbol{e}_{l-1}, \boldsymbol{b}_l|\boldsymbol{e}_l)}{P(\boldsymbol{A})} \quad (75)$$

$$= \prod_l P(\boldsymbol{b}_l|\boldsymbol{e}_{l-1}, \boldsymbol{e}_l) \quad (76)$$

with per-level posteriors

$$P(\boldsymbol{b}_l|\boldsymbol{e}_l, \boldsymbol{e}_{l+1}) = \frac{P(\boldsymbol{e}_l|\boldsymbol{e}_{l+1}, \boldsymbol{b}_l)P(\boldsymbol{b}_l)}{P(\boldsymbol{e}_l|\boldsymbol{e}_{l+1})}, \quad (77)$$

where we assume $\boldsymbol{e}_0 = A$, and $P(\boldsymbol{e}_l|\boldsymbol{e}_{l+1})$ is a normalization constant.

Therefore, a workable approach is to separately sample partitions at each level according to its individual posterior, conditioned on the remaining levels, which are kept unchanged for the time being. If we sample sufficiently often from each level in this manner (what is called "Gibbs sampling" in the statistics literature), the overall distribution will correspond to the desired full posterior of Eq. 74. Since the hierarchical levels are coupled, when moving a node at level l, we must ensure that this does not invalidate the partition at level $l+1$. Hence, we must forbid node moves between groups that

are themselves at different groups in the next level. (This constraint does not break ergodicity, since all partitions in the upper levels will be allowed to change at some point).

In more detail, we proceed as follows. At each individual level l, we perform a move proposal of node i from its current group r to a new group s, according to a probability $P(b_i^{(l)} = r \to s)$ that we will specify shortly. We compute the difference in the log-likelihood $\Delta \ln P_l$ at that level, and we accept the move according to the Metropolis-Hastings criterion [41, 42], i.e. with a probability

$$a = \min\left\{1, e^{\Delta \ln P_l} \frac{P(b_i^{(l)} = s \to r)}{P(b_i^{(l)} = r \to s)}\right\}, \tag{78}$$

where $P(b_i^{(l)} = s \to r)$ is the likelihood of the reverse move being proposed. The log-likelihood difference is computed as

$$\Delta \ln P_l = \ln \frac{P(b_i^{(l)} = s, \boldsymbol{b}_l \setminus b_i^{(l)} | \boldsymbol{e}_l, \boldsymbol{e}_{l+1})}{P(b_i^{(l)} = r, \boldsymbol{b}_l \setminus b_i^{(l)} | \boldsymbol{e}_l, \boldsymbol{e}_{l+1})}, \tag{79}$$

where $\boldsymbol{b}_l \setminus b_i^{(l)}$ means the partition of the remaining nodes excluding node i. Note that in computing Eq. 79, we do not need to determine the normalization constant in Eq. 77, and the remaining relevant terms correspond only to a subset of the full joint distribution of Eq. 44. Typically, the number of groups in the upper levels decreases exponentially, and hence the algorithmic complexity is dominated by the bottom level $l = 0$. As mentioned previously, the number of terms of the joint likelihood that are necessary to compute $\Delta \ln P_0$ is proportional only to the degree k_i of node i, and is independent of B_1, and hence can be computed quickly. Therefore, if we attempt one move for each node in the network, such a "sweep" can be completed in time $O(E)$, independent on the total number of groups.

An important element of this algorithm is the move proposal probability $P(b_i^{(l)} = r \to s)$. Any choice with nonzero probability for all values of s will preserve ergodicity, and — coupled with the Metropolis-Hastings criterion — also detailed balance. These two ingredients are sufficient to guarantee that hierarchical partitions are eventually sampled from the correct posterior distribution. However, in practice, the equilibration time will depend strongly on the move proposals, and will become shorter if they are close to the actual posterior. The simplest choice we could make is to select from all groups with equal probability

$$P(b_i^{(l)} = r \to s) = \frac{1}{B_l + 1}, \tag{80}$$

where we also account for the occupation of a new group, which if the move is accepted, will increase B_l by one (provided the node i is not the last one in its current group). Since this likelihood is always nonzero, it fulfills our requirements. However, it will lead to very large equilibration times, in particular for large values of B_l. This is because the actual posterior likelihood for node i is likely to be concentrated only in a small subset of all possible groups, and hence most such fully random proposals will simply be rejected. A better approach was developed in Ref. [9], and it consists in inspecting the current parameters of the model to provide a better guess of the posterior. It amounts to making move proposals according to

$$P(b_i^{(l)} = r \to s) = \sum_t P(t|i) \frac{e_{ts}^l + \epsilon}{e_t^l + \epsilon(B_l + 1)}, \tag{81}$$

where $P(t|i) = \sum_j A_{ij}\delta_{b_j,t}/k_i$ is the fraction of neighbors of node i that belong to group t, and $\epsilon > 0$ is an arbitrary parameter that enforces ergodicity, but with no other significant impact in the algorithm, provided it is sufficiently small. It is worthwhile to emphasize that these move proposals do not bias the partitions toward any particular mixing pattern. For example, they do not prefer assortative versus non-assortative partitions, since they inspect the neighbors of a node only to access with other groups their kinds are typically connected — which can be different from the the group assignment of the original node. Furthermore, these proposals can be generated efficiently, simply by

1. sampling a random neighbor j of node i, and inspecting its group membership $t = b_j$, and then
2. with probability $\epsilon(B_l+1)/(e_t+\epsilon(B_l+1))$ sampling a fully random group s (which can be a new group),
3. or otherwise, sampling a group label s with a probability proportional to the number of edges leading to it from group t, e_{ts}.

The above can be done in time $O(k_i)$, again independently of B_l, as long as a continuous book-keeping is made of the edges which are incident to each group, and therefore it does not affect the overall $O(E)$ time complexity. As reported in Ref. [9], these move proposals tend to significantly improve the mixing times, and remove an explicit dependency on the number of groups, that would otherwise be present with the fully random moves.

This approach is also more efficient than the rejection-free "heat bath" algorithm used in Ref. [17], since the latter requires all possible moves to be probed, incurring an additional time complexity that grows linearly with the number of groups.

In addition to the move proposals, another crucial aspect of the algorithm's efficiency is the choice of the starting state. A simple approach such as starting from a random partition can lead to metastable states, from which it takes a long time to escape. Instead, here we adopt the agglomerative initialization approach presented in Ref. [9], which amounts to putting each node in their own group, and then progressively merging groups, while

alternatingly allowing for individual node moves. This can be done for each hierarchical level iteratively, as described in detail in Ref. [10]. As reported in Ref. [9], this approach greatly reduces the tendency to get trapped in a metastable state, and serves as an initialization protocol that further reduces the overall mixing time of the MCMC.

While the above algorithm serves to sample from the posterior distribution of Eq. 74, it can be easily modified to find its maximum by introducing an "inverse-temperature" parameter β in Eq. 78 via the replacement $\Delta \ln P_l \to \beta \Delta \ln P_l$. By making $\beta \to \infty$ the algorithm is turned into a greedy heuristic that, if repeated many times, yields a reliable estimate of the maximum.

The lack of an explicit dependence on the number of groups of the algorithm above is atypical, since most other proposed Bayesian (or semi-Bayesian) algorithms have either quadratic $O(EB^2)$ [15–17, 35] or linear $O(EB)$ [14, 43] dependencies, which means that those can be applied to large networks only if the number of groups is kept small. Furthermore, the increased efficiency obtained here does not rely on any approximations made to the likelihood.

A reference implementation of the algorithm is freely available as part of the `graph-tool` library [44][45].

VII. MODEL COMPARISON

With the different model flavors available (NDC-SBM, DC-SBM with uniform degree prior or uniform hyperprior) we are left with the problem of deciding which offers the best description of a given network. This problem can be formulated in at least two ways, depending on whether we want to compare individual partitions or entire model classes, which we describe now detail.

If we wish to compare two individual partitions, obtained from the posterior distribution of two different models, wee need to consider the joint posterior likelihood $P(\{\boldsymbol{b}_l\}, \mathcal{H}|\boldsymbol{A})$, where $\mathcal{H}$ is the model class being used. For example, when comparing results from the DC-SBM and NDC-SBM, we can compute the ratio,

$$\Lambda_1 = \frac{P(\{\boldsymbol{b}_l\}, \mathcal{H}_{\text{NDC}}|\boldsymbol{A})}{P(\{\boldsymbol{b}_l\}', \mathcal{H}_{\text{DC}}|\boldsymbol{A})} \tag{82}$$

$$= \frac{P(\boldsymbol{A}, \{\boldsymbol{b}_l\}|\mathcal{H}_{\text{NDC}})}{P(\boldsymbol{A}, \{\boldsymbol{b}_l\}'|\mathcal{H}_{\text{DC}})} \times \frac{P(\mathcal{H}_{\text{NDC}})}{P(\mathcal{H}_{\text{DC}})} \tag{83}$$

$$= \exp(-\Delta\Sigma) \tag{84}$$

where in the last equation $\Delta\Sigma = \Sigma_{\text{NDC}} - \Sigma_{\text{DC}}$ is the difference in the description length, and we have assumed that both model classes are equally likely a priori, $P(\mathcal{H}_{\text{NDC}}) = P(\mathcal{H}_{\text{DC}})$. If $\Lambda_1 < 1$, we have that the data favors the particular hierarchical partition $\{\boldsymbol{b}_l\}'$ together with the degree-corrected model variant, or if $\Lambda_1 > 1$ we have the opposite case. Choosing a model according to Λ_1 is identical to employing the MDL criterion, but its value can be used to quantify the degree of confidence. E.g. a value $\Lambda_1 = 1/2$ indicates a very modest evidence supporting the DC-SBM that cannot be reliably distinguished from pure chance, whereas a value of $\Lambda_1 = 1/10^5$ would clearly indicate that it is a much better model than the alternative.

The criterion above should not be confused with the "frequentist" approach of computing the *parametric* likelihood ratio between both models, as was done in Ref. [46]. In the latter case, which does not involve any prior likelihoods, the ratio needs to be compared to the distribution obtained with the null model, which is more cumbersome to obtain. However, as is understood in general (and can also be shown for the particular case of the SBM [22]), this frequentist criterion should coincide asymptotically with the Bayesian criterion above as long as uniform priors are used. On the other hand, since here we use deeper Bayesian hierarchies, and hence nonuniform priors, these amount to different tests, with Λ_1 being more sensitive to regularities in the data, since it uses properties of the parameters themselves in the decision.

The comparison above using Λ_1 is easy to perform, since it requires one to simply inspect the result of the inference procedure. However, it may be possible that the same network admits many alternative fits with very similar posterior likelihoods. A more strict Bayesian stance would require us to treat those on an equal footing, and any statement about the generative model behind the data should be averaged over all possible fits, weighted according to the respective posterior likelihood. Hence, in this scenario we may be interested instead in comparing the entire model classes to each other, which involves evaluating the so-called *model evidence* by summing over all hierarchical partitions,

$$P(\boldsymbol{A}|\mathcal{H}) = \sum_{\{\boldsymbol{b}_l\}} P(\boldsymbol{A}, \{\boldsymbol{b}_l\}). \tag{85}$$

With this, we can again compute the posterior odds ratio, e.g.

$$\Lambda_2 = \frac{P(\mathcal{H}_{\text{NDC}}|\boldsymbol{A})}{P(\mathcal{H}_{\text{DC}}|\boldsymbol{A})} = \frac{P(\boldsymbol{A}|\mathcal{H}_{\text{NDC}})}{P(\boldsymbol{A}|\mathcal{H}_{\text{DC}})} \times \frac{P(\mathcal{H}_{\text{NDC}})}{P(\mathcal{H}_{\text{DC}})}. \tag{86}$$

If we have no prior preference towards either model, $P(\mathcal{H}_{\text{NDC}}) = P(\mathcal{H}_{\text{DC}})$, the value of Λ_2 is known as the Bayes factor [47], and like Λ_1 can be used to establish a degree of confidence in the outcome.

Unfortunately, the exact computation of the sum in Eq. 85 is intractable. We therefore resort to a variational approach, firstly by writing

$$\ln P(\boldsymbol{A}|\mathcal{H}) = \ln \sum_{\{\boldsymbol{b}_l\}} P(\boldsymbol{A}, \{\boldsymbol{b}_l\}) \tag{87}$$

$$= \sum_{\{\boldsymbol{b}_l\}} q(\{\boldsymbol{b}_l\}) \ln P(\boldsymbol{A}, \{\boldsymbol{b}_l\}) \tag{88}$$

$$- \sum_{\{\boldsymbol{b}_l\}} q(\{\boldsymbol{b}_l\}) \ln q(\{\boldsymbol{b}_l\}), \tag{89}$$

$$\tag{90}$$

with

$$q(\{\boldsymbol{b}_l\}) = \frac{P(\boldsymbol{A},\{\boldsymbol{b}_l\})}{P(\boldsymbol{A})} \tag{91}$$

being precisely the posterior distribution of for the hierarchical partition that we obtain from with the MCMC algorithm used above. (Note that so far we have not made any approximations, with the identities above holding exactly.) The first term in Eq. 88 is easy to compute, as it amounts to the average log-likelihood (or minus the description length) of the partitions we obtain with the MCMC above,

$$\langle \ln P(\boldsymbol{A},\{\boldsymbol{b}_l\})\rangle = \sum_{\{\boldsymbol{b}_l\}} q(\{\boldsymbol{b}_l\}) \ln P(\boldsymbol{A},\{\boldsymbol{b}_l\}). \tag{92}$$

On the other hand, the second term in Eq. 89 amounts to the entropy of the posterior distribution,

$$H(\{\boldsymbol{b}_l\}) = -\sum_{\{\boldsymbol{b}_l\}} q(\{\boldsymbol{b}_l\}) \ln q(\{\boldsymbol{b}_l\}), \tag{93}$$

and measures how strongly it is concentrated. For example, in the extreme (and unrealistic) case where for each model being compared only one partition occurs with probability $q(\{\boldsymbol{b}_l\}) = 1$, the entropy will be zero, and we have that $\Lambda_1 = \Lambda_2$. Otherwise the entropy $H(\{\boldsymbol{b}_l\})$ will effectively measure how many partitions contribute to the average log-likelihood, so that a model class with a larger entropy will be preferred over another with less variance, even if their posterior likelihoods are on average the same. Unfortunately, the entropy $H(\{\boldsymbol{b}_l\})$ is notoriously difficult to compute exactly, even asymptotically via MCMC algorithms, and encapsulates the difficulty of computing Eq. 85 directly. A brute force approach simply does not work, since it would require keeping track of all visited hierarchical partitions, which grow combinatorially in number with system size. Other approaches such as thermodynamic integration [48], annealed importance sampling [49] and flat-histogram methods [50] are also possible, but tend to be significantly inefficient in comparison. Instead, here we make a so-called "mean field" assumption on the shape of $q(\{\boldsymbol{b}_l\})$ which assumes that it factorizes over all levels

$$q(\{\boldsymbol{b}_l\}) \approx q_i^1(\boldsymbol{b}_1)\prod_{l>1}\prod_i q_i^l(b_i^l). \tag{94}$$

For the first level we use the so-called "Bethe approximation" [51], which takes into account the correlation between adjacent nodes in the network,

$$q^1(\boldsymbol{b}_1) \approx \prod_{i<j}\left[q_{ij}^1(b_i^1,b_j^1)\right]^{A_{ij}}\prod_i\left[q_i^1(b_i^1)\right]^{1-k_i} \tag{95}$$

with $q_i^1(r)$ and $q_{ij}^1(r,s)$ obtained from the posterior node and edge marginals

$$q_i^l(r) = P(b_i^l = r|\boldsymbol{A}) = \sum_{\{\boldsymbol{b}^l\}\setminus b_i^l} P(b_i^l = r,\{\boldsymbol{b}^l\}\setminus b_i^l|\boldsymbol{A}), \tag{96}$$

$$\begin{aligned} q_{ij}^1(r,s) &= P(b_i^1 = r, b_j^1 = s|\boldsymbol{A}) \\ &= \sum_{\{\boldsymbol{b}^l\}\setminus\{b_i^1,b_j^1\}} P(b_i^1 = r, b_j^1 = s,\{\boldsymbol{b}^l\}\setminus\{b_i^1,b_j^1\}|\boldsymbol{A}), \end{aligned} \tag{97}$$

estimated with the MCMC algorithm above. For the upper levels $l > 1$ we cannot use the same approximation since the adjacency matrices will be in general multigraphs that will keep changing throughout the algorithm. Therefore we used above a mean-field approximation where the posterior factorizes over all nodes. With this we can finally write Eq. 87 as

$$\ln P(\boldsymbol{A}) \approx \langle \ln P(\boldsymbol{A},\{\boldsymbol{b}_l\})\rangle + \sum_l H_l \tag{98}$$

where

$$\begin{aligned} H_1 = &-\sum_{i<j} A_{ij}\sum_{rs} q_{ij}^1(r,s)\ln q_{ij}^1(r,s) \\ &-\sum_i (1-k_i)\sum_r q_i^1(r)\ln q_i^1(r) \end{aligned} \tag{99}$$

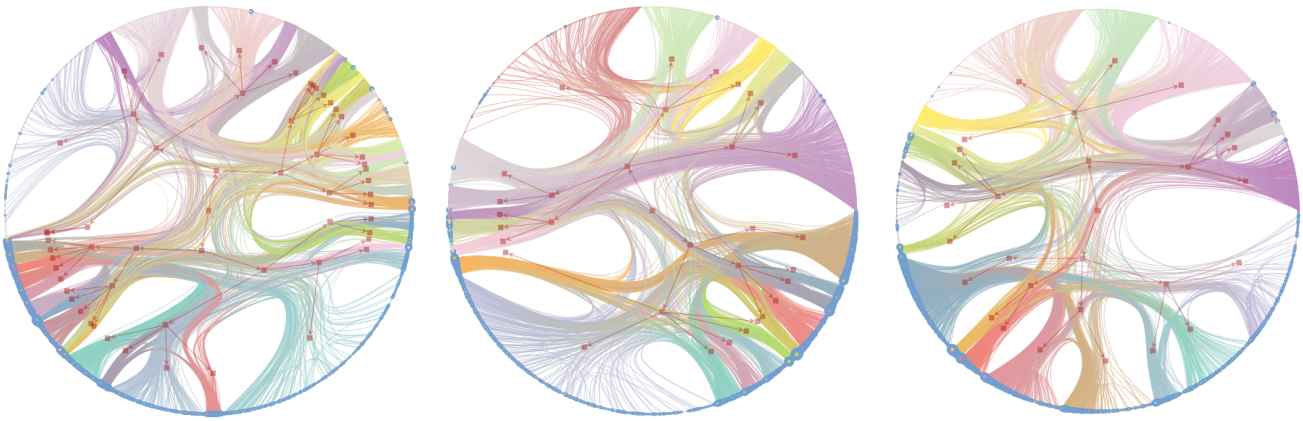
is the entropy of the first level and

$$H_l = -\sum_i\sum_r q_i^l(r)\ln q_i^l(r) \tag{100}$$

is the entropy of the remaining hierarchy levels $l > 1$. Thus, Eq. 98 can be computed simply by equilibrating the MCMC, obtaining the average log-likelihood and the node and edge posterior marginal distribution, $q_i^l(r)$ and $q_{ij}^1(r,s)$.

VIII. RESULTS FOR EMPIRICAL NETWORKS

We demonstrate the use of our approach on empirical networks (summarized in Table I), which we also use to compare different model variations. We begin with a network of political blogs compiled by Adamic et al [52] during the 2004 general election in the USA. In this network nodes are blogs, and an edge exists between two nodes if one blog cites the other (hence, the network is directed, and therefore the directed versions of the SBM were used, see Appendix B). This network was used in Ref. [11] as an example where the DC-SBM yielded more meaningful results, since it preferred a partition of the nodes that was largely compatible with the original categorization done in Ref. [52], based on the content of the blogs, into "liberal" and "conservative" sites. The NDC-SBM, on the other hand, preferred to divide the



(a) NDC-SBM, $B_1 = 42$, $\Sigma \approx 89938$ bits

(b) DC-SBM, uniform prior, $B_1 = 23$, $\Sigma \approx 87162$ bits

(c) DC-SBM, uniform hyperprior, $B_1 = 20$, $\Sigma \approx 84890$ bits

Figure 5. Most likely hierarchical partitions of a network of political blogs [52], according to the three model variants considered, as well as the number of groups B_1 at the bottom of the hierarchy, and the description length Σ. The nodes circled in blue were classified as "liberals" in Ref. [52] based on the content of the blogs. The layout is obtained with an algorithm by Holten [53].

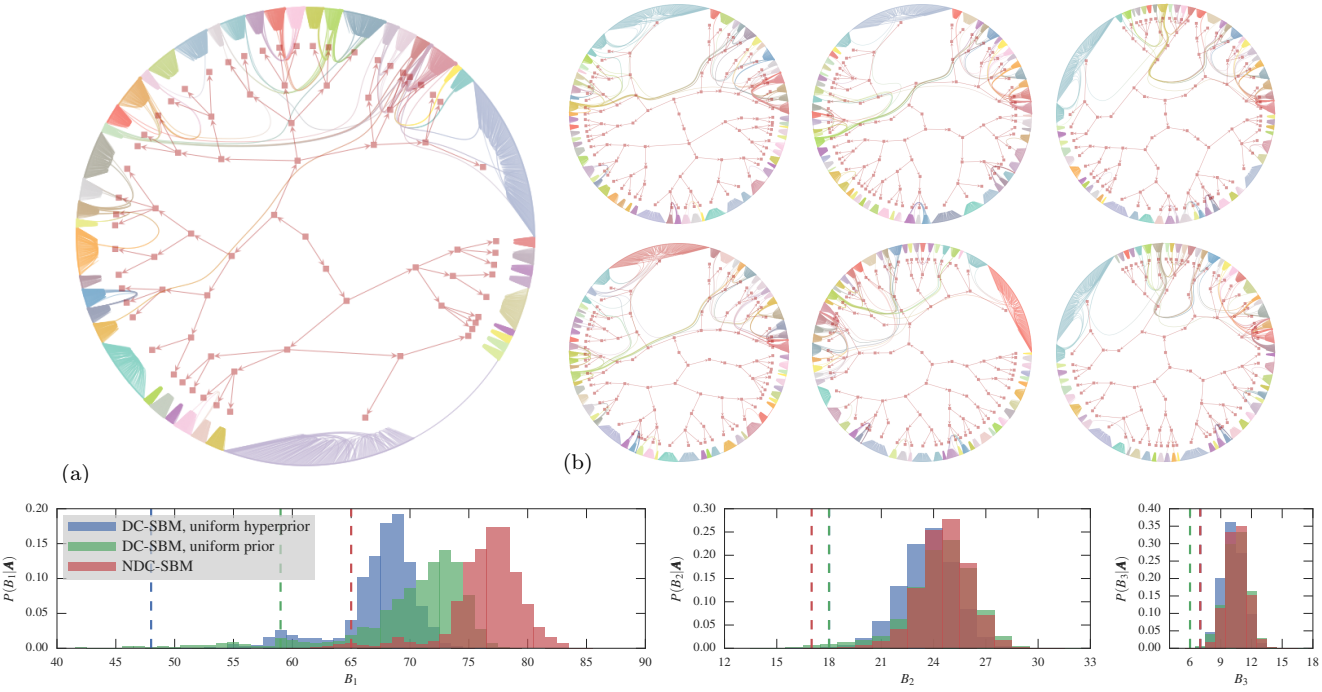


Figure 6. Hierarchical partitions of a network of collaboration between scientists [54]. (a) Most likely hierarchical partition according to the DC-SBM with a uniform hyperprior. (b) Uncorrelated samples from the posterior distribution. (c) Marginal posterior distribution of the number of groups at the first three hierarchical levels, according to the model variants described in the legend. The vertical lines mark the value obtained for the most likely partition.

nodes only according to degree. However, in that analysis the number of groups was fixed at $B = 2$. Using the nonparametric approach described here, where the number of groups is determined from data itself, the results show a less extreme amount of discrepancy, as seen in Fig. 5, which shows the most likely partition according to each model flavor. In all cases, the division of the nodes is largely compatible with the accepted one: The hierarchy branches at the top into the two political factions, and then proceeds into further sub-divisions inside each group. However, when inspecting the lower levels of the hierarchy, we see that the different variants yield distinct subdivisions inside the two main groups. The non-degree-corrected version yields the largest number

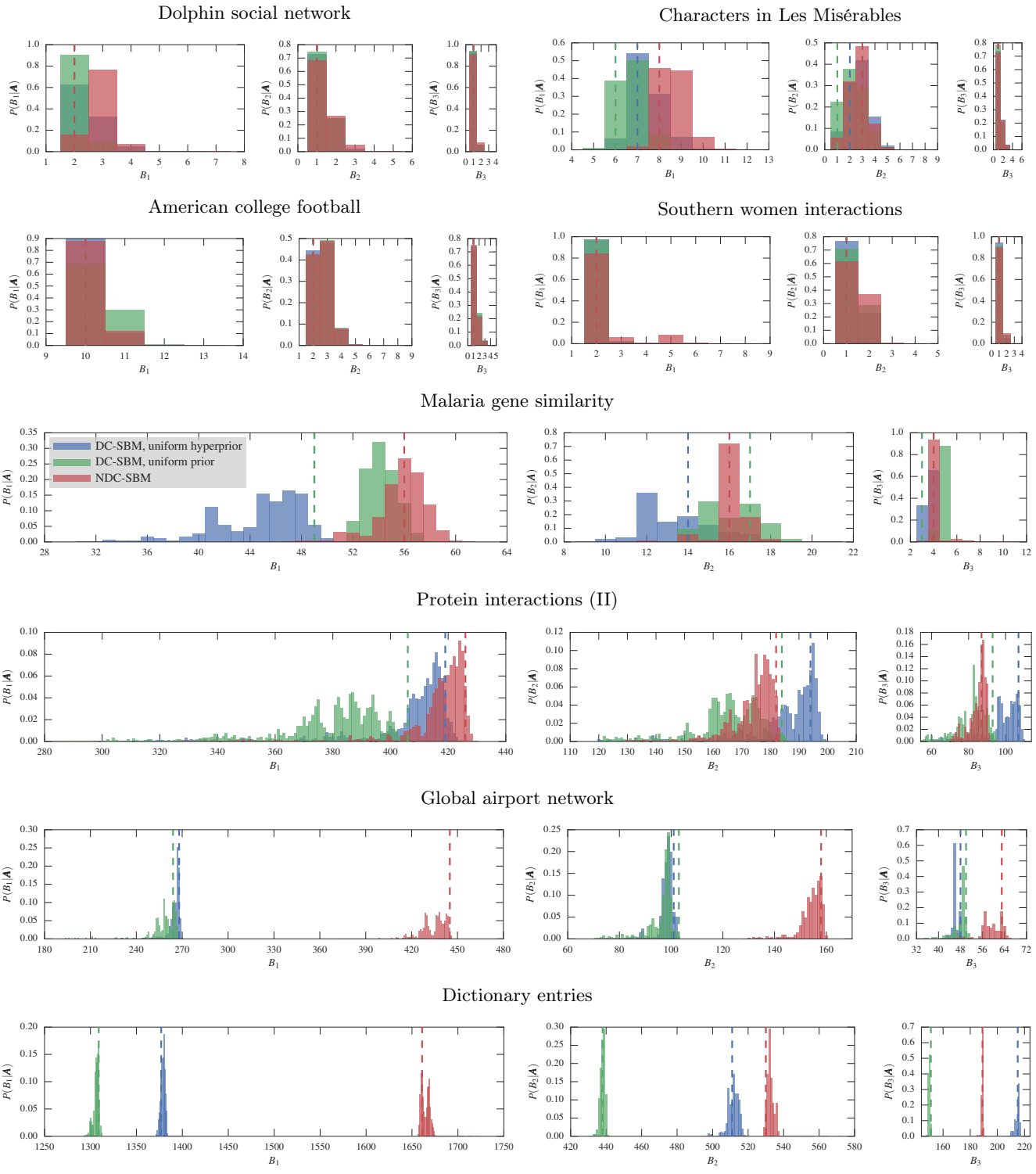


Figure 7. Marginal posterior distribution of the number of groups at the first three hierarchical levels, for several empirical networks, according to the model variants described in the legend. The vertical lines mark the value obtained for the most likely partition (the MDL criterion).

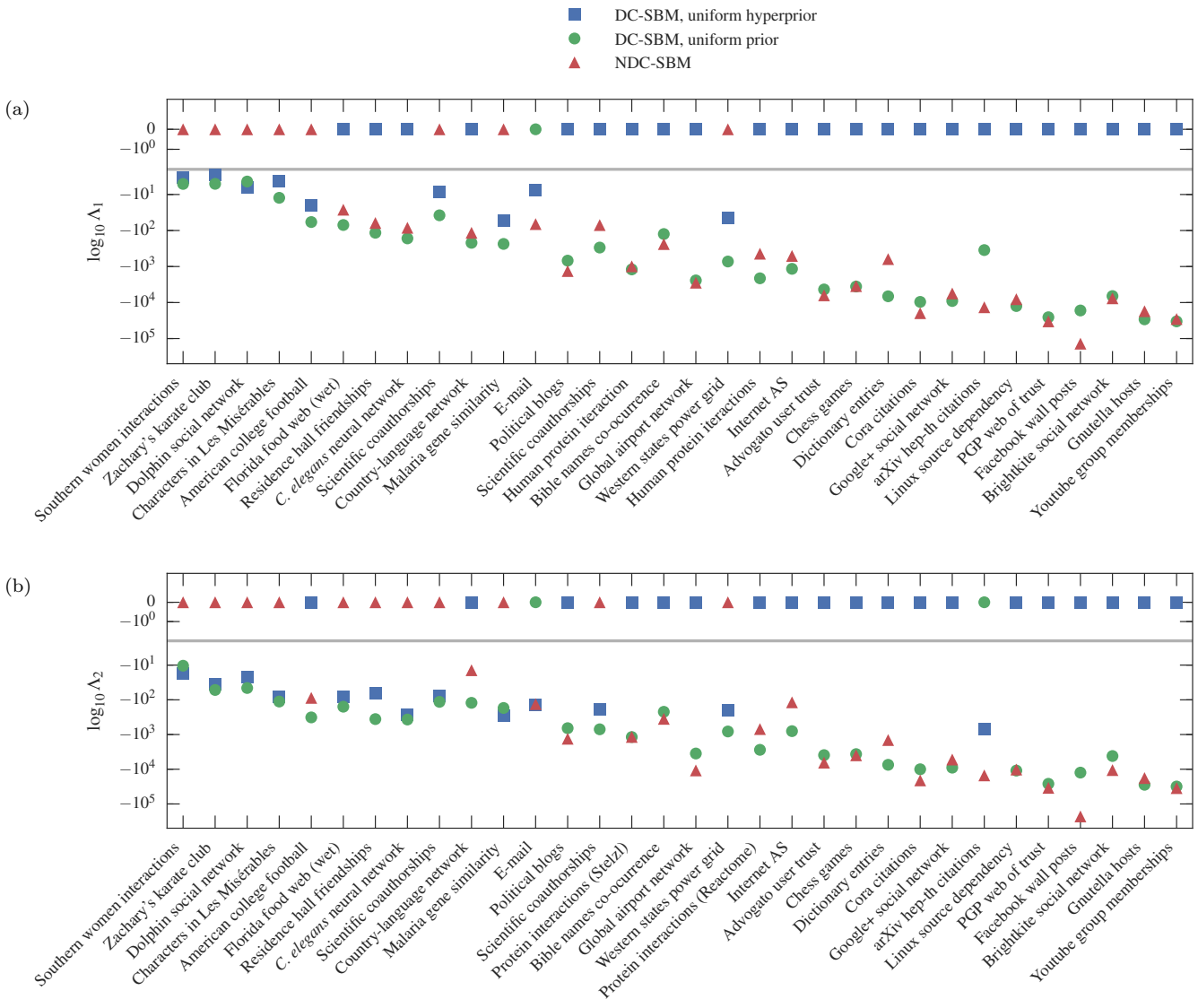


Figure 8. Posterior odds ratio relative to the best model, according to (a) the MDL criterion, Λ_1 (Eq. 82) and (b) full posterior likelihood, Λ_2 (Eq. 86) for the empirical networks listed in Table I. The solid lines mark a $\Lambda = 10^{-2}$ confidence threshold.

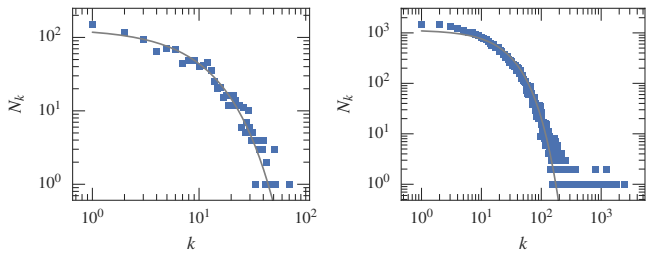


Figure 9. Degree histograms for the Email (left) and arXiv hep-th citations (right) networks. In both cases the solid lines show a geometric distribution $N_k = Np(1-p)^{k-1}$, with $p = 1/\langle k\rangle$.

Dataset	N	$\langle k\rangle$	B_1	$\langle B_1\rangle$	σ_{B_1}
Southern women interactions [55]	32	5.6	2	2.4	0.9
Zachary's karate club [56]	34	4.6	2	2.2	0.5
Dolphin social network [57]	62	5.1	2	2.9	0.5
Characters in Les Misérables [58]	77	6.6	8	8.6	0.7
American college football [59]	115	10.7	10	10.1	0.3
Florida food web (wet) [60]	128	32.9	14	14.2	0.4
Residence hall friendships [61]	217	24.6	20	20	0
C. elegans neural network [62]	297	15.9	20	13.5	0.5
Scientific coauthorships [63]	379	4.8	28	29.6	1.6
Country-language network [64]	868	2.9	4	10.1	1.9
Malaria gene similarity [65]	1,104	5.4	56	55.8	1.9
E-mail [66]	1,133	9.6	28	26.9	0.3
Political blogs [52]	1,222	31.2	15	15	0
Scientific coauthorships [63]	1,589	3.5	48	67.3	3.4
Protein iteractions (I) [67]	1,706	7.3	26	40.2	0.6
Bible names co-ocurrence [64]	1,773	10.3	63	79.1	5.3
Global airport network [10]	3,286	41.6	268	264.6	6.1
Western states power grid [68]	4,941	2.7	38	37.3	1
Protein iteractions (II) [69]	6,327	46.6	419	406.4	18.6
Internet AS [70]	6,474	4.3	40	50	7.2
Advogato user trust [71]	6,541	15.6	174	80.7	0.6
Chess games [64]	7,301	17.8	79	79	0
Dictionary entries [72]	13,356	18	1,378	1,378.9	2.3
Cora citations [73]	23,166	7.9	575	575	0.2
Google+ social network [74]	23,628	3.3	46	41.3	2.4
arXiv hep-th citations [70]	27,770	25.4	1,211	1,207.1	4
Linux source dependency [64]	30,837	13.9	448	384.7	3.1
PGP web of trust [75]	39,796	15.2	1,350	1,323.2	26.4
Facebook wall posts [76]	46,952	37.4	6,930	6,794.9	129.9
Brightkite social network [77]	58,228	7.4	171	177.4	3.8
Gnutella hosts [78]	62,586	4.7	24	24	0
Youtube group memberships [79]	124,325	4.7	273	266.7	4.7

Table I. Empirical networks used in this work, with their number of nodes N, average degree $\langle k\rangle = 2E/N$, number of groups at the lowest hierarchical level B_1 according to the MDL criterion, and the same value averaged from the posterior distribution $\langle B_1\rangle$, as well as standard deviation of the distribution, σ_{B_1}.

of groups, followed by the degree corrected one with uniform degree priors, and finally the version with uniform degree hyperpriors with the smallest number of groups. In this particular case, the models with smaller number of groups have also the smallest description length, which seems to indicate that the division into a larger number of groups are necessary for the models that are unable to otherwise properly explain the heterogeneity in the degree sequence. Thus, despite their uniform agreement with the accepted division, the MDL criterion still confirms the DC-SBM as a better model for this network.

We now move to a social network between scientist, where an edge exists if two scientists collaborated on a paper [54]. Here, we compare the results obtained by employing MDL (i.e. finding the most likely partition) and sampling many partitions from the posterior distribution, as shown in Fig. 6. We observe that while the sampled partitions share close similarities to the MDL result, there is a noticeable variance among the individual samples. Fig. 6 also shows the marginal distribution for the number of groups at the first three hierarchical levels. For all three model variants, the typical number of groups is significantly higher that what is obtained for the optimal partition (due to the low degree variability in this network, this is one of the few that are better modelled by the NDC-SBM, as seen in Fig. 8). This can be understood as an entropic effect, where the existence of a much larger number of more complex models with smaller yet comparable likelihood pushes the posterior distribution towards them. This is a good example of the bias-variance trade-off mentioned in Sec. III A, where we see that the MDL results in a more conservative partition, whereas the full posterior deposits more collective weight on larger models that are also more numerous. This seems to indicate that no single partition (and its associated model) serves as a overwhelmingly better explanation among those considered — a symptom that no specific model variant can perfectly accommodate the network structure, and thus that the SBM is possibly not a suitable generative model for this data.

This disagreement between MDL and posterior sampling is not universal, and depends strongly on the network structure. In Fig. 7 we show further results for other networks, that show a fair amount of diversity in this respect. In many cases the MDL estimate lies close to the mode of the posterior, indicating a fair amount of agreement (at least as far as the number of groups is concerned).

If we compare the different model flavors as outlined in Sec. VII, we obtain that most typically the DC-SBM with uniform degree hyperpriors provides the smallest description length for a large variety of networks, as shown in Fig. 8a. If we compare instead the whole model class, by summing over all partitions, we obtain largely consistent (though not identical) outcomes, as seen in Fig. 8b. Exceptions to this include networks where there is no significant statistical evidence to support the most complex models — either due to their small size or narrow degree distributions (e.g. Scientific coauthorships, Malaria gene similarity and Western states power grid) — and often the simpler NDC-SBM is preferred, as well as some networks for which the DC-SBM with uniform degree priors is preferred instead (E-mail, arXiv hep-th citations). A closer inspection of these networks reveal that their global degree distribution is fairly narrow, well approximated by an exponential distribution, as shown in Fig. 9. Since this is what is precisely assumed by the uniform degree prior, this model variation has the advantage in this case. It is worthwhile to observe that according to both criteria, the preference towards the DC-SBM over the NDC-SBM is sometimes only attained with the uni-

form degree hyperprior. In many cases the NDC-SBM yields a smaller description length or larger evidence than the degree-corrected variant with a uniform prior. This means that correcting for arbitrary degree frequencies — as opposed to simply the degrees but assuming uniform frequencies — can reveal important information on the structure of the network that would otherwise remain obscured.

IX. DISCUSSION

The microcanonical approach to the inference of large-scale network structures offers an opportunity to encode deeper Bayesian hierarchies into the generative models, which alleviates the underfitting problems present otherwise, while at the same time enabling the implementation of efficient inference algorithms with a complexity that is not explicitly dependent on the number of groups being inferred.

We showed how the degree-corrected SBM can be formulated in a Bayesian way, via the incorporation of priors for the degree sequence that depend on the degree distribution, and hence are more capable of decoupling modular organization from degree regularities. We have again visited the issue of the maximum number of groups that can be inferred, and determined that the hierarchical version of the model is significantly less susceptible to underfitting, by being able to uncover small groups in very large networks.

We also showed that the microcanonical model is identical to a Bayesian version of the typical canonical formulation, if we consider only its shallower version with uniform priors. Hence, the main strength of the approach presented here lies not in details of the model specification, but rather in the ease with which higher order Bayesian considerations can be incorporated.

Throughout the work we have contrasted two approaches to Bayesian inference, one where we search for the single best network parametrization (the MDL criterion), and the other where parametrizations are sampled according to their posterior likelihood. We showed that the bias-variance trade-off that these two options represent can manifest itself in practice, where a lack of quality of fit yields a disagreement between both approaches. By performing a systematic analysis of various empirical networks, we observed that the degree of discrepancy is varied, and itself serves as an indication of the suitability of the SBM in capturing the network structure.

We argue that the methods proposed here can be useful in the principled detection of large-scale network structures and in their interpretation. In particular we believe it can be used as a basis for a further understanding of the quality of the SBM family of models in capturing the properties of real networks.

[1] Santo Fortunato, "Community detection in graphs," Physics Reports **486**, 75–174 (2010).

[2] Santo Fortunato and Darko Hric, "Community detection in networks: A user guide," Physics Reports (2016), 10.1016/j.physrep.2016.09.002.

[3] Roger Guimerà, Marta Sales-Pardo, and Luís A. Nunes Amaral, "Modularity from fluctuations in random graphs and complex networks," Phys. Rev. E **70**, 025101 (2004).

[4] Santo Fortunato and Marc Barthélemy, "Resolution limit in community detection," PNAS **104**, 36–41 (2007).

[5] Benjamin H. Good, Yves-Alexandre de Montjoye, and Aaron Clauset, "Performance of modularity maximization in practical contexts," Phys. Rev. E **81**, 046106 (2010).

[6] Darko Hric, Richard K. Darst, and Santo Fortunato, "Community detection in networks: structural clusters versus ground truth," arXiv:1406.0146 [physics, q-bio] (2014), arXiv: 1406.0146.

[7] Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt, "Stochastic blockmodels: First steps," Social Networks **5**, 109–137 (1983).

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

[9] Tiago P. Peixoto, "Efficient Monte Carlo and greedy heuristic for the inference of stochastic block models," Phys. Rev. E **89**, 012804 (2014).

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

[11] Brian Karrer and M. E. J. Newman, "Stochastic blockmodels and community structure in networks," Phys. Rev. E **83**, 016107 (2011).

[12] B. Bollobás, "A probabilistic proof of an asymptotic formula for the number of labeled regular graphs." Eur. J. Comb. **1**, 311–316 (1980).

[13] Joseph Blitzstein and Persi Diaconis, "A Sequential Importance Sampling Algorithm for Generating Random Graphs with Prescribed Degrees," Internet Math. **6**, 489–522 (2010).

[14] Roger Guimerà and Marta Sales-Pardo, "Missing and spurious interactions and the reconstruction of complex networks," Proceedings of the National Academy of Sciences **106**, 22073 –22078 (2009).

[15] Xiaoran Yan, Yaojia Zhu, Jean-Baptiste Rouquier, and Cristopher Moore, "Active Learning for Hidden Attributes in Networks," arXiv:1005.0794 (2010).

[16] Etienne Côme and Pierre Latouche, "Model selection and clustering in stochastic block models based on the exact integrated complete data likelihood," Statistical Modelling **15**, 564–589 (2015).

[17] M. E. J. Newman and Gesine Reinert, "Estimating the Number of Communities in a Network," Phys. Rev. Lett. **117**, 078301 (2016).

[18] J. Rissanen, "Modeling by shortest data description," Automatica **14**, 465–471 (1978).

[19] Peter D. Grünwald, *The Minimum Description Length Principle* (The MIT Press, 2007).

[20] Pierre Latouche, Etienne Birmelé, and Christophe Am-

broise, "Bayesian Methods for Graph Clustering," in *Advances in Data Analysis, Data Handling and Business Intelligence*, Studies in Classification, Data Analysis, and Knowledge Organization, edited by Andreas Fink, Berthold Lausen, Wilfried Seidel, and Alfred Ultsch (Springer Berlin Heidelberg, 2009) pp. 229–239.

[21] Jake M. Hofman and Chris H. Wiggins, "Bayesian Approach to Network Modularity," Phys. Rev. Lett. **100**, 258701 (2008).

[22] Xiaoran Yan, "Bayesian Model Selection of Stochastic Block Models," arXiv:1605.07057 [cs, stat] (2016), arXiv: 1605.07057.

[23] Martin Rosvall and Carl T. Bergstrom, "An information-theoretic framework for resolving community structure in complex networks," PNAS **104**, 7327–7331 (2007).

[24] Tiago P. Peixoto, "Model Selection and Hypothesis Testing for Large-Scale Network Models with Overlapping Groups," Phys. Rev. X **5**, 011033 (2015).

[25] Note that this is not an issue when we are strictly maximizing the posterior, since the most likely partition will never contain empty groups.

[26] Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman, "Power-Law Distributions in Empirical Data," SIAM Rev. **51**, 661–703 (2009).

[27] George E. Andrews, *The Theory of Partitions* (Cambridge University Press, Cambridge, 1984).

[28] G. Szekeres, "An Asymptotic Formula in the Theory of Partitions," Q J Math **2**, 85–108 (1951).

[29] G. Szekeres, "Some Asymptotic Formulae in the Theory of Partitions (ii)," Q J Math **4**, 96–111 (1953).

[30] E. Rodney Canfield, "From Recursions to Asymptotics: On Szekeres' Formula for the Number of Partitions," The Electronic Journal of Combinatorics **4**, R6 (1996).

[31] Paul Erdős and Joseph Lehner, "The distribution of the number of summands in the partitions of a positive integer," Duke Math. J **8**, 335–345 (1941).

[32] G. H. Hardy and S. Ramanujan, "Une formule asymptotique pour le nombres des partitions de n," Comptes Rendus Acad. Sci. Paris, Sér. A **2** (1917).

[33] G. H. Hardy and S. Ramanujan, "Asymptotic Formulaae in Combinatory Analysis," Proceedings of the London Mathematical Society **s2-17**, 75–115 (1918).

[34] Tiago P. Peixoto, "Entropy of stochastic blockmodel ensembles," Phys. Rev. E **85**, 056122 (2012).

[35] Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová, "Inference and Phase Transitions in the Detection of Modules in Sparse Networks," Phys. Rev. Lett. **107**, 065701 (2011).

[36] Ginestra Bianconi, "Entropy of network ensembles," Phys. Rev. E **79**, 036114 (2009).

[37] Tiziano Squartini, Joey de Mol, Frank den Hollander, and Diego Garlaschelli, "Breaking of Ensemble Equivalence in Networks," Phys. Rev. Lett. **115**, 268701 (2015).

[38] Diego Garlaschelli, Frank den Hollander, and Andrea Roccaverde, "Ensemble nonequivalence in random graphs with modular structure," arXiv:1603.08759 (2016).

[39] Anne Condon and Richard M. Karp, "Algorithms for graph partitioning on the planted partition model," Random Structures & Algorithms **18**, 116–140 (2001).

[40] Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová, "Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications," Phys. Rev. E **84**, 066106 (2011).

[41] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller, "Equation of State Calculations by Fast Computing Machines," J. Chem. Phys. **21**, 1087 (1953).

[42] W. K. Hastings, "Monte Carlo sampling methods using Markov chains and their applications," Biometrika **57**, 97 –109 (1970).

[43] Prem K. Gopalan and David M. Blei, "Efficient discovery of overlapping communities in massive networks," PNAS **110**, 14534–14539 (2013).

[44] Tiago P. Peixoto, "The graph-tool python library," figshare (2014), 10.6084/m9.figshare.1164194.

[45] Available at `https://graph-tool.skewed.de`.

[46] Xiaoran Yan, Cosma Shalizi, Jacob E. Jensen, Florent Krzakala, Cristopher Moore, Lenka Zdeborová, Pan Zhang, and Yaojia Zhu, "Model selection for degree-corrected block models," J. Stat. Mech. **2014**, P05007 (2014).

[47] Sir Harold Jeffreys, *The Theory of Probability* (Oxford University Press, 1998).

[48] Daan Frenkel and Berend Smit Professor, *Understanding Molecular Simulation: From Algorithms to Applications*, 2nd ed. (Academic Press, San Diego, 2001).

[49] Radford M. Neal, "Annealed importance sampling," Statistics and Computing **11**, 125–139 (2001).

[50] Fugao Wang and D. P. Landau, "Efficient, Multiple-Range Random Walk Algorithm to Calculate the Density of States," Phys. Rev. Lett. **86**, 2050–2053 (2001).

[51] Marc Mezard and Andrea Montanari, *Information, Physics, and Computation* (Oxford University Press, 2009).

[52] Lada A. Adamic and Natalie 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, USA, 2005) pp. 36–43.

[53] D. Holten, "Hierarchical Edge Bundles: Visualization of Adjacency Relations in Hierarchical Data," IEEE Transactions on Visualization and Computer Graphics **12**, 741–748 (2006).

[54] M. E. J. Newman, "Finding community structure in networks using the eigenvectors of matrices," Phys. Rev. E **74**, 036104 (2006).

[55] Allison Davis and Burleigh B. Gardner, *Deep South: A Social Anthropological Study of Caste and Class*, revised ed. edition ed. (University of South Carolina Press, Columbia, S.C, 2009).

[56] Wayne W. Zachary, "An Information Flow Model for Conflict and Fission in Small Groups," Journal of Anthropological Research **33**, 452–473 (1977).

[57] David Lusseau, Karsten Schneider, Oliver J. Boisseau, Patti Haase, Elisabeth Slooten, and Steve M. Dawson, "The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations," Behav Ecol Sociobiol **54**, 396–405 (2003).

[58] Donald E. Knuth, *The Stanford GraphBase: A Platform for Combinatorial Computing*, 1st ed. (Addison-Wesley Professional, New York, N.Y. : Reading, Mass, 1993).

[59] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," Proceedings of the National Academy of Sciences **99**, 7821 –7826 (2002).

[60] Robert E. Ulanowicz and Donald L. DeAngelis, "Network analysis of trophic dynamics in south florida ecosystems," US Geological Survey Program on the South Florida Ecosystem **114** (2005).

[61] Linton C Freeman, Cynthia M Webster, and Deirdre M Kirke, "Exploring social structure using dynamic three-dimensional color images," Social Networks **20**, 109–118 (1998).
[62] J. G. White, E. Southgate, J. N. Thomson, and S. Brenner, "The structure of the nervous system of the nematode Caenorhabditis elegans," Philos. Trans. R. Soc. Lond., B, Biol. Sci. **314**, 1–340 (1986).
[63] M. E. J. Newman, "Modularity and community structure in networks," PNAS **103**, 8577–8582 (2006).
[64] Jérôme Kunegis, "KONECT: The Koblenz Network Collection," in *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13 Companion (ACM, New York, NY, USA, 2013) pp. 1343–1350.
[65] Daniel B. Larremore, Aaron Clauset, and Caroline O. Buckee, "A Network Approach to Analyzing Highly Recombinant Malaria Parasite Genes," PLOS Comput Biol **9**, e1003268 (2013).
[66] 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 (2003).
[67] Ulrich Stelzl, Uwe Worm, Maciej Lalowski, Christian Haenig, Felix H. Brembeck, Heike Goehler, Martin Stroedicke, Martina Zenkner, Anke Schoenherr, Susanne Koeppen, Jan Timm, Sascha Mintzlaff, Claudia Abraham, Nicole Bock, Silvia Kietzmann, Astrid Goedde, Engin Toksöz, Anja Droege, Sylvia Krobitsch, Bernhard Korn, Walter Birchmeier, Hans Lehrach, and Erich E. Wanker, "A Human Protein-Protein Interaction Network: A Resource for Annotating the Proteome," Cell **122**, 957–968 (2005).
[68] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," Nature **393**, 409–10 (1998).
[69] G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono, B. Jassal, G. R. Gopinath, G. R. Wu, L. Matthews, S. Lewis, E. Birney, and L. Stein, "Reactome: a knowledgebase of biological pathways," Nucl. Acids Res. **33**, D428–D432 (2005).
[70] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos, "Graph evolution: Densification and shrinking diameters," ACM Trans. Knowl. Discov. Data **1** (2007), 10.1145/1217299.1217301.
[71] P. Massa, M. Salvetti, and D. Tomasoni, "Bowling Alone and Trust Decline in Social Network Sites," in *Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing, 2009. DASC '09* (2009) pp. 658–663.
[72] Vladimir Batagelj, Andrej Mrvar, and Matjaz Zaversnik, "Network Analysis of texts," (2002).
[73] Lovro Šubelj and Marko Bajec, "Model of Complex Networks Based on Citation Dynamics," in *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13 Companion (ACM, New York, NY, USA, 2013) pp. 527–530.
[74] Jure Leskovec and Julian J. Mcauley, "Learning to Discover Social Circles in Ego Networks," in *Advances in Neural Information Processing Systems 25*, edited by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Curran Associates, Inc., 2012) pp. 539–547.
[75] Oliver Richters and Tiago P. Peixoto, "Trust Transitivity in Social Networks," PLoS ONE **6**, e18384 (2011).
[76] Bimal Viswanath, Alan Mislove, Meeyoung Cha, and Krishna P. Gummadi, "On the Evolution of User Interaction in Facebook," in *Proceedings of the 2Nd ACM Workshop on Online Social Networks*, WOSN '09 (ACM, New York, NY, USA, 2009) pp. 37–42.
[77] Eunjoon Cho, Seth A. Myers, and Jure Leskovec, "Friendship and mobility: user movement in location-based social networks," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11 (ACM, New York, NY, USA, 2011) pp. 1082–1090.
[78] Matei Ripeanu and Ian Foster, "Mapping the Gnutella Network: Macroscopic Properties of Large-Scale Peer-to-Peer Systems," in *Peer-to-Peer Systems*, Lecture Notes in Computer Science No. 2429, edited by Peter Druschel, Frans Kaashoek, and Antony Rowstron (Springer Berlin Heidelberg, 2002) pp. 85–93.
[79] Alan E. Mislove, *Online social networks: measurement, analysis, and applications to distributed information systems* (ProQuest, 2009).

Appendix A: Asymptotic degree distributions sampled from uniform priors and hyperpriors

We can easily obtain the expected degree distribution when using the uniform prior for the degree sequence in Eq. 24, if we relax the ensemble to allow the total number of edges to fluctuate, with the global constraint being enforced only on average. This canonical ensemble is not identical to the microcanonical one used in the main text, but will approach it asymptotically in the the thermodynamic limit, i.e. when the number of nodes and edges become sufficiently large.

If we focus on only one group with N nodes and E half edges on average, a degree sequence $\boldsymbol{k}$ will be sampled with probability

$$P(\boldsymbol{k}) = \frac{e^{-\lambda \sum_i k_i}}{Z} \tag{A1}$$

with the canonical partition function given by

$$Z = \sum_{\boldsymbol{k}} e^{-\lambda \sum_i k_i} = \left(1 - e^{-\lambda}\right)^{-N}, \tag{A2}$$

with $\lambda = \ln(1 + N/E)$ obtained by enforcing the constraint $E = \sum_i k_i = -\partial \ln Z/\partial\lambda$. From the above, we obtain immediately that the probability of a given node i having a degree k is

$$P(k_i = k) = e^{-\lambda k} \frac{e^{-\lambda \sum_{j\neq i} k_j}}{Z} = (1 - e^{-\lambda})e^{-\lambda k}. \tag{A3}$$

This is a geometric distribution, more commonly parametrized as

$$P(k) = (1 - p)p^k, \tag{A4}$$

with an average $\langle k\rangle = (1 - p)/p = E/N$.

We can use the same approach to obtain the expected degree distribution generated from the uniform hyperprior of Eq. 27, which is somewhat more involved, but it

is still quite feasible. We want to consider the ensemble of non-negative integer counts $\{n_k\}$, subject to a normalization constraint $\sum_{k=0}^{\infty} n_k = N$ and a fixed average $\sum_{k=0}^{\infty} k n_k = E$. The canonical partition function of this ensemble is

$$Z = \sum_{\{n_k\}} e^{-\lambda \sum_k n_k - \mu \sum_k k n_k} = \prod_k Z_k, \tag{A5}$$

with

$$Z_k = \frac{1}{1 - \exp(-\lambda - \mu k)}. \tag{A6}$$

The expected degree counts are given by

$$\langle n_k \rangle = -\frac{\partial \ln Z_k}{\partial \lambda} = \frac{1}{\exp(\lambda + \mu k) - 1}, \tag{A7}$$

which is the Bose-Einstein distribution. The parameters λ and μ are determined via the imposed constraints,

$$\sum_{k=0}^{\infty} \frac{1}{\exp(\lambda + \mu k) - 1} = N, \tag{A8}$$

$$\sum_{k=0}^{\infty} \frac{k}{\exp(\lambda + \mu k) - 1} = E. \tag{A9}$$

For sufficiently large E and N, the sums may be approximated by integrals, and using the polylogarithm function, $\mathrm{Li}_s(z) = \Gamma(s)^{-1} \int_0^{\infty} [t^{s-1}/(e^t/z - 1)] \mathrm{d}t$, we have

$$\int_0^{\infty} \frac{\mathrm{d}k}{\exp(\lambda + \mu k) - 1} = \frac{\mathrm{Li}_1(e^{-\lambda})}{\mu} = N, \tag{A10}$$

$$\int_0^{\infty} \frac{k\,\mathrm{d}k}{\exp(\lambda + \mu k) - 1} = \frac{\mathrm{Li}_2(e^{-\lambda})}{\mu^2} = E. \tag{A11}$$

Eq. A10 can be solved for λ as $e^{-\lambda} = 1 - \exp(-N/\mu)$, but the same cannot be done for Eq. A11 in closed form. However, for $N \gg \mu$, we have $\lambda \to 0$, and hence $\mu \approx \sqrt{\mathrm{Li}_2(1)/E} = \sqrt{\zeta(2)/E}$, with $\zeta(s)$ being the Riemann zeta function. This yields the asymptotic distribution,

$$\langle n_k \rangle \approx \frac{1}{\exp\left(k\sqrt{\zeta(2)/E}\right) - 1}. \tag{A12}$$

Its variance can be obtained from the second moment,

$$N\langle k^2 \rangle = \int_0^{\infty} \frac{k^2\,\mathrm{d}k}{\exp(\lambda + \mu k) - 1} = \frac{\mathrm{Li}_3(e^{-\lambda})}{2\mu^3}, \tag{A13}$$

which leads to

$$\langle k^2 \rangle = \frac{\zeta(3)}{2} \left(\frac{\langle k \rangle}{\zeta(2)}\right)^{3/2} \sqrt{N}, \tag{A14}$$

which diverges in the limit $N \gg 1$. For degrees $k \ll \sqrt{E}$, we have $\exp(k\sqrt{\zeta(2)/E}) \approx 1 + k\sqrt{\zeta(2)/E}$, and hence the expected distribution of Eq. A12 will follow a power law $1/k$ for small arguments, with an exponential cut-off for larger arguments,

$$\langle n_k \rangle \approx \begin{cases} \sqrt{E/\zeta(2)}/k & \text{for } k \ll \sqrt{E}, \\ \exp(-k\sqrt{\zeta(2)/E}) & \text{for } k \gg \sqrt{E}. \end{cases} \tag{A15}$$

This gives us a purely entropic explanation for broad discrete distributions, which is synonymous with the fact that most of them will approach asymptotically the above shape, if their average is constrained.

Appendix B: Directed networks

Although in the main text we focused on undirected networks, directed model variants are easy to obtain, as we summarize here. For the directed DC-SBM we have the model likelihood

$$P(\boldsymbol{A}|\boldsymbol{k}, \boldsymbol{e}, \boldsymbol{b}) = \frac{\prod_i k_i^+! k_i^-! \prod_{rs} e_{rs}!}{\prod_r e_r^+! e_r^-! \prod_{ij} A_{ij}!}, \tag{B1}$$

with $k_i^+ = \sum_j A_{ji}$, $k_i^- = \sum_j A_{ij}$, $e_r^+ = \sum_s e_{sr}$, $e_r^- = \sum_s e_{rs}$. For the hierarchical prior of edge counts, we have to treat the multigraphs as directed,

$$P(\boldsymbol{e}_l|\boldsymbol{e}_{l+1}, \boldsymbol{b}_l) = \prod_{rs} \left(\!\!\binom{n_r^l n_s^l}{e_{rs}^{l+1}}\!\!\right)^{-1}. \tag{B2}$$

The uniform degree prior is the product of two priors, for the in- and out-degree sequences,

$$P(\boldsymbol{k}|\boldsymbol{e}, \boldsymbol{b}) = \prod_r \left(\!\!\binom{n_r}{e_r^+}\!\!\right)^{-1} \left(\!\!\binom{n_r}{e_r^-}\!\!\right)^{-1}. \tag{B3}$$

Analogously for the conditioned degree prior we need to account for the joint (in, out)-degree distribution,

$$P(\boldsymbol{k}|\boldsymbol{\eta}) = \prod_r \frac{\prod_{k^+,k^-} \eta^r_{k^+,k^-}!}{n_r!} \tag{B4}$$

and an uniform hyperprior

$$P(\boldsymbol{\eta}|\boldsymbol{e}, \boldsymbol{b}) = \prod_r q(e_r^+, n_r)^{-1} q(e_r^-, n_r)^{-1}. \tag{B5}$$

The NDC-SBM is also entirely analogous, corresponding to a degree likelihood

$$P(\boldsymbol{k}|\boldsymbol{e}, \boldsymbol{b}) = \prod_r \frac{e_r^+!}{n_r^{e_r+} \prod_{i\in r} k_i^+!} \prod_r \frac{e_r^-!}{n_r^{e_r-} \prod_{i\in r} k_i^-!}, \tag{B6}$$

which yields the model likelihood

$$P(\boldsymbol{A}|\boldsymbol{e}, \boldsymbol{b}) = \frac{\prod_{rs} e_{rs}!}{\prod_r n_r^{e_r^+} n_r^{e_r^-}} \times \frac{1}{\prod_{ij} A_{ij}!}. \tag{B7}$$