

# Bayesian stochastic blockmodeling<sup>a</sup>

Tiago P. Peixoto<sup>†</sup>

*Department of Mathematical Sciences and Centre for Networks  
and Collective Behaviour, University of Bath, United Kingdom and*

*ISI Foundation, Turin, Italy*

This chapter provides a self-contained introduction to the use of Bayesian inference to extract large-scale modular structures from network data, based on the stochastic blockmodel (SBM), as well as its degree-corrected and overlapping generalizations. We focus on non-parametric formulations that allow their inference in a manner that prevents overfitting, and enables model selection. We discuss aspects of the choice of priors, in particular how to avoid underfitting via increased Bayesian hierarchies, and we contrast the task of sampling network partitions from the posterior distribution with finding the single point estimate that maximizes it, while describing efficient algorithms to perform either one. We also show how inferring the SBM can be used to predict missing and spurious links, and shed light on the fundamental limitations of the detectability of modular structures in networks.

arXiv:1705.10225v7 [stat.ML] 26 Nov 2018

---

<sup>a</sup>To appear in “Advances in Network Clustering and Blockmodeling,” edited by P. Doreian, V. Batagelj, A. Ferligoj, (Wiley, New York, 2019 [forthcoming]).

<sup>†</sup> t.peixoto@bath.ac.uk

**CONTENTS**

I. Introduction	3
II. Structure versus randomness in networks	3
III. The stochastic blockmodel (SBM)	5
IV. Bayesian inference: the posterior probability of partitions	7
V. Microcanonical models and the minimum description length principle (MDL)	11
VI. The “resolution limit” underfitting problem, and the nested SBM	13
VII. Model variations	17
A. Model selection	18
B. Degree correction	18
C. Group overlaps	22
D. Further model extensions	25
VIII. Efficient inference using Markov chain Monte Carlo (MCMC)	26
IX. To sample or to optimize?	28
X. Generalization and prediction	31
XI. Fundamental limits of inference: the detectability-indetectability phase transition	35
XII. Conclusion	38
References	39

## I. INTRODUCTION

Since the past decade and a half there has been an ever-increasing demand to analyze network data, in particular those stemming from social, biological and technological systems. Often these systems are very large, comprising millions of even billions of nodes and edges, such as the World Wide Web, and the global-level social interactions among humans. A particular challenge that arises is how to describe the large-scale structures of these systems, in a way that abstracts away from low-level details, allowing us to focus instead on “the big picture.” Differently from systems that are naturally embedded in some low-dimensional space — such as the population density of cities or the physiology of organisms — we are unable just to “look” at a network and readily extract its most salient features. This has prompted a flurry of activity in developing algorithmic approaches to extract such global information in a well-defined manner, many of which are described in the remaining chapters of this book. Most of them operate on a rather simple ansatz, where we try to divide the network into “building blocks,” which then can be described at an aggregate level in a simplified manner. The majority of such methods go under the name “community detection,” “network clustering” or “blockmodeling.” In this chapter we consider the situation where the ultimate objective when analyzing network data in this way is to *model* it, i.e. we want to make statements about possible generative mechanisms that are responsible for the network formation. This overall aim sets us in a well-defined path, where we get to formulate probabilistic models for network structure, and use principled and robust methods of statistical inference to fit our models to data. Central to this approach is the ability to distinguish structure from randomness, so that we do not fool ourselves into believing that there are elaborate structures in our data which are in fact just the outcome of stochastic fluctuations — which tends to be the Achilles’ heel of alternative nonstatistical approaches. In addition to providing a description of the data, the models we infer can also be used to generalize from observations, and make statements about what has *not* yet been observed, yielding something more tangible than mere interpretations. In what follows we will give an introduction to this inference approach, which includes recent developments that allow us to perform it in a consistent, versatile and efficient manner.

## II. STRUCTURE VERSUS RANDOMNESS IN NETWORKS

If we observe a random string of characters we will eventually encounter every possible substring, provided the string is long enough. This leads to the famous thought experiment of a large number of monkeys with typewriters: Assuming that they type randomly, for a sufficiently large number of monkeys any output can be observed, including, for example, the very text you are reading. Therefore, if we are ever faced with this situation, we should not be surprised if a such a text is in fact produced, and most importantly, we should not offer its simian author a place in a university department, as this occurrence is unlikely to be repeated. However, this example is of little practical relevance, as the number of monkeys necessary to type the text “blockmodeling” by chance is already of the order of  $10^{18}$ , and there are simply not that many monkeys.

Networks, however, are different from random strings. The network analogue of a random string is an Erdős-Rényi random graph [1] where each possible edge can occur with the same probability. But differently from a random string, a random graph can contain a wealth of structure before it becomes astronomically large — specially if we *search* for it. An example of this is shown in Fig. 1 for a modest network of 5,000 nodes, where its adjacency matrix is visualized using three different node orderings. Two of the orderings seem to reveal patterns of large-scale connections that are tantalizingly clear, and indeed would be eagerly captured by many network clustering

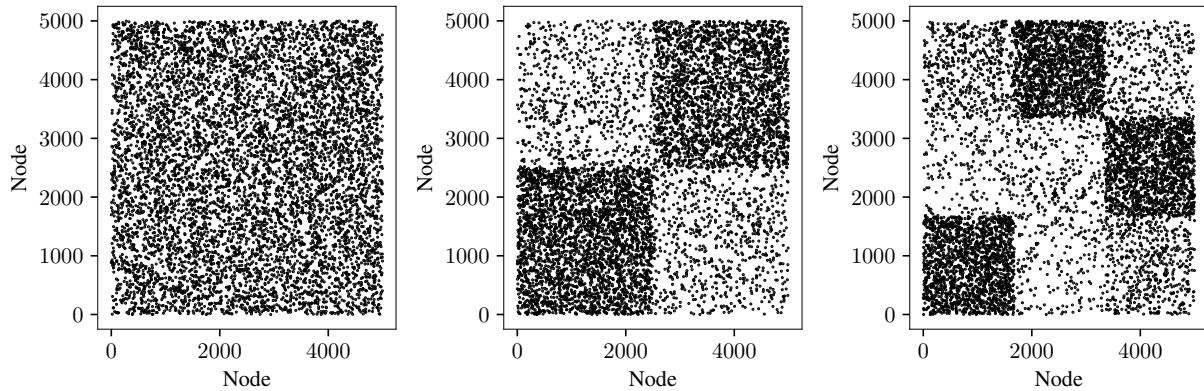


Figure 1. The three panels show the same adjacency matrix, with the only difference between them being the ordering of the nodes. The different orderings show seemingly clear, albeit very distinct patterns of modular structure. However, the adjacency matrix in question corresponds to an instance of a fully random Erdős-Rényi model, where each edge has the same probability  $p = \langle k \rangle / (N - 1)$  of occurring, with  $\langle k \rangle = 3$ . Although the patterns seen in the second and third panels are not mere fabrications — as they are really there in the network — they are also not meaningful descriptions of this network, since they arise purely out of random fluctuations. Therefore, the node groups that are identified via these patterns bear no relation to the generative process that produced the data. In other words, the second and third panels correspond each to an *overfit* of the data, where stochastic fluctuations are misrepresented as underlying structure. This pitfall can lead to misleading interpretations of results from clustering methods that do not account for statistical significance.

methods [2]. In particular, they seem to show groupings of nodes that have distinct probabilities of connections to each other — in direct contradiction to actual process that generated the network, where all connections had the same probability of occurring. What makes matters even worse is that in Fig. 1 is shown only a very small subset of all orderings that have similar patterns, but are otherwise very distinct from each other. Naturally, in the same way we should not confuse a monkey with a proper scientist in our previous example, we should not use any of these node groupings to explain why the network has its structure. Doing so should be considering *overfitting* it, i.e. mistaking random fluctuations for generative structure, yielding an overly complicated and ultimately wrong explanation for the data.

The remedy to this problem is to think probabilistically. We need to ascribe to each possible explanation of the data a probability that it is correct, which takes into account modeling assumptions, the statistical evidence available in the data, as well any source of prior information we may have. Imbued in the whole procedure must be the principle of parsimony — or Occam’s razor — where a simpler model is preferred if the evidence is not sufficient to justify a more complicated one.

In order to follow this path, before we look at any network data, we must first look in the “forward” direction, and decide on which mechanisms generate networks in the first place. Based on this, we will finally be able to look “backwards,” and tell which particular mechanism generated a given observed network.

### III. THE STOCHASTIC BLOCKMODEL (SBM)

As mentioned in the introduction, we wish to decompose networks into “building blocks,” by grouping together nodes that have a similar role in the network. From a generative point of view, we wish to work with models that are based on a partition of  $N$  nodes into  $B$  such building blocks, given by the vector  $\mathbf{b}$  with entries

$$b_i \in \{1, \dots, B\},$$

specifying the group membership of node  $i$ . We wish to construct a generative model that takes this division of the nodes as parameters, and generates networks with a probability

$$P(\mathbf{A}|\mathbf{b}),$$

where where  $\mathbf{A} = \{A_{ij}\}$  is the adjacency matrix. But what shape should  $P(\mathbf{A}|\mathbf{b})$  have? If we wish to impose that nodes that belong to the same group are statistically indistinguishable, our ensemble of networks should be fully characterized by the number of edges that connects nodes of two groups  $r$  and  $s$ ,

$$e_{rs} = \sum_{ij} A_{ij} \delta_{b_i,r} \delta_{b_j,s}, \quad (1)$$

or twice that number if  $r = s$ . If we take these as conserved quantities, the ensemble that reflects our maximal indifference towards any other aspect is the one that maximizes the entropy [3]

$$S = - \sum_{\mathbf{A}} P(\mathbf{A}|\mathbf{b}) \ln P(\mathbf{A}|\mathbf{b}) \quad (2)$$

subject to the constraint of Eq. 1. If we relax somewhat our requirements, such that Eq. 1 is obeyed only on expectation, then we can obtain our model using the method of Lagrange multipliers, using the Lagrangian function

$$F = S - \sum_{r \leq s} \mu_{rs} \left( \sum_{\mathbf{A}} P(\mathbf{A}|\mathbf{b}) \sum_{i < j} A_{ij} \delta_{b_i,r} \delta_{b_j,s} - \langle e_{rs} \rangle \right) - \lambda \left( \sum_{\mathbf{A}} P(\mathbf{A}|\mathbf{b}) - 1 \right) \quad (3)$$

where  $\langle e_{rs} \rangle$  are constants independent of  $P(\mathbf{A}|\mathbf{b})$ , and  $\boldsymbol{\mu}$  and  $\lambda$  are multipliers that enforce our desired constraints and normalization, respectively. Obtaining the saddle point  $\partial F / \partial P(\mathbf{A}|\mathbf{b}) = 0$ ,  $\partial F / \partial \mu_{rs} = 0$  and  $\partial F / \partial \lambda = 0$  gives us the maximum entropy ensemble with the desired properties. If we constrain ourselves to simple graphs, i.e.  $A_{ij} \in \{0, 1\}$ , without self-loops, we have as our maximum entropy model

$$P(\mathbf{A}|\mathbf{p}, \mathbf{b}) = \prod_{i < j} p_{b_i, b_j}^{A_{ij}} (1 - p_{b_i, b_j})^{1 - A_{ij}}. \quad (4)$$

with  $p_{rs} = e^{-\mu_{rs}} / (1 + e^{-\mu_{rs}})$  being the probability of an edge existing between any two nodes belonging to group  $r$  and  $s$ . This model is called the **stochastic blockmodel** (SBM), and has its roots in the social sciences and statistics [4–7], but has appeared repeatedly in the literature under a variety of different names [8–13]. By selecting the probabilities  $\mathbf{p} = \{p_{rs}\}$  appropriately, we can achieve arbitrary mixing patterns between the groups of nodes, as illustrated in Fig. 2. We stress that while the SBM can perfectly accommodate the usual “community structure” pattern [14], i.e. when the diagonal entries of  $\mathbf{p}$  are dominant, it can equally well describe a large variety of other

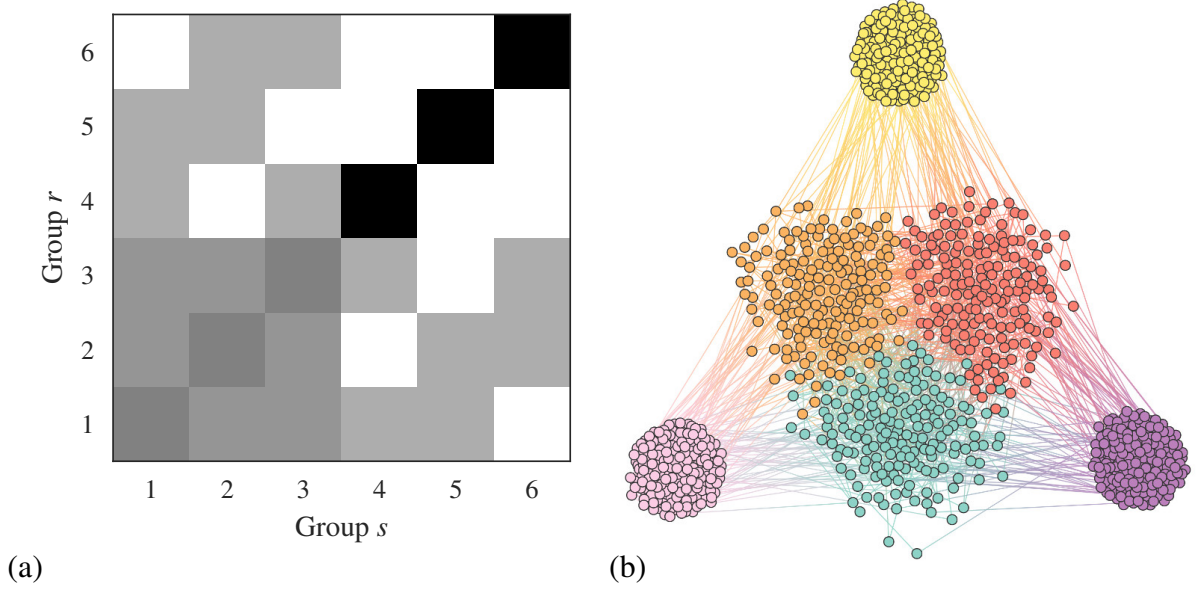


Figure 2. The stochastic blockmodel (SBM): (a) The matrix of probabilities between groups  $p_{rs}$  defines the large-scale structure of generated networks; (b) a sampled network corresponding to (a), where the node colors indicate the group membership.

patterns, such as bipartiteness, core-periphery, and many others.

Instead of simple graphs, we may consider *multigraphs* by allowing multiple edges between nodes, i.e.  $A_{ij} \in \mathbb{N}$ . Repeating the same procedure, we obtain in this case

$$P(\mathbf{A}|\boldsymbol{\lambda}, \mathbf{b}) = \prod_{i < j} \frac{\lambda_{b_i, b_j}^{A_{ij}}}{(\lambda_{b_i, b_j} + 1)^{A_{ij} + 1}}, \quad (5)$$

with  $\lambda_{rs} = e^{-\mu_{rs}} / (1 - e^{-\mu_{rs}})$  being the *average number* of edges existing between any two nodes belonging to group  $r$  and  $s$ . Whereas the placement of edges in Eq. 4 is given by a Bernoulli distribution, in Eq. 5 it is given by a geometric distribution, reflecting the different nature of both kinds of networks. Although these models are not the same, there is in fact little difference between the networks they generate in the *sparse limit* given by  $p_{rs} = \lambda_{rs} = O(1/N)$  with  $N \gg 1$ . We see this by noticing how their log-probabilities become asymptotically identical in this limit, i.e.

$$\ln P(\mathbf{A}|\mathbf{p}, \mathbf{b}) \approx \frac{1}{2} \sum_{rs} e_{rs} \ln p_{rs} - n_r n_s p_{rs} + O(1), \quad (6)$$

$$\ln P(\mathbf{A}|\boldsymbol{\lambda}, \mathbf{b}) \approx \frac{1}{2} \sum_{rs} e_{rs} \ln \lambda_{rs} - n_r n_s \lambda_{rs} + O(1). \quad (7)$$

Therefore, since most networks that we are likely to encounter are sparse [15], it does not matter which model we use, and we may prefer whatever is more convenient for our calculations. With this in mind, we may consider yet another variant, which uses instead a Poisson distribution to

sample edges [16],

$$P(\mathbf{A}|\boldsymbol{\lambda}, \mathbf{b}) = \prod_{i<j} \frac{e^{-\lambda_{b_i, b_j}} \lambda_{b_i, b_j}^{A_{ij}}}{A_{ij}!} \times \prod_i \frac{e^{-\lambda_{b_i, b_i}/2} (\lambda_{b_i, b_i}/2)^{A_{ii}/2}}{(A_{ii}/2)!}, \quad (8)$$

where now we also allow for self-loops. Like the geometric model, the Poisson model generates multigraphs, and it is easy to verify that it also leads to Eq. 7 in the sparse limit. This model is easier to use in some of the calculations that we are going to make, in particular when we consider important extensions of the SBM, therefore we will focus on it.<sup>1</sup>

The model above generates undirected networks. It can be very easily modified to generate directed networks instead, by making  $\lambda_{rs}$  an asymmetric matrix, and adjusting the model likelihood accordingly. The same is true for all model variations that are going to be used in the following sections. However, for the sake of conciseness we will focus only on the undirected case. We point out that the corresponding expressions for the directed case are readily available in the literature (e.g. Refs. [17–19]).

Now that we have defined how networks with prescribed modular structure are generated, we need to develop the reverse procedure, i.e. how to infer the modular structure from data.

#### IV. BAYESIAN INFERENCE: THE POSTERIOR PROBABILITY OF PARTITIONS

Instead of generating networks, our nominal task is to determine which partition  $\mathbf{b}$  generated an observed network  $\mathbf{A}$ , assuming this was done via the SBM. In other words, we want to obtain the probability  $P(\mathbf{b}|\mathbf{A})$  that a node partition  $\mathbf{b}$  was responsible for a network  $\mathbf{A}$ . By evoking elementary properties of conditional probabilities, we can write this probability as

$$P(\mathbf{b}|\mathbf{A}) = \frac{P(\mathbf{A}|\mathbf{b})P(\mathbf{b})}{P(\mathbf{A})} \quad (9)$$

with

$$P(\mathbf{A}|\mathbf{b}) = \int P(\mathbf{A}|\boldsymbol{\lambda}, \mathbf{b})P(\boldsymbol{\lambda}|\mathbf{b})d\boldsymbol{\lambda} \quad (10)$$

being the *marginal likelihood* integrated over the remaining model parameters, and

$$P(\mathbf{A}) = \sum_{\mathbf{b}} P(\mathbf{A}|\mathbf{b})P(\mathbf{b}) \quad (11)$$

is called the *evidence*, i.e. the total probability of the data under the model, which serves as a normalization constant in Eq. 9. Eq. 9 is known as **Bayes' rule**, and far from being only a simple mathematical step, it encodes how our prior beliefs about the model, i.e. before we observe any data — in the above represented by the *prior distributions*  $P(\mathbf{b})$  and  $P(\boldsymbol{\lambda}|\mathbf{b})$  — are affected by the observation, yielding the so-called *posterior distribution*  $P(\mathbf{b}|\mathbf{A})$ . The overall approach outlined above has been proposed to the problem of network inference by several authors [18–33],

<sup>1</sup>Although the Poisson model is not strictly a maximum entropy ensemble, the generative process behind it is easy to justify. We can imagine it as the random placement of exactly  $E$  edges into the  $N(N-1)/2$  entries of the matrix  $\mathbf{A}$ , each with a probability  $q_{ij}$  of attracting an edge, with  $\sum_{i<j} q_{ij} = 1$ , yielding a multinomial distribution  $P(\mathbf{A}|\mathbf{q}, E) = E! \prod_{i<j} q_{ij}^{A_{ij}} / A_{ij}!$  — where, differently from Eq. 8, the edge placements are not conditionally independent. But if we now sample the total number of edges  $E$  from a Poisson distribution  $P(E|\bar{E})$  with average  $\bar{E}$ , by exploiting the relationship between the multinomial and Poisson distributions, we have  $P(\mathbf{A}|\mathbf{q}) = \sum_E P(\mathbf{A}|\mathbf{q}, E)P(E|\bar{E}) = \prod_{i<j} e^{-\omega_{ij}} \omega_{ij}^{A_{ij}} / A_{ij}!$ , where  $\omega_{ij} = q_{ij}/\bar{E}$ , which does amount to conditionally independent edge placements. Making  $q_{ij} = \bar{E} \lambda_{b_i, b_j}$ , and allowing self-loops, we arrive at Eq. 8.

with different implementations that vary in some superficial details in the model specification, approximations used, and in particular in the choice of priors. Here we will not review or compare all approaches in detail, but rather focus on the most important aspects, while choosing a particular path that makes exact calculations possible.

The prior probabilities are a crucial element of the inference procedure, as they will affect the shape of the posterior distribution, and ultimately, our inference results. In more traditional scenarios, the choice of priors would be guided by previous observations of data that are believed to come from the same model. However, this is not an applicable scenario when considering networks, which are typically *singletons*, i.e. they are unique objects, instead of coming from a population (e.g. there is only one internet, one network of trade between countries, etc).<sup>2</sup> In the absence of such empirical prior information, we should try as much as possible to be guided by well defined principles and reasonable assumptions about our data, rather than *ad hoc* choices. A central proposition we will be using is the *principle of maximum indifference* about the model before we observe any data. This will lead us to so-called *uninformative* priors,<sup>3</sup> that are maximum entropy distributions that ascribe the same probability to each possible parameter combination [3]. These priors have the property that they do not bias the posterior distribution in any particular way, and thus let the data “speak for itself.” But as we will see in the following, the naive application of this principle will lead to adverse effects in many cases, and upon closer inspection we will often be able to identify aspects of the model that we should not be agnostic about. Instead, a more meaningful approach will be to describe higher-order aspects of the model with their own models. This can be done in a manner that preserves the unbiased nature of our results, while being able to provide a more faithful representation of the data.

We begin by choosing the prior for the partition,  $\mathbf{b}$ . The most direct uninformative prior is the “flat” distribution where all partitions into at most  $B = N$  groups are equally likely, namely

$$P(\mathbf{b}) = \frac{1}{\sum_{\mathbf{b}'} 1} = \frac{1}{a_N} \quad (12)$$

where  $a_N$  are the ordered Bell numbers [42], given by

$$a_N = \sum_{B=1}^N \left\{ \begin{matrix} N \\ B \end{matrix} \right\} B! \quad (13)$$

where  $\left\{ \begin{matrix} N \\ B \end{matrix} \right\}$  are the Stirling numbers of the second kind [43], which count the number of ways to partition a set of size  $N$  into  $B$  indistinguishable and nonempty groups (the  $B!$  in the above equation recovers the distinguishability of the groups, which we require). However, upon closer inspection we often find that such flat distributions are not a good choice. In this particular case, since there are many more partitions into  $B + 1$  groups than there are into  $B$  groups (if  $B$  is sufficiently smaller than  $N$ ), Eq. 12 will typically prefer partitions with a number of groups that is comparable to the number of nodes. Therefore, this uniform assumption seems to betray the principle of parsimony that we stated in the introduction, since it favors large models with many groups, before we even observe the data.<sup>4</sup> Instead, we may wish to be agnostic about the *number of groups itself*, by first

<sup>2</sup>One could argue that most networks change in time, and hence belong to a time series, thus possibly allowing priors to be selected from earlier observations of the same network. This is a potentially useful way to proceed, but also opens a Pandora’s box of dynamical network models, where simplistic notions of statistical stationarity are likely to be contradicted by data. Some recent progress has been made on the inference of dynamic networks [34–41], but this field is still in relative infancy.

<sup>3</sup>The name “uninformative” is something of a misnomer, as it is not really possible for priors to truly carry “no information” to the posterior distribution. In our context, the term is used simply to refer to *maximum entropy priors*, conditioned on specific constraints.

<sup>4</sup>Using constant priors such as Eq. 12 makes the posterior distribution proportional to the likelihood. Maximizing



sampling it from its own uninformative distribution  $P(B) = 1/N$ , and then sampling the partition conditioned on it

$$P(\mathbf{b}|B) = \frac{1}{\left\{ \begin{matrix} N \\ B \end{matrix} \right\} B!}, \quad (14)$$

since  $\left\{ \begin{matrix} N \\ B \end{matrix} \right\} B!$  is the number of ways to partition  $N$  nodes into  $B$  labelled groups.<sup>5</sup> Since  $\mathbf{b}$  is a parameter of our model, the number of groups  $B$  is called a *hyperparameter*, and its distribution  $P(B)$  is called a *hyperprior*. But once more, upon closer inspection we can identify further problems: If we sample from Eq. 14, most partitions of the nodes will occupy all the groups approximately equally, i.e. all group sizes will be the approximately the same. Is this something we want to assume before observing any data? Instead, we may wish to be agnostic about this aspect as well, and choose to sample first the distribution of group sizes  $\mathbf{n} = \{n_r\}$ , where  $n_r$  is the number of nodes in group  $r$ , forbidding empty groups,

$$P(\mathbf{n}|B) = \binom{N-1}{B-1}^{-1}, \quad (15)$$

since  $\binom{N-1}{B-1}$  is the number of ways to divide  $N$  nonzero counts into  $B$  nonempty bins. Given these randomly sampled sizes as a constraint, we sample the partition with a uniform probability

$$P(\mathbf{b}|\mathbf{n}) = \frac{\prod_r n_r!}{N!}. \quad (16)$$

This gives us finally

$$P(\mathbf{b}) = P(\mathbf{b}|\mathbf{n})P(\mathbf{n}|B)P(B) = \frac{\prod_r n_r!}{N!} \binom{N-1}{B-1}^{-1} N^{-1}. \quad (17)$$

At this point the reader may wonder if there is any particular reason to stop here. Certainly we can find some higher-order aspect of the group sizes  $\mathbf{n}$  that we may wish to be agnostic about, and introduce a “*hyperhyperprior*”, and so on, indefinitely. The reason why we should not keep recursively being more and more agnostic about higher-order aspects of our model is that it brings increasingly diminishing returns. In this particular case, if we assume that the individual group sizes are sufficiently large, we obtain asymptotically

$$\ln P(\mathbf{b}) \approx -NH(\mathbf{n}) + O(\ln N) \quad (18)$$

where  $H(\mathbf{n}) = -\sum_r (n_r/N) \ln(n_r/N)$  is the entropy of the group size distribution. The value  $\ln P(\mathbf{b}) \rightarrow -NH(\mathbf{n})$  is an information-theoretical limit that cannot be surpassed, regardless of how we choose  $P(\mathbf{n}|B)$ . Therefore, the most we can optimize by being more refined is a marginal factor  $O(\ln N)$  in the log-probability, which would amount to little practical difference in most cases.

In the above, we went from a purely flat uninformative prior distribution for  $\mathbf{b}$ , to a Bayesian hierarchy with three levels, where we sample first the number of groups, the groups sizes, and then finally the partition. In each of the levels we used maximum entropy distributions that are

such a posterior distribution is therefore entirely equivalent to a “non-Bayesian” maximum likelihood approach, and nullifies our attempt to prevent overfitting.

<sup>5</sup>We could have used simply  $P(\mathbf{b}|B) = 1/B^N$ , since  $B^N$  is the number of partitions of  $N$  nodes into  $B$  groups, which are allowed to be empty. However, this would force us to distinguish between the nominal and the actual number of groups (discounting empty ones) during inference [33], which becomes unnecessary if we simply forbid empty groups in our prior.

constrained by parameters that are themselves sampled from their own distributions at a higher level. In doing so, we removed some intrinsic assumptions about the model (in this case, number and sizes of groups), thereby postponing any decision on them until we observe the data. This will be a general strategy we will use for the remaining model parameters.

Having dealt with  $P(\mathbf{b})$ , this leaves us with the prior for the group to group connections,  $\boldsymbol{\lambda}$ . A good starting point is an uninformative prior conditioned on a global average,  $\bar{\lambda}$ , which will determine the expected density of the network. For a continuous variable  $x$ , the maximum entropy distribution with a constrained average  $\bar{x}$  is the exponential,  $P(x) = e^{-x/\bar{x}}/\bar{x}$ . Therefore, for  $\boldsymbol{\lambda}$  we have

$$P(\boldsymbol{\lambda}|\mathbf{b}) = \prod_{r \leq s} e^{-n_r n_s \lambda_{rs}/(1+\delta_{rs})\bar{\lambda}} n_r n_s / (1+\delta_{rs})\bar{\lambda}, \quad (19)$$

with  $\bar{\lambda} = 2E/B(B+1)$  determining the expected total number of edges,<sup>6</sup> where we have assumed the local average  $\langle \lambda_{rs} \rangle = \bar{\lambda}(1+\delta_{rs})/n_r n_s$ , such that that the expected number of edges  $e_{rs} = \lambda_{rs} n_r n_s / (1+\delta_{rs})$  will be equal to  $\bar{\lambda}$ , irrespective of the group sizes  $n_r$  and  $n_s$  [19]. Combining this with Eq. 8, we can compute the integrated marginal likelihood of Eq. 10 as

$$P(\mathbf{A}|\mathbf{b}) = \frac{\bar{\lambda}^E}{(\bar{\lambda}+1)^{E+B(B+1)/2}} \times \frac{\prod_{r < s} e_{rs}! \prod_r e_{rr}!!}{\prod_r n_r^{e_r} \prod_{i < j} A_{ij}! \prod_i A_{ii}!!}. \quad (20)$$

Just as with the node partition, the uninformative assumption of Eq. 19 also leads to its own problems, but we postpone dealing with them to Sec. VI. For now, we have everything we need to write the posterior distribution, with the exception of the model evidence  $P(\mathbf{A})$  given by Eq. 11. Unfortunately, since it involves a sum over all possible partitions, it is not tractable to compute the evidence exactly. However, since it is just a normalization constant, we will not need to determine it when optimizing or sampling from the posterior, as we will see in Sec. VIII. The numerator of Eq. 9, which is comprised of the terms that we can compute exactly, already contains all the information we need to proceed with the inference, and also has a special interpretation, as we will see in the next section.

The posterior of Eq. 9 will put low probabilities on partitions that are not backed by sufficient statistical evidence in the network structure, and it will not lead us to spurious partitions such as those depicted in Fig. 1. Inferring partitions from this posterior amounts to a so-called *non-parametric* approach; not because it lacks the estimation of parameters, but because the number of parameters itself, a.k.a. the *order* or *dimension* of the model, will be inferred as well. More specifically, the number of groups  $B$  itself will be an outcome of the inference procedure, which will be chosen in order to accommodate the structure in the data, without overfitting. The precise reasons why the latter is guaranteed might not be immediately obvious for those unfamiliar with Bayesian inference. In the following section we will provide an explanation by making a straightforward connection with information theory. The connection is based on a different interpretation of our model, which allow us to introduce some important improvements.

<sup>6</sup>More strictly, we should treat  $\bar{\lambda}$  just as another hyperparameter and integrate over its own distribution. But since this is just a global parameter, not affected by the dimension of the model, we can get away with setting its value directly from the data. It means we are pretending we know precisely the density of the network we are observing, which is not a very strong assumption. Nevertheless, readers that are uneasy with this procedure can rest assured that this can be completely amended once we move to microcanonical models in Sec. V (see footnote 15).

## V. MICROCANONICAL MODELS AND THE MINIMUM DESCRIPTION LENGTH PRINCIPLE (MDL)

We can re-interpret the integrated marginal likelihood of Eq. 20 as the joint likelihood of a *microcanonical* model given by<sup>7</sup>

$$P(\mathbf{A}|\mathbf{b}) = P(\mathbf{A}|\mathbf{e}, \mathbf{b})P(\mathbf{e}|\mathbf{b}), \quad (21)$$

where

$$P(\mathbf{A}|\mathbf{e}, \mathbf{b}) = \frac{\prod_{r<s} e_{rs}! \prod_r e_{rr}!!}{\prod_r n_r^{e_r} \prod_{i<j} A_{ij}! \prod_i A_{ii}!!}, \quad (22)$$

$$P(\mathbf{e}|\mathbf{b}) = \prod_{r<s} \frac{\bar{\lambda}^{e_{rs}}}{(\bar{\lambda} + 1)^{e_{rs}+1}} \prod_r \frac{\bar{\lambda}^{e_{rs}/2}}{(\bar{\lambda} + 1)^{e_{rs}/2+1}} = \frac{\bar{\lambda}^E}{(\bar{\lambda} + 1)^{E+B(B+1)/2}}, \quad (23)$$

and  $\mathbf{e} = \{e_{rs}\}$  is the matrix of edge counts between groups. The term “microcanonical” — borrowed from statistical physics — means that model parameters correspond to “hard” constraints that are *strictly* imposed on the ensemble, as opposed to “soft” constraints that are obeyed only on average. In the particular case above,  $P(\mathbf{A}|\mathbf{e}, \mathbf{b})$  is the probability of generating a multigraph  $\mathbf{A}$  where Eq. 1 is always fulfilled, i.e. the total number of edges between groups  $r$  and  $s$  is always exactly  $e_{rs}$ , without any fluctuation allowed between samples (see Ref. [19] for a combinatorial derivation). This contrasts with the parameter  $\lambda_{rs}$  in Eq. 8, which determines only the *average* number of edges between groups, which fluctuates between samples. Conversely, the prior for the edge counts  $P(\mathbf{e}|\mathbf{b})$  is a mixture of geometric distributions with average  $\bar{\lambda}$ , which does allow the edge counts to fluctuate, guaranteeing the overall equivalence. The fact that Eq. 21 holds is rather remarkable, since it means that — at least for the basic priors we used — these two kinds of model (“canonical” and microcanonical) cannot be distinguished from data, since their marginal likelihoods (and hence the posterior probability) are identical<sup>8</sup>.

With this microcanonical interpretation in mind, we may frame the posterior probability in an information-theoretical manner as follows. If a discrete variable  $x$  occurs with a probability mass  $P(x)$ , the asymptotic amount of information necessary to describe it is  $-\log_2 P(x)$  (if we choose bits as the unit of measurement), by using an optimal lossless coding scheme such as Huffman’s algorithm [44]. With this in mind, we may write the numerator of the posterior distribution in Eq. 9 as

$$P(\mathbf{A}|\mathbf{b})P(\mathbf{b}) = P(\mathbf{A}|\mathbf{e}, \mathbf{b})P(\mathbf{e}, \mathbf{b}) = 2^{-\Sigma}, \quad (24)$$

where the quantity

$$\Sigma = -\log_2 P(\mathbf{A}, \mathbf{e}, \mathbf{b}) \quad (25)$$

$$= -\log_2 P(\mathbf{A}|\mathbf{e}, \mathbf{b}) - \log_2 P(\mathbf{e}, \mathbf{b}) \quad (26)$$

is called the *description length* of the data [45, 46]. It corresponds to the asymptotic amount of information necessary to encode the data  $\mathbf{A}$  *together* with the model parameters  $\mathbf{e}$  and  $\mathbf{b}$ . There-

<sup>7</sup>Some readers may wonder why Eq. 21 should not contain a sum, i.e.  $P(\mathbf{A}|\mathbf{b}) = \sum_{\mathbf{e}} P(\mathbf{A}|\mathbf{e}, \mathbf{b})P(\mathbf{e}|\mathbf{b})$ . Indeed, that is the proper way to write a marginal likelihood. However, for the microcanonical model there is only one element of the sum that fulfills the constraint of Eq. 1, and thus yields a nonzero probability, making the marginal likelihood identical to the joint, as expressed in Eq. 21. The same is true for the partition prior of Eq. 17. We will use this fact in our notation throughout, and omit sums when they are unnecessary.

<sup>8</sup>This equivalence occurs for a variety of Bayesian models. For instance, if we flip a coin with a probability  $p$  of coming up heads, the integrated likelihood under a uniform prior after  $N$  trials in which  $m$  heads were observed is  $P(\mathbf{x}) = \int_0^1 p^m (1-p)^{N-m} dp = (N-m)!m!/(N+1)!$ . This is the same as the “microcanonical” model  $P(\mathbf{x}) = P(\mathbf{x}|m)P(m)$  with  $P(\mathbf{x}|m) = \binom{N}{m}^{-1}$  and  $P(m) = 1/(N+1)$ , i.e. the number of heads is sampled from a uniform distribution, and the coin flips are sampled randomly among those that have that exact number of heads.

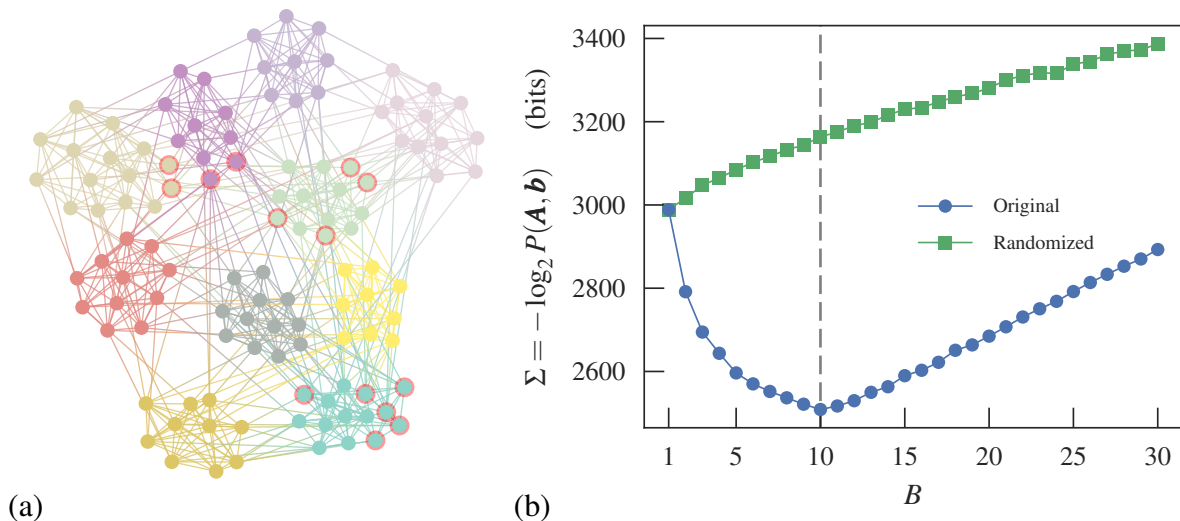


Figure 3. Bayesian inference of the SBM for a network of American college football teams [53]: (a) The partition that maximizes the posterior probability of Eq. 9, or equivalently, minimizes the description length of Eq. 24. Nodes marked in red are not classified according to the known division into “conferences.” (b) Description length as a function of the number of groups of the corresponding optimal partition, both for the original and randomized data.

fore, if we find a network partition that maximizes the posterior distribution of Eq. 20, we are also automatically finding one which minimizes the description length.<sup>9</sup> With this, we can see how the Bayesian approach outlined above prevents overfitting: As the size of the model increases (via a larger number of occupied groups), it will constrain itself better to the data, and the amount of information necessary to describe it when the model is known,  $-\log_2 P(\mathbf{A}|\mathbf{e}, \mathbf{b})$ , will decrease. At the same time, the amount of information necessary to describe the model itself,  $-\log_2 P(\mathbf{e}, \mathbf{b})$ , will increase as it becomes more complex. Therefore, the latter will function as a *penalty*<sup>10</sup> that prevents the model from becoming overly complex, and the optimal choice will amount to a proper balance between both terms.<sup>11</sup> Among other things, this approach will allow us to properly estimate the dimension of the model — represented by the number of groups  $B$  — in a parsimonious way.

We now illustrate this approach with a real-world dataset of American college football teams [53], where a node is a team and an edge exists if two teams played against each other in a season. If we find the partition that maximizes the posterior distribution, we uncover  $B = 10$  groups, as can be seen in Fig. 3a. If we compare this partition with the known division of the teams into “conferences” [54, 55], we find that they match with a high degree of precision, with the exception of only a few nodes.<sup>12</sup> In Fig. 3b we show the description length of the optimal partitions if we constrain them to have a pre-specified number of groups, which allows us to see how the approach penalizes

<sup>9</sup>Sometimes the minimum description length principle (MDL) is considered as an alternative method to Bayesian inference. Although it is possible to apply MDL in a manner that makes the connection with Bayesian inference difficult, as for example with the normalized maximum likelihood scheme [47, 48], in its more direct and tractable form it is fully equivalent to the Bayesian approach [46]. Note also that we do not in fact require the connection with microcanonical models made here, as the description length can be defined directly as  $\Sigma = -\log_2 P(\mathbf{A}, \mathbf{b})$ , without referring explicitly to internal model parameters.

<sup>10</sup>Some readers may notice the similarity between Eq. 26 and other penalty-based criteria, such as BIC [49] and AIC [50]. Although all these criteria share the same overall interpretation, BIC and AIC rely on specific assumptions about the asymptotic shape of the model likelihood, which are known to be invalid for the SBM [51], unlike Eq. 26 which is exact.

<sup>11</sup>An important result in information theory states that compressing random data is asymptotically impossible [52]. This lies at the heart of the effectiveness of the MDL approach in preventing overfitting, as incorporating randomness into the model description cannot be used to reduce the description length.

<sup>12</sup>Care should be taken when comparing with “known” divisions in this manner, as there is no guarantee that the available metadata is in fact relevant for the network structure. See Refs. [56–58] for more detailed discussions.

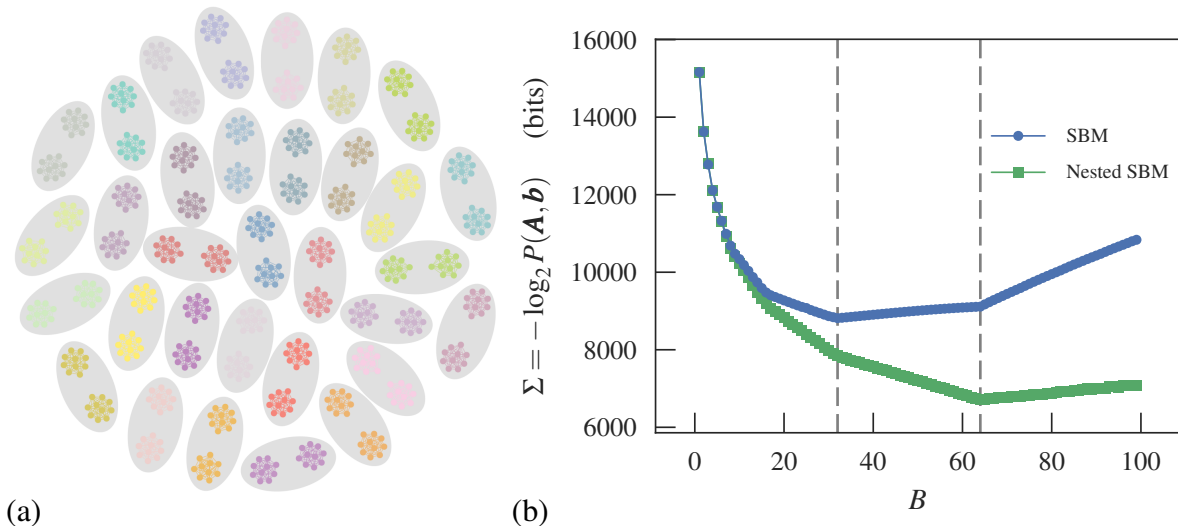


Figure 4. Inference of the SBM on a simple artificial network composed of 64 cliques of size 10, illustrating the underfitting problem: (a) The partition that maximizes the posterior probability of Eq. 9, or equivalently, minimizes the description length of Eq. 24. The 64 cliques are grouped into 32 groups composed of two cliques each. (b) Minimum description length as a function of the number of groups of the corresponding partition, both for the SBM and its nested variant, which is less susceptible to underfitting, and puts all 64 cliques in their own groups.

both too simple and too complex models, with a global minimum at  $B = 10$  — corresponding to the most compressive partition. Importantly, if we now *randomize* the network, by placing all its edges in a completely random fashion, we obtain instead a trivial partition into  $B = 1$  group — indicating that the best model for this data is indeed a fully random graph. Hence, we see that this approach completely avoids the pitfall discussed in Sec. II and does not identify groups in fully random networks, and that the division shown in Fig. 3a points to a statistically significant structure in the data, that cannot be explained simply by random fluctuations.

## VI. THE “RESOLUTION LIMIT” UNDERFITTING PROBLEM, AND THE NESTED SBM

Although the Bayesian approach outlined above is in general protected against overfitting, it is still susceptible to *underfitting*, i.e. when we mistake statistically significant structure for randomness, resulting in the inference of an overly simplistic model. This happens whenever there is a large discrepancy between our prior assumptions and what is observed in the data. We illustrate this problem with a simple example: Consider a network formed of 64 isolated cliques of size 10, as shown in Fig. 4a. If we employ the approach described in the previous section, and maximize the posterior of Eq. 9, we obtain a partition into  $B = 32$  groups, where each group is composed of two cliques. This is a fairly unsatisfying characterization of this network, and also somewhat perplexing, since the probability that the inferred SBM will generate the observed network — i.e. each of the 32 groups will simultaneously and spontaneously split in two disjoint cliques — is vanishingly small. Indeed, intuitively it seems we should do significantly better with this rather obvious example, and that the best fit would be to put each of the cliques in their own group. In order to see what went wrong, we need to revisit our prior assumptions, in particular our choice for  $P(\boldsymbol{\lambda}|\mathbf{b})$  in Eq. 19, or equivalently, our choice of  $P(\mathbf{e}|\mathbf{b})$  in Eq. 23 for the microcanonical for-

mulation. In both cases, they correspond to uninformative priors, which put approximately equal weight on all allowed types of large-scale structures. As argued before, this seems reasonable at first, since we should not bias our model before we observe the data. However, the implication of this choice is that we expect *a priori* the structure of the network at the aggregate group level, i.e. considering only the groups and the edges between them (not the individual nodes), to be fully random. This is indeed not the case in the simple example of Fig. 4, and in fact it is unlikely to be the case for most networks that we encounter, which will probably be structured at a higher level as well. The unfavorable outcome of the uninformative assumption can also be seen by inspecting its effect on the description length of Eq. 24. If we revisit our simple model with  $C$  cliques of size  $m$ , grouped uniformly into  $B$  groups of size  $C/B$ , and we assume that these values are sufficiently large so that Stirling's factorial approximation  $\ln x! \approx x \ln x - x$  can be used, the description length becomes

$$\Sigma \approx -(E - N) \log_2 B + \frac{B(B+1)}{2} \log_2 E, \quad (27)$$

where  $N = Cm$  is the total number of nodes and  $E = C \binom{m}{2}$  is the total number of edges, and we have omitted terms that do not depend on  $B$ . From this, we see that if we increase the number of groups  $B$ , this incurs a quadratic penalty in the description length given by the second term of Eq. 27, which originates precisely from our expression of  $P(\mathbf{e}|\mathbf{b})$ : It corresponds to the amount of information necessary to describe all entries of a symmetric  $B \times B$  matrix that takes independent values between 0 and  $E$ . Indeed, a slightly more careful analysis of the scaling of the description length [19, 29] reveals that this approach is unable to uncover a number of groups that is larger than  $B_{\max} \propto \sqrt{N}$ , even if their existence is obvious, as in our example of Fig. 4.<sup>13</sup>

Trying to avoid this limitation might seem like a conundrum, since replacing the uninformative prior for  $P(\mathbf{e}|\mathbf{b})$  amounts to making a more definite statement on the most likely large-scale structures that we expect to find, which we might hesitate to stipulate, as this is precisely what we want to discover from the data in the first place, and we want to remain unbiased. Luckily, there is in fact a general approach available to us to deal with this problem: We postpone our decision about the higher-order aspects of the model until we observe the data. In fact, we already saw this approach in action when we decided on the prior for the partitions; We do so by replacing the uninformative prior with a *parametric* distribution, whose parameters are in turn modelled by a another distribution, i.e. a *hyperprior*. The parameters of the prior then become *latent variables* that are learned from data, allowing us to uncover further structures, while remaining unbiased.

The microcanonical formulation allows us to proceed in this direction in a straightforward manner, as we can interpret the matrix of edge counts  $\mathbf{e}$  as the adjacency matrix of a multigraph where each of the groups is represented as a single node. Within this interpretation, an elegant solution presents itself, where we describe the matrix  $\mathbf{e}$  with *another* SBM, i.e. we partition each of the groups into meta-groups, and the edges between groups are placed according to the edge counts between meta-groups. For this second SBM, we can proceed in the same manner, and model it by a third SBM, and so on, forming a nested hierarchy, as illustrated in Fig. 5 [61]. More precisely, if we denote by  $B_l$ ,  $\mathbf{b}_l$  and  $\mathbf{e}_l$  the number of groups, the partition and the matrix of edge counts at level  $l \in \{0, \dots, L\}$ , we have

$$P(\mathbf{e}_l | \mathbf{b}_{l-1}, \mathbf{e}_{l+1}, \mathbf{b}_l) = \prod_{r < s} \left( \binom{n_r^l n_s^l}{e_{rs}^{l+1}} \right)^{-1} \prod_r \left( \binom{n_r^l (n_r^l + 1) / 2}{e_{rs}^{l+1} / 2} \right)^{-1}, \quad (28)$$

<sup>13</sup>This same problem occurs for slight variations of the SBM and corresponding priors, provided they are uninformative, such as those in Refs. [30, 31, 33], and also with other penalty based approaches that rely on a functional form similar to Eq. 27 [59]. Furthermore, this limitation is conspicuously similar to the “resolution limit” present in the popular heuristic of modularity maximization [60], although it is not yet clear if a deeper connection exists between both phenomena.

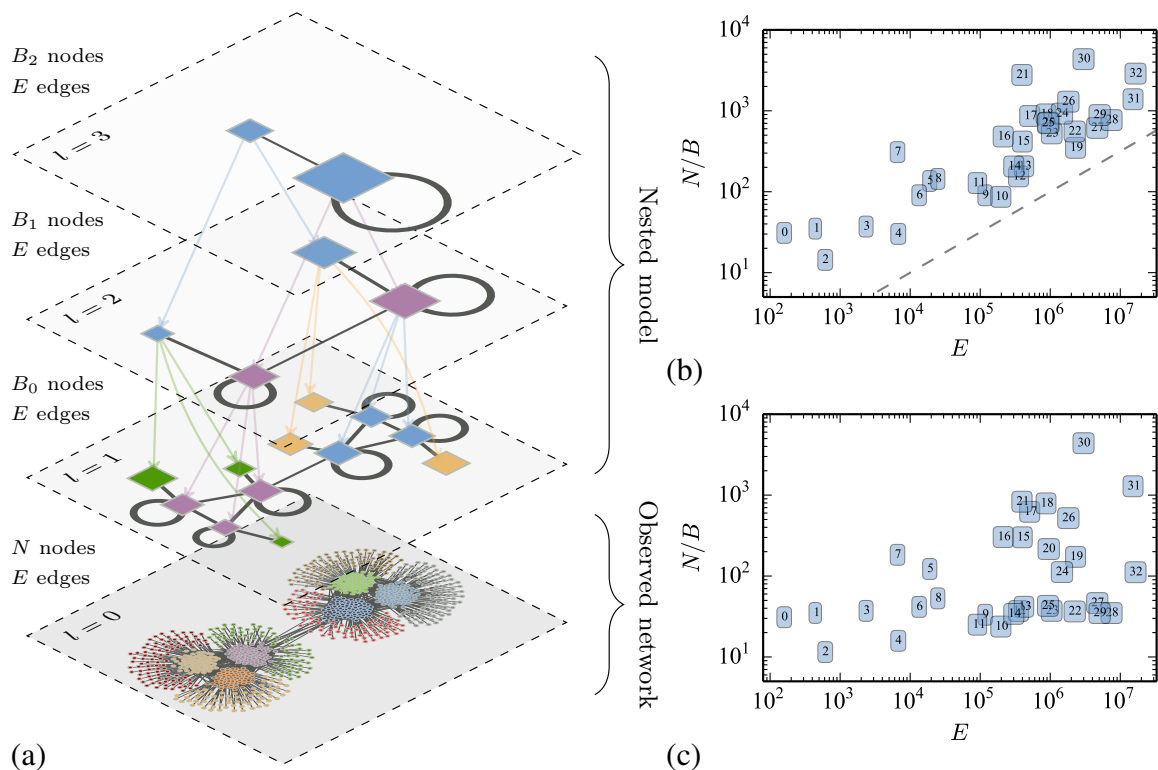


Figure 5. (a) Diagrammatic representation of the nested SBM described in the text, with  $L = 3$  levels, adapted from Ref. [61]. (b) Average group sizes  $N/B$  obtained with the SBM using uninformative priors, for a variety of empirical networks, listed in Ref. [61]. The dashed line shows a slope  $\sqrt{E}$ , highlighting the systematic underfitting problem. (c) The same as in (b), but using the nested SBM, where the underfitting has virtually disappeared, with datasets randomly scattered in the allowed range.

with  $\binom{n}{m} = \binom{n+m-1}{m}$  counting the number of  $m$ -combinations with repetitions from a set of size  $n$ . Eq. 28 is the likelihood of a maximum-entropy multigraph SBM, i.e. every multigraph occurs with the same probability, provided they fulfill the imposed constraints<sup>14</sup> [17]. The prior for the partitions is again given by Eq. 17,

$$P(\mathbf{b}_l) = \frac{\prod_r n_r!}{B_{l-1}!} \left( \frac{B_{l-1} - 1}{B_{l-1}} \right)^{-1} B_{l-1}^{-1}, \quad (29)$$

with  $B_{-1} = N$ , so that the joint probability of the data, edge counts and the hierarchical partition  $\{\mathbf{b}_l\}$  becomes

$$P(\mathbf{A}, \{\mathbf{e}_l\}, \{\mathbf{b}_l\} | L) = P(\mathbf{A} | \mathbf{e}_1, \mathbf{b}_0) P(\mathbf{b}_0) \prod_{l=1}^L P(\mathbf{e}_l | \mathbf{b}_{l-1}, \mathbf{e}_{l+1}, \mathbf{b}_l) P(\mathbf{b}_l), \quad (30)$$

where we impose the boundary conditions  $B_L = 1$  and  $P(\mathbf{b}_L) = 1$ . We can treat the hierarchy depth  $L$  as a latent variable as well, by placing a prior on it  $P(L) = 1/L_{\max}$ , where  $L_{\max}$  is the maximum value allowed. But since this only contributes to an overall multiplicative constant, it

<sup>14</sup>Note that we cannot use in the upper levels exactly the same model we use in the bottom level, given by Eq. 22, as most terms in the subsequent levels will cancel out. This happens because the model in Eq. 22 is based on the uniform generation of configurations, not multigraphs [19]. However, we are free to use Eq. 28 in the bottom level as well.

has no effect on the posterior distribution, and thus can be omitted. If we impose  $L = 1$ , we recover the uninformative prior for  $\mathbf{e} = \mathbf{e}_1$ ,

$$P(\mathbf{e}|\mathbf{b}_0) = \left( \binom{B(B+1)/2}{E} \right)^{-1}, \quad (31)$$

which is different from Eq. 23 only in that the number of edges  $E$  is not allowed to fluctuate.<sup>15</sup> The inference of this model is done in the same manner as the uninformative one, by obtaining the posterior distribution of the hierarchical partition

$$P(\{\mathbf{b}_l\}|\mathbf{A}) = \frac{P(\mathbf{A}, \{\mathbf{b}_l\})}{P(\mathbf{A})} = \frac{P(\mathbf{A}, \{\mathbf{e}_l\}, \{\mathbf{b}_l\})}{P(\mathbf{A})}, \quad (32)$$

and the description length is given analogously by

$$\Sigma = -\log_2 P(\mathbf{A}|\{\mathbf{e}_l\}, \{\mathbf{b}_l\}) - \log_2 P(\{\mathbf{e}_l\}, \{\mathbf{b}_l\}). \quad (33)$$

This approach has a series of advantages; in particular, we remain *a priori* agnostic with respect to what kind of large-scale structure is present in the network, having constrained ourselves simply in that it can be represented as a SBM at a higher level, and with the uninformative prior as a special case. Despite this, we are able to overcome the underfitting problem encountered with the uninformative approach: If we apply this model to the example of Fig. 4, we can successfully distinguish all 64 cliques, and provide a lower overall description length for the data, as can be seen in Fig. 4b. More generally, by investigating the properties of the model likelihood, it is possible to show that the maximum number of groups that can be uncovered with this model scales as  $B_{\max} \propto N/\log N$ , which is significantly larger than the limit with uninformative priors [19, 61]. The difference between both approaches manifests itself very often in practice, as shown in Fig. 5b, where systematic underfitting is observed for a wide variety of network datasets, which disappears with the nested model, as seen in Fig. 5c. Crucially, we achieve this decreased tendency to underfit without sacrificing our protection against overfitting: Despite the more elaborate model specification, the inference of the nested SBM is completely nonparametric, and the same Bayesian and information-theoretical principles still hold. Furthermore, as we already mentioned, the uninformative case is a special case of the nested SBM, i.e. when  $L = 1$ , and hence it can only improve the inference (e.g. by reducing the description length), with no drawbacks. We stress that the number of hierarchy levels, as with any other dimension of the model, such as the number of groups in each level, is inferred from data, and does not need to be determined a priori.

In addition to the above, the nested model also gives us the capacity of describing the data at multiple scales, which could potentially exhibit different mixing patterns. This is particularly useful for large networks, where the SBM might still give us a very complex description, which becomes easier to interpret if we concentrate first on the upper levels of the hierarchy. A good example is the result obtained for the internet topology at the autonomous systems level, shown in Fig. 6. The lowest level of the hierarchy shows a division into a large number of groups, with a fairly complicated structure, whereas the higher levels show an increasingly simplified picture, culminating in a core-periphery organization as the dominating pattern.

<sup>15</sup>The prior of Eq. 31 and the hierarchy in Eq. 30 are conditioned on the total number of edges  $E$ , which is typically unknown before we observe the data. Similarly to the parameter  $\bar{\lambda}$  in the canonical model formulation, the strictly correct approach would be to consider this quantity as an additional model parameter, with its prior distribution  $P(E)$ . However, in the microcanonical model there is no integration involved, and  $P(E)$  — regardless of how we specify it — would contribute to an overall multiplicative constant that disappears from the posterior distribution after normalization. Therefore we can simply omit it.



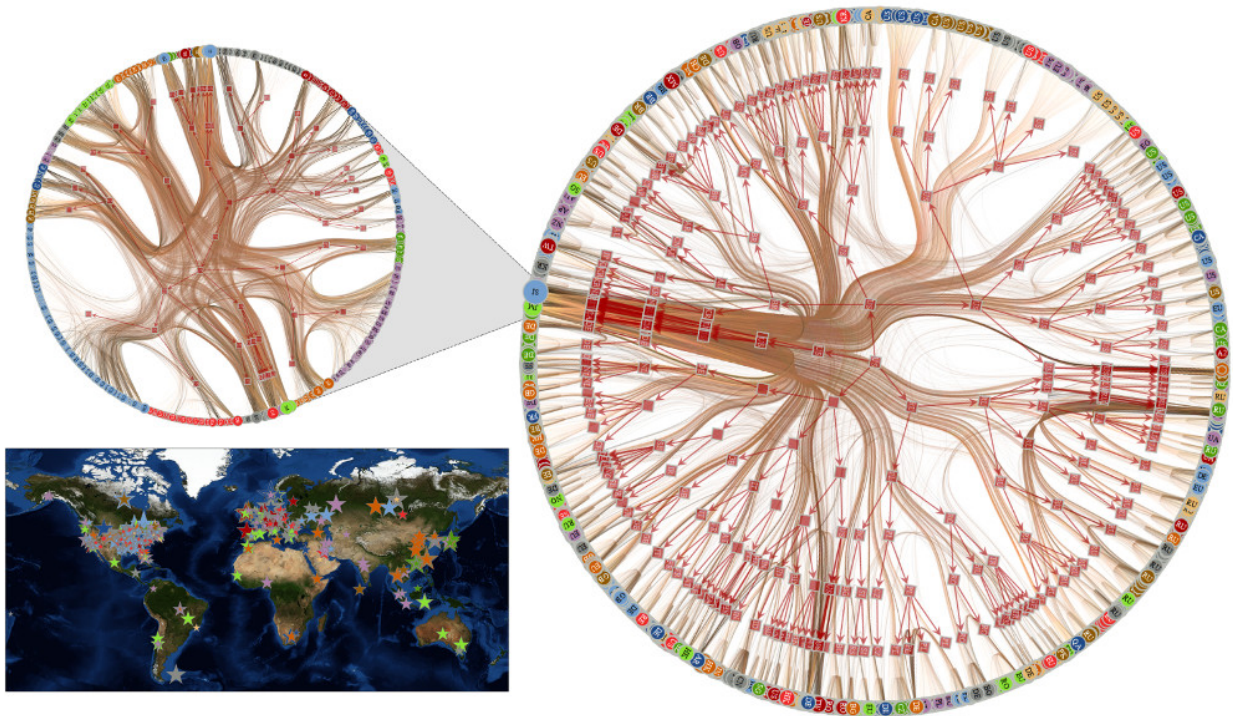


Figure 6. Fit of the (degree-corrected) nested SBM for the internet topology at the autonomous systems level, adapted from Ref. [61]. The hierarchical division reveals a core-periphery organization at the higher levels, where most routes go through a relatively small number of nodes (shown in the inset and in the map). The lower levels reveal a more detailed picture, where a large number of groups of nodes are identified according to their routing patterns (amounting largely to distinct geographical regions). The layout is obtained with an edge bundling algorithm by Holten [62], which uses the hierarchical partition to route the edges.

## VII. MODEL VARIATIONS

Varying the number of groups and building hierarchies are not the only ways we have of adapting the complexity of the model to the data. We may also change the internal structure of the model, and how the division into groups affect the placement of edges. In fact, the basic ansatz of the SBM is very versatile, and many variations have been proposed in the literature. In this section we review two important ones — SBMs with degree correction and group overlap — and review other model flavors in a summarized manner.

Before we go further into the model variations, we point out that the multiplicity of models is a strength of the inference approach. This is different from the broader field of network clustering, where a large number of available algorithms often yield conflicting results for the same data, leaving practitioners lost in how to select between them [63, 64]. Instead, within the inference framework we can in fact compare different models in a principled manner and select the best one according to the statistical evidence available. We proceed with a general outline of the model selection procedure, before following with specific model variations.

### A. Model selection

Suppose we define two versions of the SBM, labeled  $\mathcal{C}_1$  and  $\mathcal{C}_2$ , each with their own posterior distribution of partitions,  $P(\mathbf{b}|\mathbf{A}, \mathcal{C}_1)$  and  $P(\mathbf{b}|\mathbf{A}, \mathcal{C}_2)$ . Suppose we find the most likely partitions  $\mathbf{b}_1$  and  $\mathbf{b}_2$ , according to  $\mathcal{C}_1$  and  $\mathcal{C}_2$ , respectively. How do we decide which partition is more representative of the data? The consistent approach is to obtain the so-called posterior odds ratio [3, 65]

$$\Lambda = \frac{P(\mathbf{b}_1, \mathcal{C}_1|\mathbf{A})}{P(\mathbf{b}_2, \mathcal{C}_2|\mathbf{A})} = \frac{P(\mathbf{A}|\mathbf{b}_1, \mathcal{C}_1)P(\mathbf{b}_1)P(\mathcal{C}_1)}{P(\mathbf{A}|\mathbf{b}_2, \mathcal{C}_2)P(\mathbf{b}_2)P(\mathcal{C}_2)}, \quad (34)$$

where  $P(\mathcal{C})$  is our prior belief that variant  $\mathcal{C}$  is valid. A value of  $\Lambda > 1$  indicates that the choice  $(\mathbf{b}_1, \mathcal{C}_1)$  is  $\Lambda$  times more plausible as an explanation for the data than the alternative,  $(\mathbf{b}_2, \mathcal{C}_2)$ . If we are *a priori* agnostic with respect to which model flavor is best, i.e.  $P(\mathcal{C}_1) = P(\mathcal{C}_2)$ , we have then

$$\Lambda = \frac{P(\mathbf{A}|\mathbf{b}_1, \mathcal{C}_1)P(\mathbf{b}_1)}{P(\mathbf{A}|\mathbf{b}_2, \mathcal{C}_2)P(\mathbf{b}_2)} = 2^{-\Delta\Sigma}, \quad (35)$$

where  $\Delta\Sigma = \Sigma_1 - \Sigma_2$  is the description length difference between both choices. Hence, we should generally prefer the model choice that is most compressive, i.e. with the smallest description length. However, if the value of  $\Lambda$  is close to 1, we should refrain from forcefully rejecting the alternative, as the evidence in data would not be strongly decisive either way. I.e. the actual value of  $\Lambda$  gives us the confidence with which we can choose the preferred model. The final decision, however, is subjective, since it depends on what we might consider plausible. A value of  $\Lambda = 2$ , for example, typically cannot be used to forcefully reject the alternative hypothesis, whereas a value of  $\Lambda = 10^{100}$  might.

An alternative test we can make is to decide which model *class* is most representative of the data, when averaged over all possible partitions. In this case, we proceed in a an analogous way by computing the posterior odds ratio

$$\Lambda' = \frac{P(\mathcal{C}_1|\mathbf{A})}{P(\mathcal{C}_2|\mathbf{A})} = \frac{P(\mathbf{A}|\mathcal{C}_1)P(\mathcal{C}_1)}{P(\mathbf{A}|\mathcal{C}_2)P(\mathcal{C}_2)}, \quad (36)$$

where

$$P(\mathbf{A}|\mathcal{C}) = \sum_{\mathbf{b}} P(\mathbf{A}|\mathbf{b}, \mathcal{C})P(\mathbf{b}) \quad (37)$$

is the model evidence. When  $P(\mathcal{C}_1) = P(\mathcal{C}_2)$ ,  $\Lambda'$  is called the *Bayes factor*, with an interpretation analogous to  $\Lambda$  above, but where the statement is made with respect to all possible partitions, not only the most likely one. Unfortunately, as mentioned previously, the evidence  $P(\mathbf{A}|\mathcal{C})$  cannot be computed exactly for the models we are interested in, making this criterion more difficult to employ in practice (although approximations have been proposed, see e.g. Ref [19]). We return to the issue of when it should we optimize or sample from the posterior distribution in Sec. IX, and hence which of the two criteria should be used.

### B. Degree correction

The underlying assumption of all variants of the SBM considered so far is that nodes that belong to the same group are statistically equivalent. As it turns out, this fundamental aspect results in a very unrealistic property. Namely, this generative process implies that all nodes that belong to the

same group receive on average the same number of edges. However, a common property of many empirical networks is that they have very heterogeneous degrees, often broadly distributed over several orders of magnitudes [15]. Therefore, in order for this property to be reproduced by the SBM, it is necessary to group nodes according to their degree, which may lead to some seemingly odd results. An example of this was given in Ref. [16] and is shown in Fig. 7a. It corresponds to a fit of the SBM to a network of political blogs recorded during the 2004 American presidential election campaign [66], where an edge exists between two blogs if one links to the other. If we guide ourselves by the layout of the figure, we identify two assortative groups, which happen to be those aligned with the Republican and Democratic parties. However, inside each group there is a significant variation in degree, with a few nodes with many connections and many with very few. Because of what just has been explained, if we perform a fit of the SBM using only  $B = 2$  groups, it prefers to cluster the nodes into high-degree and low-degree groups, completely ignoring the party alliance.<sup>16</sup> Arguably, this is a bad fit of this network, since — similarly to the underfitting example of Fig. 4 — the probability of the fitted SBM generating a network with such a party structure is vanishingly small. In order to solve this undesired behavior, Karrer and Newman [16] proposed a modified model, which they dubbed the degree-corrected SBM (DC-SBM). In this variation, each node  $i$  is attributed with a parameter  $\theta_i$  that controls its expected degree, independently of its group membership. Given this extra set of parameters, a network is generated with probability

$$P(\mathbf{A}|\boldsymbol{\lambda}, \boldsymbol{\theta}, \mathbf{b}) = \prod_{i < j} \frac{e^{-\theta_i \theta_j \lambda_{b_i, b_j}} (\theta_i \theta_j \lambda_{b_i, b_j})^{A_{ij}}}{A_{ij}!} \times \prod_i \frac{e^{-\theta_i^2 \lambda_{b_i, b_i} / 2} (\theta_i^2 \lambda_{b_i, b_i} / 2)^{A_{ii} / 2}}{(A_{ii} / 2)!}, \quad (38)$$

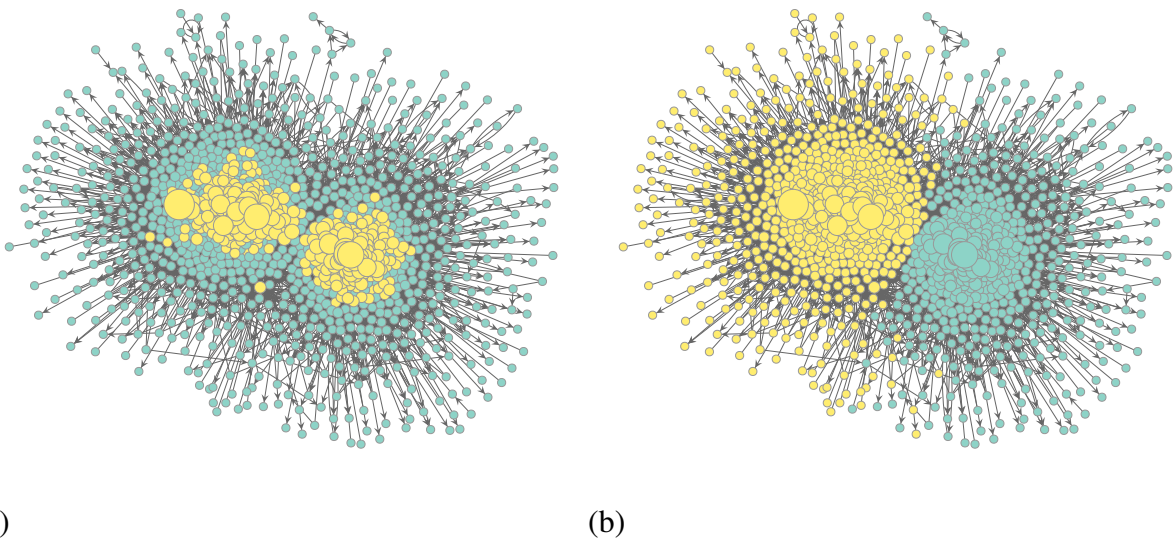


Figure 7. Inferred partition for a network of political blogs [66] using (a) the SBM and (b) the DC-SBM, in both cases forcing  $B = 2$  groups. The node sizes are proportional to the node degrees. The SBM divides the network into low and high-degree groups, whereas the DC-SBM prefers the division into political factions.

<sup>16</sup>It is possible that unexpected results of this kind inhibited the initial adoption of SBM methods in the network science community, which focused instead on more heuristic community detection methods, save for a few exceptions (e.g. [20, 22, 23, 25, 67, 68]).

where  $\lambda_{rs}$  again controls the expected number of edges between groups  $r$  and  $s$ . Note that since the parameters  $\lambda_{rs}$  and  $\theta_i$  always appear multiplying each other in the likelihood, their individual values may be arbitrarily scaled, provided their products remain the same. If we choose the parametrization  $\sum_i \theta_i \delta_{b_i,r} = 1$  for every group  $r$ , then they acquire a simple interpretation:  $\lambda_{rs}$  is the expected number of edges between groups  $r$  and  $s$ ,  $\lambda_{rs} = \langle e_{rs} \rangle$ , and  $\theta_i$  is proportional to the expected degree of node  $i$ ,  $\theta_i = \langle k_i \rangle / \sum_s \lambda_{b_i,s}$ .

When inferring this model from the political blogs data — again forcing  $B = 2$  — we obtain a much more satisfying result, where the two political factions are neatly identified, as seen in Fig. 7b. As this model is capable of fully decoupling the community structure from the degrees, which are captured separately by the parameters  $\boldsymbol{\lambda}$  and  $\boldsymbol{\theta}$ , respectively, the degree heterogeneity of the network does not interfere with the identification of the political factions.

Based on the above example, and on the knowledge that most networks possess heterogeneous degrees, we could expect the DC-SBM to provide a better fit for most of them. However, before we jump to this conclusion, we must first acknowledge that the seemingly increased quality of fit obtained with the SBM came at the expense of adding an extra set of parameters,  $\boldsymbol{\theta}$  [51]. However intuitive we might judge the improvement brought on by degree correction, simply adding more parameters to a model is an almost sure recipe for overfitting. Therefore, a more prudent approach is once more to frame the inference problem in a Bayesian way, by focusing on the posterior distribution  $P(\mathbf{b}|\mathbf{A})$ , and on the description length. For this, we must include a prior for the node propensities  $\boldsymbol{\theta}$ . The uninformative choice is the one which ascribes the same probability to all possible choices,

$$P(\boldsymbol{\theta}|\mathbf{b}) = \prod_r (n_r - 1)! \delta(\sum_i \theta_i \delta_{b_i,r} - 1). \quad (39)$$

Using again an uninformative prior for  $\boldsymbol{\lambda}$ ,

$$P(\boldsymbol{\lambda}|\mathbf{b}) = \prod_{r \leq s} e^{-\lambda_{rs}/(1+\delta_{rs})\bar{\lambda}} / (1 + \delta_{rs})\bar{\lambda} \quad (40)$$

with  $\bar{\lambda} = 2E/B(B+1)$ , the marginal likelihood now becomes

$$\begin{aligned} P(\mathbf{A}|\mathbf{b}) &= \int P(\mathbf{A}|\boldsymbol{\lambda}, \boldsymbol{\theta}, \mathbf{b}) P(\boldsymbol{\lambda}|\mathbf{b}) P(\boldsymbol{\theta}|\mathbf{b}) d\boldsymbol{\lambda} d\boldsymbol{\theta} \\ &= \frac{\bar{\lambda}^E}{(\bar{\lambda} + 1)^{E+B(B+1)/2}} \times \frac{\prod_{r < s} e_{rs}! \prod_r e_{rr}!!}{\prod_{i < j} A_{ij}! \prod_i A_{ii}!!} \times \prod_r \frac{(n_r - 1)!}{(e_r + n_r - 1)!} \times \prod_i k_i!, \end{aligned} \quad (41)$$

where  $k_i = \sum_j A_{ij}$  is the degree of node  $i$ , which can be used in the same way to obtain a posterior for  $\mathbf{b}$ , via Eq. 9. Once more, the model above is equivalent to a microcanonical formulation [19], given by

$$P(\mathbf{A}|\mathbf{b}) = P(\mathbf{A}|\mathbf{k}, \mathbf{e}, \mathbf{b}) P(\mathbf{k}|\mathbf{e}, \mathbf{b}) P(\mathbf{e}|\mathbf{b}), \quad (42)$$

with

$$P(\mathbf{A}|\mathbf{k}, \mathbf{e}, \mathbf{b}) = \frac{\prod_{r < s} e_{rs}! \prod_r e_{rr}!! \prod_i k_i!}{\prod_{i < j} A_{ij}! \prod_i A_{ii}!! \prod_r e_r!!}, \quad (43)$$

$$P(\mathbf{k}|\mathbf{e}, \mathbf{b}) = \prod_r \left( \binom{n_r}{e_r} \right)^{-1}, \quad (44)$$

and  $P(\mathbf{e}|\mathbf{b})$  given by Eq. 23. In the model above,  $P(\mathbf{A}|\mathbf{k}, \mathbf{e}, \mathbf{b})$  is the probability of generating a

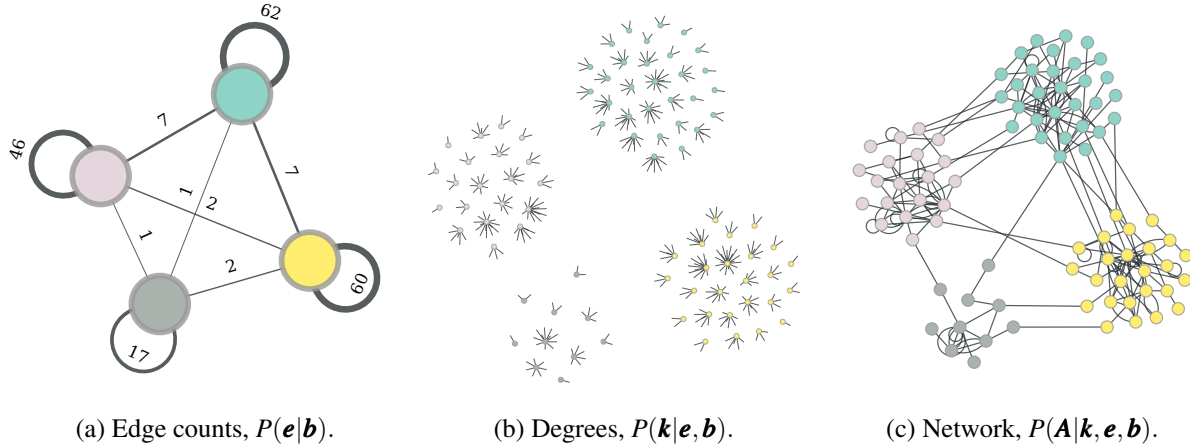


Figure 8. Illustration of the generative process of the microcanonical DC-SBM. Given a partition of the nodes, the edge counts between groups are sampled (a), followed by the degrees of the nodes (b) and finally the network itself (c). Adapted from Ref. [19].

multigraph where the edge counts between groups *as well as* the degrees  $\mathbf{k}$  are fixed to specific values.<sup>17</sup> The prior  $P(\mathbf{k}|\mathbf{e}, \mathbf{b})$  is the uniform probability of generating a degree sequence, where all possibilities that satisfy the constraints imposed by the edge counts  $\mathbf{e}$ , namely  $\sum_i k_i \delta_{b_i, r} = e_r$ , occur with the same probability. The description length of this model is then given by

$$\Sigma = -\log_2 P(\mathbf{A}, \mathbf{b}) = -\log_2 P(\mathbf{A}|\mathbf{k}, \mathbf{e}, \mathbf{b}) - \log_2 P(\mathbf{k}, \mathbf{e}, \mathbf{b}). \quad (45)$$

Because uninformative priors were used to derive the above equations, we are once more subject to the same underfitting problem described previously. Luckily, from the microcanonical model we can again derive a nested DC-SBM, by replacing  $P(\mathbf{e})$  by a nested sequence of SBMs, exactly in the same way as was done before [19, 61]. We also have the opportunity of replacing the uninformative prior for the degrees in Eq. 44 with a more realistic option. As was argued in Ref. [19], degree sequences generated by Eq. 44 result in exponential degree distributions, which are not quite as heterogeneous as what is often encountered in practice. A more refined approach, which is already familiar to us at this point, is to increase the Bayesian hierarchy, and choose a prior that is conditioned on a higher-order aspect of the data, in this case the *frequency* of degrees, i.e.

$$P(\mathbf{k}|\mathbf{e}, \mathbf{b}) = P(\mathbf{k}|\mathbf{e}, \mathbf{b}, \boldsymbol{\eta})P(\boldsymbol{\eta}|\mathbf{e}, \mathbf{b}), \quad (46)$$

where  $\boldsymbol{\eta} = \{\eta_k^r\}$ , with  $\eta_k^r$  being the number of nodes of degree  $k$  in group  $r$ . In the above,  $P(\boldsymbol{\eta}|\mathbf{e}, \mathbf{b})$  is a uniform distribution of frequencies, and  $P(\mathbf{k}|\mathbf{e}, \mathbf{b}, \boldsymbol{\eta})$  generates the degrees according to the sampled frequencies (we omit the respective expressions for brevity, and refer to Ref. [19] instead). Thus, this model is capable of using regularities in the degree distribution to inform the division into groups, and is generally capable of better fits than the uniform model of Eq. 44.

If we apply this nonparametric approach to the same political blog network of Ref. 9, we find a much more detailed picture of its structure, revealing many more than two groups, as shown in Fig. 9, for three model variants: the nested SBM, the nested DC-SBM and the nested DC-SBM

<sup>17</sup>The ensemble equivalence of Eq. 42 is in some ways more remarkable than for the traditional SBM. This is because a direct equivalence between the ensembles of Eqs. 38 and 43 is not satisfied even in the asymptotic limit of large networks [17, 69], which does happen for Eqs. 8 and 22. Equivalence is observed only if the individual degrees  $k_i$  also become asymptotically large. However, when the parameters  $\boldsymbol{\lambda}$  and  $\boldsymbol{\theta}$  are integrated out, the equivalence becomes exact for networks of any size.

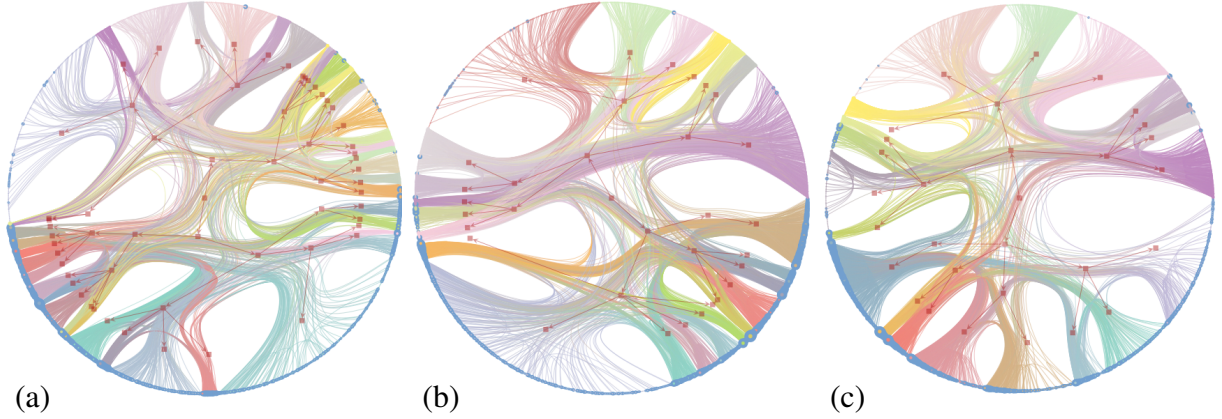


Figure 9. Most likely hierarchical partitions of a network of political blogs [66], according to the three model variants considered, as well as the inferred number of groups  $B_1$  at the bottom of the hierarchy, and the description length  $\Sigma$ : (a) NDC-SBM,  $B_1 = 42$ ,  $\Sigma \approx 89,938$  bits, (b) DC-SBM,  $B_1 = 23$ ,  $\Sigma \approx 87,162$  bits, (c) DC-SBM with the degree prior of Eq. 46,  $B_1 = 20$ ,  $\Sigma \approx 84,890$  bits. The nodes circled in blue were classified as “liberals” and the remaining ones as “conservatives” in Ref. [66] based on the blog contents. Adapted from Ref. [19].

with the degree prior of Eq. 46. All three model variants are in fact capable of identifying the same Republican/Democrat division at the topmost hierarchical level — showing that the non-degree-corrected SBM is not as inept in capturing this aspect of the data as the result obtained by forcing  $B = 2$  might suggest. However the internal divisions of both factions that they uncover are very distinct from each other. If we inspect the obtained values of the description length with each model we see that the DC-SBM (in particular when using Eq. 46) results in a smaller value, indicating that it better captures the structure of the data, despite the increased number of parameters. Indeed, a systematic analysis carried out in Ref. [19] showed that the DC-SBM does in fact yield shorter description lengths for a majority of empirical datasets, thus ultimately confirming the original intuition behind the model formulation.

### C. Group overlaps

Another way we can change the internal structure of the model is to allow the groups to overlap, i.e. we allow a node to belong to more than one group at the same time. The connection patterns of the nodes are then assumed to be a mixture of the “pure” groups, which results in a richer type of model [24]. Following Ball et al. [70], we can adapt the Poisson formulation to overlapping SBMs in a straightforward manner,

$$P(\mathbf{A}|\boldsymbol{\kappa}, \boldsymbol{\lambda}) = \prod_{i<j} \frac{e^{-\lambda_{ij}} \lambda_{ij}^{A_{ij}}}{A_{ij}!} \prod_i \frac{e^{-\lambda_{ii}/2} (\lambda_{ii}/2)^{A_{ii}/2}}{A_{ii}/2!}, \quad (47)$$

with

$$\lambda_{ij} = \sum_{rs} \kappa_{ir} \lambda_{rs} \kappa_{js}, \quad (48)$$

where  $\kappa_{ir}$  is the probability with which node  $i$  is chosen from group  $r$ , so that  $\sum_i \kappa_{ir} = 1$ , and  $\lambda_{rs}$  is once more the expected number of edges between groups  $r$  and  $s$ . The parameters  $\boldsymbol{\kappa}$  re-

place the disjoint partition  $\mathbf{b}$  we have been using so far by a “soft” clustering into overlapping categories<sup>18</sup>. Note, however, that this model is a direct generalization of the non-overlapping DC-SBM of Eq. 38, which is recovered simply by choosing  $\kappa_{ir} = \theta_i \delta_{r,b_i}$ . The Bayesian formulation can also be performed by using an uninformative prior for  $\boldsymbol{\kappa}$ ,

$$P(\boldsymbol{\kappa}) = \prod_r (n-1)! \delta(\sum_i \kappa_{ir} - 1), \quad (49)$$

in addition to the same prior for  $\boldsymbol{\lambda}$  in Eq. 40. Unfortunately, computing the marginal likelihood using Eq. 47 directly,

$$P(\mathbf{A}|\boldsymbol{\kappa}) = \int P(\mathbf{A}|\boldsymbol{\kappa}, \boldsymbol{\lambda}) P(\boldsymbol{\lambda}) d\boldsymbol{\lambda}, \quad (50)$$

is not tractable, which prevents us from obtaining the posterior  $P(\boldsymbol{\kappa}|\mathbf{A})$ . Instead, it is more useful to consider the auxiliary labelled matrix, or tensor,  $\mathbf{G} = \{G_{ij}^{rs}\}$ , where  $G_{ij}^{rs}$  is a particular decomposition of  $A_{ij}$  where the two edge endpoints — or “half-edges” — of an edge  $(i, j)$  are labelled with groups  $(r, s)$ , such that

$$A_{ij} = \sum_{rs} G_{ij}^{rs}. \quad (51)$$

Since a sum of Poisson variables is also distributed according to a Poisson, we can write Eq. 47 as

$$P(\mathbf{A}|\boldsymbol{\kappa}, \boldsymbol{\lambda}) = \sum_{\mathbf{G}} P(\mathbf{G}|\boldsymbol{\kappa}, \boldsymbol{\lambda}) \prod_{i \leq j} \delta_{\sum_{rs} G_{ij}^{rs}, A_{ij}}, \quad (52)$$

with each half-edge labelling being generated by

$$P(\mathbf{G}|\boldsymbol{\kappa}, \boldsymbol{\lambda}) = \prod_{i < j} \prod_{rs} \frac{e^{-\kappa_{ir} \lambda_{rs} \kappa_{js}} (\kappa_{ir} \lambda_{rs} \kappa_{js})^{G_{ij}^{rs}}}{G_{ij}^{rs}!} \times \prod_i \prod_{rs} \frac{e^{-\kappa_{ir} \lambda_{rs} \kappa_{is}/2} (\kappa_{is} \lambda_{rs} \kappa_{is}/2)^{G_{ii}^{rs}/2}}{(G_{ii}^{rs}/2)!}. \quad (53)$$

We can now compute the marginal likelihood as

$$\begin{aligned} P(\mathbf{G}) &= \int P(\mathbf{G}|\boldsymbol{\kappa}, \boldsymbol{\lambda}) P(\boldsymbol{\kappa}) P(\boldsymbol{\lambda}|\bar{\lambda}) d\boldsymbol{\kappa} d\boldsymbol{\lambda}, \\ &= \frac{\bar{\lambda}^E}{(\bar{\lambda} + 1)^{E+B(B+1)/2}} \frac{\prod_{r < s} e_{rs}! \prod_r e_{rr}!!}{\prod_{rs} \prod_{i < j} G_{ij}^{rs}! \prod_i G_{ii}^{rs}!!} \times \prod_r \frac{(N-1)!}{(e_r + N - 1)!} \times \prod_{ir} k_i^r!, \end{aligned} \quad (54)$$

which is very similar to Eq. 41 for the DC-SBM. With the above, and knowing from Eq. 51 that there is only one choice of  $\mathbf{A}$  that is compatible with any given  $\mathbf{G}$ , i.e.

$$P(\mathbf{A}|\mathbf{G}) = \prod_{i \leq j} \delta_{\sum_{rs} G_{ij}^{rs}, A_{ij}}, \quad (55)$$

we can sample from (or maximize) the posterior distribution of the half-edge labels  $\mathbf{G}$ , just like we did for the node partition  $\mathbf{b}$  in the nonoverlapping models,

$$P(\mathbf{G}|\mathbf{A}) = \frac{P(\mathbf{A}|\mathbf{G})P(\mathbf{G})}{P(\mathbf{A})} \propto P(\mathbf{G}) \times \prod_{i \leq j} \delta_{\sum_{rs} G_{ij}^{rs}, A_{ij}}, \quad (56)$$

<sup>18</sup>Note that, differently from the non-overlapping case, here it is possible for a node not to belong to any group, in which case it will never receive an incident edge.



Figure 10. Network of co-purchases of books about US politics [71], with groups inferred using (a) the non-overlapping DC-SBM, with description length  $\Sigma \approx 1,938$  bits, (b) the overlapping SBM with description length  $\Sigma \approx 1,931$  bits and (c) the overlapping SBM forcing only  $B = 2$  groups, with description length  $\Sigma \approx 1,946$  bits.

where the product in the last term only accounts for choices of  $\mathbf{G}$  which are compatible with  $\mathbf{A}$ , i.e. fulfill Eq. 51. Once more, the model of Eq. 54 is equivalent to its microcanonical analogue [18],

$$P(\mathbf{G}) = P(\mathbf{G}|\mathbf{k}, \mathbf{e})P(\mathbf{k}|\mathbf{e})P(\mathbf{e}), \quad (57)$$

where

$$P(\mathbf{G}|\mathbf{k}, \mathbf{e}) = \frac{\prod_{r<s} e_{rs}! \prod_r e_{rr}!! \prod_{ir} k_i^r!}{\prod_{rs} \prod_{i<j} G_{ij}^{rs}! \prod_i G_{ii}^{rs}!! \prod_r e_r!}, \quad (58)$$

$$P(\mathbf{k}|\mathbf{e}) = \prod_r \binom{e_r}{N}^{-1} \quad (59)$$

and  $P(\mathbf{e})$  given by Eq. 23. The variables  $\mathbf{k} = \{k_i^r\}$  are the labelled degrees of the labelled network  $\mathbf{G}$ , where  $k_i^r$  is the number of incident edges of type  $r$  a node  $i$  has. The description length becomes likewise

$$\Sigma = -\log_2 P(\mathbf{G}) = -\log_2 P(\mathbf{G}|\mathbf{k}, \mathbf{e}) - \log_2 P(\mathbf{k}|\mathbf{e}) - \log_2 P(\mathbf{e}). \quad (60)$$

The nested variant can be once more obtained by replacing  $P(\mathbf{e})$  in the same manner as before, and  $P(\mathbf{k}|\mathbf{e})$  in a manner that is conditioned on the labelled degree frequencies and degree of overlap, as described in detail in Ref. [18].

Equipped with this more general model, we may ask ourselves again if it provides a better fit of most networks, like we did for the DC-SBM in the previous section. Indeed, since the model is more general, we might conclude that this is an inevitability. However, this could be a fallacy, since more general models also include more parameters and hence are more likely to overfit. Indeed, previous claims about the existence of “pervasive overlap” in networks, based on nonstatistical methods [72], seemed to be based to some extent on this problematic logic. Claims about community overlaps are very different from, for example, the statement that networks possess heterogeneous degrees, since community overlap is not something that can be observed directly; instead it is something that must be *inferred*, which is precisely what our Bayesian approach is designed to do in a methodologically correct manner. An example of such a comparison is shown in Fig 10, for a small network of political books. This network, when analyzed using the nonoverlapping SBM, seems to be composed of three groups, easily interpreted as “left wing,” “right wing” and “center,” as the available metadata corroborates. If we fit the overlapping SBM, we observe



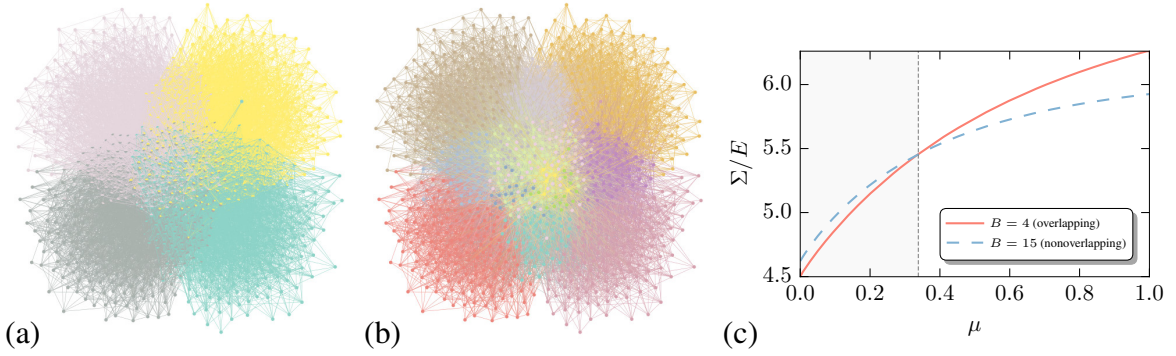


Figure 11. (a) Artificial network sampled from an assortative overlapping SBM with  $B = 4$  groups and expected mixture sizes given by  $n_{\bar{b}} \propto \mu^{|\bar{b}|}$ , with  $\mu \in [0, 1]$  controlling the degree of overlap (see Ref. [73] for details). (b) the same network as in (a), but generated according to an equivalent nonoverlapping SBM with  $B = 15$  groups. (c) Description length per edge  $\Sigma/E$  for the same models in (a) and (b), as a function of the degree of overlap  $\mu$ , showing a cross-over where the nonoverlapping model is preferred. Adapted from Ref. [73].

a mixed division into the same kinds of group. If we force the inference of only two groups, we see that some of the “center” nodes are split between “right wing” or “left wing.” The latter might seem like a more pleasing interpretation, but looking at the description length reveals that it does not improve the description of the data. The best model in this case does seem to be the overlapping SBM with  $B = 3$  groups. However, the difference in the description length between all model variants is not very large, making it difficult to fully reject any of the three variants. A more systematic analysis done in Ref. [18] revealed that for most empirical networks, in particular larger ones, the overlapping models do not provide the best fits in the majority of cases, and yield larger description lengths than the nonoverlapping variants. Hence it seems that the idea of overlapping groups is less pervasive than that of degree heterogeneity — at least according to our modeling ansatz.

It should be emphasized that we can always represent a network generated by an overlapping SBM by one generated with the nonoverlapping SBM with a larger number of groups representing the individual types of mixtures. Although model selection gives us the most parsimonious choice between the two, it does not remove the equivalence. In Fig. 11 we show how networks generated by the overlapping SBM can be better represented by the nonoverlapping SBM (i.e. with a smaller description length) as long as the overlapping regions are sufficiently large.

#### D. Further model extensions

The simple and versatile nature of the SBM has spawned a large family of extensions and generalizations incorporating various types of more realistic features. This includes, for example, versions of the SBM that are designed for networks with continuous edge covariates (a.k.a. edge weights) [74, 75], multilayer networks that are composed of different types of edges [73, 76–79], networks that evolve in time [34–41], networks that possess node attributes [80] or are annotated with metadata [57, 58], networks with uncertain structure [81], as well as networks that do not possess a discrete modular structure at all, and are instead embedded in generalized continuous spaces [82]. These model variations are too numerous to be described here in any detail. But it suffices to say that the general Bayesian approach outlined here, including model selection, also

applicable to these variations, without any conceptual difficulty.

### VIII. EFFICIENT INFERENCE USING MARKOV CHAIN MONTE CARLO (MCMC)

Although we can write exact expressions for the posterior probability of Eq. 9 (up to a normalization constant) for a variety of model variants, the resulting distributions are not simple enough to allow us to sample from them — much less find their maximum — in a direct manner. In fact, fully characterizing the posterior distribution or finding its maximum is, for most models like the SBM, typically a NP-hard problem. What we can do, however, is to employ Markov chain Monte Carlo (MCMC) [83], which can be done efficiently, and in an asymptotically exact manner, as we now show. The central idea is to sample from  $P(\mathbf{b}|\mathbf{A})$  by first starting from some initial configuration  $\mathbf{b}_0$  (in principle arbitrary), and making move proposals  $\mathbf{b} \rightarrow \mathbf{b}'$  with a probability  $P(\mathbf{b}'|\mathbf{b})$ , such that, after a sufficiently long time, the equilibrium distribution is given exactly by  $P(\mathbf{b}|\mathbf{A})$ . In particular, given any arbitrary move proposals  $P(\mathbf{b}'|\mathbf{b})$  — with the only condition that they fulfill ergodicity, i.e. that they allow every state to be visited eventually — we can guarantee that the desired posterior distribution is eventually reached by employing the Metropolis-Hastings criterion [84, 85], which dictates we should accept a given move proposal  $\mathbf{b} \rightarrow \mathbf{b}'$  with a probability  $a$  given by

$$a = \min \left( 1, \frac{P(\mathbf{b}'|\mathbf{A}) P(\mathbf{b}|\mathbf{b}')}{P(\mathbf{b}|\mathbf{A}) P(\mathbf{b}'|\mathbf{b})} \right), \quad (61)$$

otherwise the proposal is rejected. The ratio  $P(\mathbf{b}|\mathbf{b}')/P(\mathbf{b}'|\mathbf{b})$  in Eq. 61 enforces a property known as *detailed balance* or *reversibility*, i.e.

$$T(\mathbf{b}'|\mathbf{b})P(\mathbf{b}|\mathbf{A}) = T(\mathbf{b}|\mathbf{b}')P(\mathbf{b}'|\mathbf{A}), \quad (62)$$

where  $T(\mathbf{b}'|\mathbf{b})$  are the final transition probabilities after incorporating the acceptance criterion of Eq. 61. The detailed balance condition of Eq. 62 together with the ergodicity property guarantee that the Markov chain will converge to the desired equilibrium distribution  $P(\mathbf{b}|\mathbf{A})$ . Importantly, we note that when computing the ratio  $P(\mathbf{b}'|\mathbf{A})/P(\mathbf{b}|\mathbf{A})$  in Eq. 61, we do not need to determine the intractable normalization constant of Eq. 9, since it cancels out, and thus it can be performed exactly.

The above gives a generic protocol that we can use to sample from the posterior whenever we can compute the numerator of Eq. 9. If instead we are interested in maximizing the posterior, we can introduce an “inverse temperature” parameter  $\beta$ , by changing  $P(\mathbf{b}|\mathbf{A}) \rightarrow P(\mathbf{b}|\mathbf{A})^\beta$  in the above equations, and making  $\beta \rightarrow \infty$  in slow increments; what is known as *simulated annealing* [86]. The simplest implementation of this protocol for the inference of SBMs is to start from a random partition  $\mathbf{b}_0$ , and use move proposals where a node  $i$  is randomly selected, and then its new group membership  $b'_i$  is chosen randomly between all  $B + 1$  choices (where the remaining choice means we populate a new group),

$$P(b'_i|\mathbf{b}) = \frac{1}{B+1}. \quad (63)$$

By inspecting Eqs. 20, 41, 54 and 17 for all SBM variants considered, we notice that the ratio  $P(\mathbf{b}'|\mathbf{A})/P(\mathbf{b}|\mathbf{A})$  can be computed in time  $O(k_i)$ , where  $k_i$  is the degree of node  $i$ , independently of other properties of the model such as the number of groups  $B$ . Note that this is not true for all alternative formulations of the SBM; e.g. for the models in Refs. [30, 31, 33, 87, 88] computing

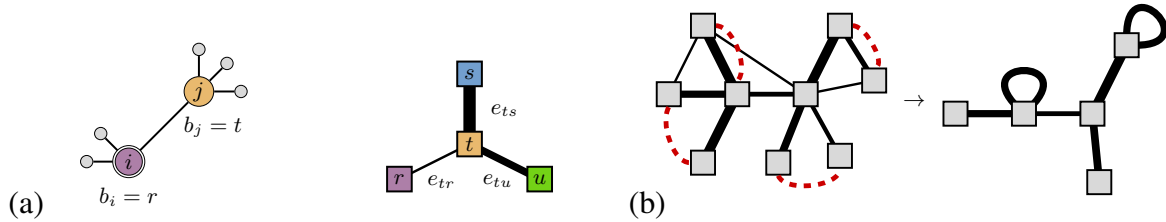


Figure 12. Efficient MCMC strategies: (a) Move proposals are made by inspecting the neighborhood of node  $i$  and selecting a random neighbor  $j$ . Based on its group membership  $t = b_j$ , the edge counts between groups are inspected (right), and the move proposal  $b_i = s$  is made with probability proportional to  $e_{ts}$ . (b) The initial state of the MCMC is obtained with an agglomerative heuristic, where groups are merged together using the same proposals described in (a).

such an update requires  $O(k_i + B)$  time [the heat-bath move proposals of Ref. [33] increases this even further to  $O(B(k_i + B))$ ], thus making them very inefficient for large networks, where the number of groups can reach the order of thousands or more. Hence, when using these move proposals, a full sweep of all  $N$  nodes in the network can be done in time  $O(E)$ , independent of  $B$ .

Although fairly simple, the above algorithm suffers from some shortcomings that can seriously degrade its performance in practice. In fact, it is typical for naive implementations of the Metropolis-Hastings algorithm to perform very badly, despite its theoretical guarantees. This is because the asymptotic properties of the Markov chain may take a very long time to be realized, and the equilibrium distribution is never observed in practical time. Generally, we should expect good convergence times only when: 1. The initial state  $\mathbf{b}_0$  is close enough to the most likely states of the posterior and 2. the move proposals  $P(\mathbf{b}'|\mathbf{b})$  resemble the shape of the posterior. Indeed, it is a trivial (and not very useful) fact that if the starting state  $\mathbf{b}_0$  is sampled directly from the posterior, and the move proposals match the posterior exactly,  $P(\mathbf{b}'|\mathbf{b}) = P(\mathbf{b}'|\mathbf{A})$ , the Markov chain would be instantaneously equilibrated. Hence if we can approach this ideal scenario, we should be able to improve the inference speeds. Here we describe two simple strategies in achieving such an improvement which have been shown to yield a significance performance impact [89]. The first one is to replace the fully random move proposals of Eq. 63 by a more informative choice. Namely, we use the current information about the model being inferred to guide our next move. We do so by selecting the membership of a node  $i$  being moved according to

$$P(b_i = r|\mathbf{b}) = \sum_s P(s|i) \frac{e_{sr} + \epsilon}{e_s + \epsilon(B+1)}, \quad (64)$$

where  $P(s|i) = \sum_j A_{ij} \delta_{b_j, s} / k_i$  is the fraction of neighbors of node  $i$  that belong to group  $s$ , and  $\epsilon > 0$  is an arbitrary parameter that enforces ergodicity, but with no other significant impact in the algorithm, provided it is sufficiently small (however, if  $\epsilon \rightarrow \infty$  we recover the fully random moves of Eq. 63). What this move proposal means is that we inspect the local neighborhood of the node  $i$ , and see which groups  $s$  are connected to this node, and we use the typical neighborhood  $r$  of the groups  $s$  to guide our placement of node  $i$  (see Fig. 12a). The purpose of these move proposals is not to waste time with attempted moves that will almost surely be rejected, as will typically happen with the fully random version. We emphasize that the move proposals of Eq. 64 do not bias the partitions toward any specific kind of mixing pattern; in particular they do not prefer assortative versus non-assortative partitions. Furthermore, these proposals can be generated efficiently, simply by following three steps: 1. sampling a random neighbor  $j$  of node  $i$ , and

inspecting its group membership  $s = b_j$ , and then; 2. with probability  $\varepsilon(B+1)/(e_s + \varepsilon(B+1))$  sampling a fully random group  $r$  (which can be a new group); 3. or otherwise, sampling a group label  $r$  with a probability proportional to the number of edges leading to it from group  $s$ ,  $e_{sr}$ . These steps can be performed in time  $O(k_i)$ , again independently of  $B$ , 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.

The second strategy is to choose a starting state that lies close to the mode of the posterior. We do so by performing a Fibonacci search [90] on the number of groups  $B$ , where for each value we obtain the best partition from a larger partition with  $B' > B$  using an agglomerative heuristic, composed of the following steps taken alternatively: 1. We attempt the moves of Eq. 64 until no improvement to the posterior is observed, 2. We merge groups together, achieving a smaller number of groups  $B'' \in [B, B']$ , stopping when  $B'' = B$ . We do the last step by treating each group as a single node and using Eq. 64 as a merge proposal, and selecting the ones that least decrease the posterior (see Fig 12b). As shown in Ref. [89], the overall complexity of this initialization algorithm is  $O(E \log^2 N)$ , and thus can be employed for very large networks.

The approach above can be adapted to the overlapping model of Sec. VII C, where instead of the partition  $\mathbf{b}$ , the move proposals are made with respect to the individual half-edge labels [18]. For the nested model, we have instead a hierarchical partition  $\{b_l\}$ , and we proceed in each step of the Markov chain by randomly choosing a level  $l$  and performing the proposals of Eq. 64 on that level, as described in Ref. [19].

The combination of the two strategies described above makes the inference procedure quite scalable, and has been successfully employed on networks on the order of  $10^7$  to  $10^8$  edges, and up to  $B = N$  groups. The MCMC algorithm described in this section, for all model variants described, is implemented in the `graph-tool` library [91], freely available under the GPL license at <http://graph-tool.skewed.de>.

## IX. TO SAMPLE OR TO OPTIMIZE?

In the examples so far, we have focused on obtaining the most likely partition from the posterior distribution, which is the one that minimizes the description length of the data. But is this in fact the best approach? In order to answer this, we need first to quantify how well our inference is doing, by comparing our estimate  $\hat{\mathbf{b}}$  of the partition to the true partition that generated the data  $\mathbf{b}^*$ , by defining a so-called *loss function*. For example, if we choose to be very strict, we may reject any partition that is strictly different from  $\mathbf{b}^*$  on equal measure, using the indicator function

$$\Delta(\hat{\mathbf{b}}, \mathbf{b}^*) = \prod_i \delta_{\hat{b}_i, b_i^*}, \quad (65)$$

so that  $\Delta(\hat{\mathbf{b}}, \mathbf{b}^*) = 1$  only if  $\hat{\mathbf{b}} = \mathbf{b}^*$ , otherwise  $\Delta(\hat{\mathbf{b}}, \mathbf{b}^*) = 0$ . If the observed data  $\mathbf{A}$  and parameters  $\mathbf{b}$  are truly sampled from the model and priors, respectively, the best assessment we can make for  $\mathbf{b}^*$  is given by the posterior distribution  $P(\mathbf{b}|\mathbf{A})$ . Therefore, the average of the indicator over the posterior is given by

$$\bar{\Delta}(\hat{\mathbf{b}}) = \sum_{\mathbf{b}} \Delta(\hat{\mathbf{b}}, \mathbf{b}) P(\mathbf{b}|\mathbf{A}). \quad (66)$$

If we maximize  $\bar{\Delta}(\hat{\mathbf{b}})$  with respect to  $\hat{\mathbf{b}}$ , we obtain the so-called maximum *a posteriori* (MAP) estimator

$$\hat{\mathbf{b}} = \underset{\mathbf{b}}{\operatorname{argmax}} P(\mathbf{b}|\mathbf{A}), \quad (67)$$

which is precisely what we have been using so far, and it is equivalent to employing the MDL principle. However, using this estimator is arguably overly optimistic, as we are unlikely to find the true partition with perfect accuracy in any but the most ideal cases. Instead, we may relax our expectations and consider instead the overlap function

$$d(\hat{\mathbf{b}}, \mathbf{b}^*) = \frac{1}{N} \sum_i \delta_{\hat{b}_i, b_i^*}, \quad (68)$$

which measures the *fraction* of nodes that are correctly classified. If we maximize now the average of the overlap over the posterior distribution

$$\bar{d}(\hat{\mathbf{b}}) = \sum_{\mathbf{b}} d(\hat{\mathbf{b}}, \mathbf{b}) P(\mathbf{b}|\mathbf{A}), \quad (69)$$

we obtain the *marginal estimator*

$$\hat{b}_i = \underset{r}{\operatorname{argmax}} \pi_i(r), \quad (70)$$

where

$$\pi_i(r) = \sum_{\mathbf{b} \setminus b_i} P(b_i = r, \mathbf{b} \setminus b_i | \mathbf{A}) \quad (71)$$

is the marginal distribution of the group membership of node  $i$ , summed over all remaining nodes.<sup>19</sup> The marginal estimator is notably different from the MAP estimator in that it leverages information from the entire posterior distribution to inform the classification of any single node. If the posterior is tightly concentrated around its maximum, both estimators will yield compatible answers. In this situation the structure in the data is very clear, and both estimators agree. Otherwise, the estimators will yield different aspects of the data, in particular if the posterior possesses many local maxima. For example, if the data has indeed been sampled from the model we are using, the multiplicity of local maxima can be just a reflection of the randomness in the data, and the marginal estimator will be able to average over them and provide better accuracy [92, 93].

In view of the above, one could argue that the marginal estimator should be generally preferred over MAP. However, the situation is more complicated for data which are not sampled from model being used for inference (i.e. the model is *misspecified*). In this situation, multiple peaks of the distribution can point to very different partitions that are all statistically significant. These different peaks function as alternative explanations for the data that must be accepted on equal footing, according to their posterior probability. The marginal estimator will in general mix the properties of all peaks into a consensus classification that is not representative of any single hypothesis, whereas the MAP estimator will concentrate only on the most likely one (or an arbitrary choice if they are all equally likely). An illustration of this is given by the well-known Zachary's karate club network [95], which captures the social interactions between members of a karate club amidst a

<sup>19</sup>The careful reader will notice that we must have in fact a trivial constant marginal  $\pi_i(r) = 1/B$  for every node  $i$ , since there is a symmetry of the posterior distribution with respect to re-labelling of the groups, in principle rendering this estimator useless. In practice, however, our samples from the posterior distribution (e.g. using MCMC) will not span the whole space of label permutations in any reasonable amount of time, and instead will concentrate on a mode around one of the possible permutations. Since the modes around the label permutations are entirely symmetric, the node marginals obtained in this manner can be meaningfully used. However, for networks where some of the groups are not very large, *local* permutations of individual group labels are statistically possible during MCMC inference, leading to degeneracies in the marginal  $\pi_i(r)$  of the affected nodes, resulting in artefacts when using the marginal estimator. This problem is exacerbated when the number of groups changes during MCMC sampling.

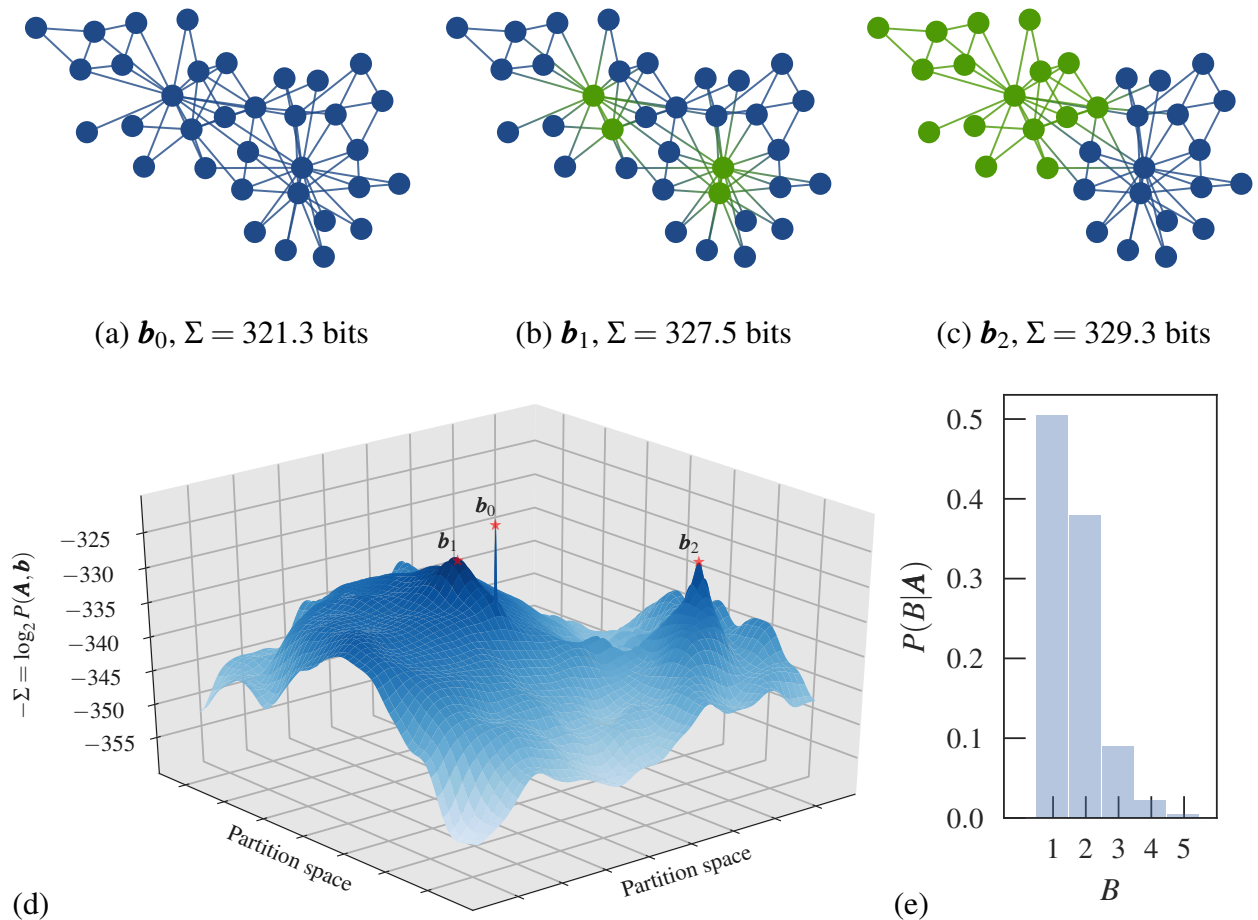


Figure 13. Posterior distribution of partitions of Zachary’s karate club network using the DC-SBM. Panels (a) to (c) show three modes of the distribution and their respective description lengths; (d) 2D projection of the posterior obtained using multidimensional scaling [94]; (e) Marginal posterior distribution of the number of groups  $B$ .

conflict between the club’s administrator and an instructor, which lead to a split of the club in two disjoint groups. The measurement of the network was done before the final split actually happened, and it is very often used as an example of a network exhibiting community structure. If we analyze this network with the DC-SBM, we obtain three partitions that occur with very high probability from the posterior distribution: A trivial  $B = 1$  partition, corresponding to the configuration model without communities (Fig. 13a), a “leader-follower” division into  $B = 2$  groups, separating the administrator and instructor, together with two close allies, from the rest of the network (Fig. 13b), and finally a  $B = 2$  division into the aforementioned factions that anticipated the split (Fig. 13c). If we would guide ourselves strictly by the MDL principle (i.e. using the MAP estimator), the preferred partition would be the trivial  $B = 1$  one, indicating that the most likely explanation of this network is a fully random graph with a pre-specified degree sequence, and that the observed community structure emerged spontaneously. However, if we inspect the posterior distribution more closely, we see that other divisions into  $B > 1$  groups amount to around 50% of the posterior probability (see Fig. 13e). Therefore, if we consider all  $B > 1$  partitions *collectively*, they give us little reason to completely discard the possibility that the network does in fact possess some group

structure. Inspecting the posterior distribution even more closely, as shown in Fig. 13d, reveals a multimodal structure clustered around the three aforementioned partitions, giving us three very different explanations for the data, none of which can be decisively discarded in favor of the others — at least not according to the evidence available in the network structure alone.

The situation encountered for the karate club network is a good example of the so-called *bias-variance trade-off* that we are often forced to face: If we choose to single-out a single partition as a unique representation of the data, we must invariably *bias* our result toward any of the three most likely scenarios, discarding the remaining ones at some loss of useful information. Otherwise, if we choose to eliminate the bias by incorporating the entire posterior distribution in our representation, by the same token it will incorporate a larger variance, i.e. it will simultaneously encompass diverging explanations of the data, leaving us without an unambiguous and clear interpretation. The only situation where this trade-off is not required is when the model is a perfect fit to the data, such that the posterior is tightly peaked around a single partition. Therefore, the variance of the posterior serves as a good indication of the quality of fit of the model, providing another reason to include it in the analysis.

It should also be remarked that when using a nonparametric approach, where the dimension of the model is also inferred from the posterior distribution, the potential bias incurred when obtaining only the most likely partition usually amounts to an *underfit* of the data, since the uncertainty in the posterior typically translates into the existence of a more conservative partition with fewer groups.<sup>20</sup> Instead, if we sample from the posterior distribution, we will average over many alternative fits, including those that model the data more closely with a larger number of groups. However, each individual sample of the posterior will tend to incorporate more randomness from the data, which will disappear only if we average over all samples. This means that single samples will tend to overfit the data, and hence we must resist looking at them individually. It is only in the aforementioned limit of a perfect fit that we are guaranteed not to be misled one way or another. An additional example of this is shown in Fig. 14 for a network of collaborations among scientists. If we infer the best nested SBM, we find a specific hierarchical division of the network. However, if we sample hierarchical divisions from the posterior distribution, we typically encounter larger models — with a larger number of groups and deeper hierarchy. Each individual sample from the posterior is likely to be an overfit, but collectively they give a more accurate picture of the network in comparison with the most likely partition, which probably over-simplifies it. As already mentioned, this discrepancy, observed for all three SBM versions, tells us that neither of them is an ideal fit for this network.

The final decision on which approach to take depends on the actual objective and resources available. In general, sampling from the posterior will be more suitable when the objective is to generalize from observation and make predictions (see next section and Ref. [97]), and when computational resources are ample. Conversely, if the objective is to make a precise statement about the data, e.g. in order to summarize and interpret it, and the computational resources are scarce, maximizing the posterior tends to be more adequate.

## X. GENERALIZATION AND PREDICTION

When we fit a model like the SBM to a network, we are doing more than simply dividing the nodes into statistically equivalent groups; we are also making a statement about a possible

<sup>20</sup>This is different from *parametric* posteriors, where the dimension of the model is externally imposed in the prior, and the MAP estimator tends to overfit [92, 93].

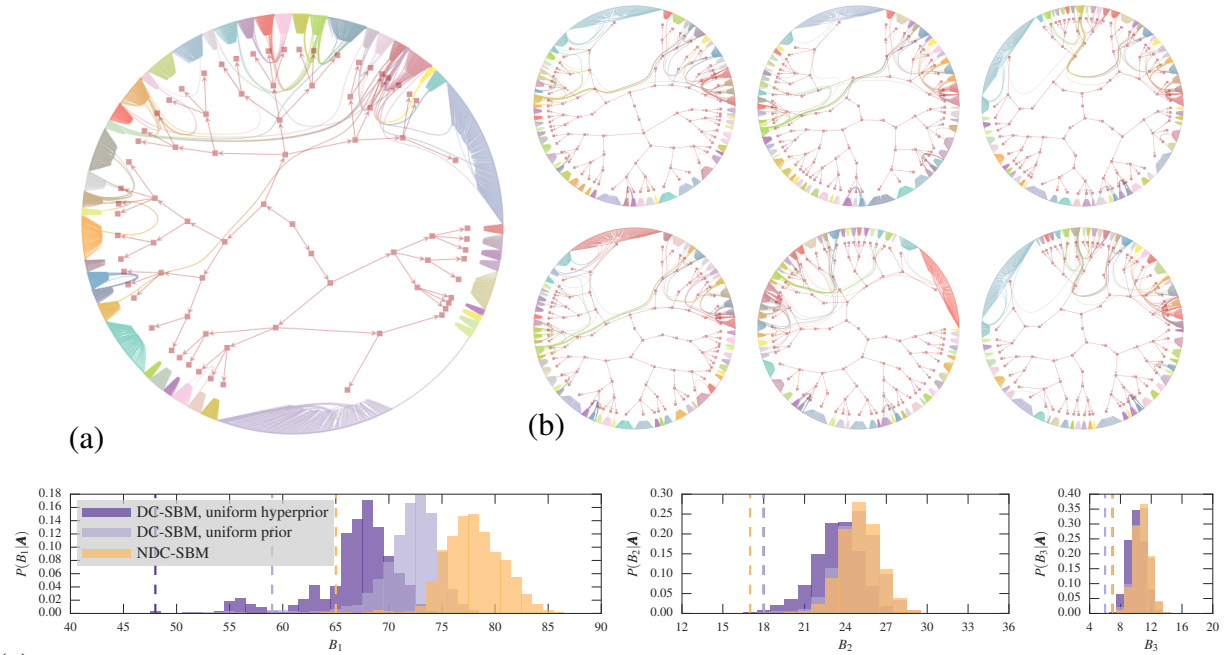


Figure 14. Hierarchical partitions of a network of collaboration between scientists [96]. (a) Most likely hierarchical partition according to the DC-SBM with a uniform hyperprior. (b) Uncorrelated samples from the posterior distribution. (c) Marginal posterior distribution of the number of groups at 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. Adapted from Ref. [19].

mechanism that generated the network. This means that, to the extent that the model is a good representation of the data, we can use it generalize and make predictions about what has *not* been observed. This has been most explored for the prediction of missing and spurious links [25, 68]. This represents the situation where we know or stipulate that the observed data is noisy, and may contain edges that in fact do not exist, or does not contain edges that do exist. With a generative model like the SBM, we are able to ascribe probabilities to existing and non-existing edges of being spurious or missing, respectively, as we now describe.

Following Ref. [97], the scenario we will consider is the situation where there exists a complete network  $\mathbf{G}$  which is decomposed in two parts,

$$\mathbf{G} = \mathbf{A}^O + \delta\mathbf{A} \quad (72)$$

where  $\mathbf{A}^O$  is the network that we observe, and the  $\delta\mathbf{A}$  is the set of missing and spurious edges that we want to predict, where an entry  $\delta A_{ij} > 0$  represents a missing edge, and  $\delta A_{ij} < 0$  a spurious one. Hence, our task is to obtain the posterior distribution

$$P(\delta\mathbf{A}|\mathbf{A}^O). \quad (73)$$

The central assumption we will make is that the complete network  $\mathbf{G}$  has been generated using some arbitrary version of the SBM, with a marginal distribution

$$P_G(\mathbf{G}|\mathbf{b}). \quad (74)$$



Given a generated network  $\mathbf{G}$ , we then select  $\delta\mathbf{A}$  from some arbitrary distribution that models our source of errors

$$P_{\delta\mathbf{A}}(\delta\mathbf{A}|\mathbf{G}). \quad (75)$$

With the above model for the generation of the complete network and its missing and spurious edges, we can proceed to compute the posterior of Eq. 73. We start from the joint distribution

$$P(\mathbf{A}^O, \delta\mathbf{A}|\mathbf{G}) = P(\mathbf{A}^O|\delta\mathbf{A}, \mathbf{G})P_{\delta\mathbf{A}}(\delta\mathbf{A}|\mathbf{G}) \quad (76)$$

$$= \delta(\mathbf{G} - (\mathbf{A}^O + \delta\mathbf{A}))P_{\delta\mathbf{A}}(\delta\mathbf{A}|\mathbf{G}), \quad (77)$$

where we have used the fact  $P(\mathbf{A}^O|\delta\mathbf{A}, \mathbf{G}) = \delta(\mathbf{G} - (\mathbf{A}^O + \delta\mathbf{A}))$  originating from Eq. 72. For the joint distribution conditioned on the partition, we sum the above over all possible graphs  $\mathbf{G}$ , sampled from our original model,

$$P(\mathbf{A}^O, \delta\mathbf{A}|\mathbf{b}) = \sum_{\mathbf{G}} P(\mathbf{A}^O, \delta\mathbf{A}|\mathbf{G})P_{\mathbf{G}}(\mathbf{G}|\mathbf{b}) \quad (78)$$

$$= P_{\delta\mathbf{A}}(\delta\mathbf{A}|\mathbf{A}^O + \delta\mathbf{A})P_{\mathbf{G}}(\mathbf{A}^O + \delta\mathbf{A}|\mathbf{b}). \quad (79)$$

The final posterior distribution of Eq. 73 is therefore

$$P(\delta\mathbf{A}|\mathbf{A}^O) = \frac{\sum_{\mathbf{b}} P(\mathbf{A}^O, \delta\mathbf{A}|\mathbf{b})P(\mathbf{b})}{P(\mathbf{A}^O)} \quad (80)$$

$$= \frac{P_{\delta\mathbf{A}}(\delta\mathbf{A}|\mathbf{A}^O + \delta\mathbf{A})\sum_{\mathbf{b}} P_{\mathbf{G}}(\mathbf{A}^O + \delta\mathbf{A}|\mathbf{b})P(\mathbf{b})}{P(\mathbf{A}^O)}, \quad (81)$$

with  $P(\mathbf{A}^O)$  being a normalization constant, independent of  $\delta\mathbf{A}$ . This expression gives a general recipe to compute the posterior, where one averages the marginal likelihood  $P_{\mathbf{G}}(\mathbf{A}^O + \delta\mathbf{A}|\mathbf{b})$  obtained by sampling partitions from the prior  $P(\mathbf{b})$ . However, this procedure will typically take an astronomical time to converge to the correct asymptotic value, since the largest values of  $P_{\mathbf{G}}(\mathbf{A}^O + \delta\mathbf{A}|\mathbf{b})$  will be far away from most values of  $\mathbf{b}$  sampled from  $P(\mathbf{b})$ . A much better approach is to perform importance sampling, by rewriting the posterior as

$$P(\delta\mathbf{A}|\mathbf{A}^O) \propto P_{\delta\mathbf{A}}(\delta\mathbf{A}|\mathbf{A}^O + \delta\mathbf{A}) \sum_{\mathbf{b}} P_{\mathbf{G}}(\mathbf{A}^O + \delta\mathbf{A}|\mathbf{b}) \frac{P_{\mathbf{G}}(\mathbf{A}^O|\mathbf{b})}{P_{\mathbf{G}}(\mathbf{A}^O|\mathbf{b})} P(\mathbf{b}) \quad (82)$$

$$\propto P_{\delta\mathbf{A}}(\delta\mathbf{A}|\mathbf{A}^O + \delta\mathbf{A}) \sum_{\mathbf{b}} \frac{P_{\mathbf{G}}(\mathbf{A}^O + \delta\mathbf{A}|\mathbf{b})}{P_{\mathbf{G}}(\mathbf{A}^O|\mathbf{b})} P_{\mathbf{G}}(\mathbf{b}|\mathbf{A}^O), \quad (83)$$

where  $P_{\mathbf{G}}(\mathbf{b}|\mathbf{A}^O)$  is the posterior of partitions obtained by pretending that the observed network came directly from the SBM. We can sample from this posterior using MCMC as described in Sec. VIII. As the number of entries in  $\delta\mathbf{A}$  is typically much smaller than the number of observed edges, this importance sampling approach will tend to converge much faster. This allows us to compute  $P(\delta\mathbf{A}|\mathbf{A}^O)$  in practical manner — up to a normalization constant. However, if we want to compare the relative probability between specific sets of missing/spurious edges,  $\{\delta\mathbf{A}_i\}$ , via the

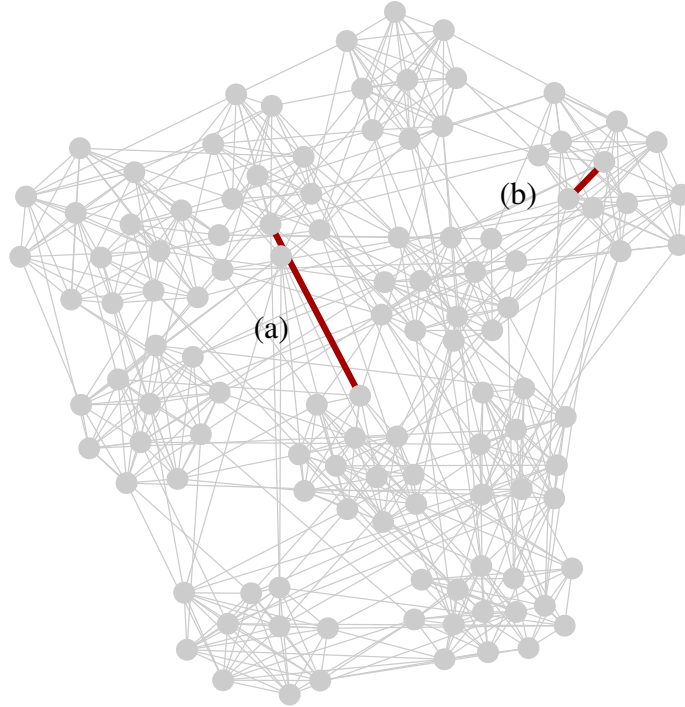


Figure 15. Two hypothetical missing edges in the network of American college football teams. The edge (a) connects teams of different conferences, whereas (b) connects teams of the same conference. According to the nested DC-SBM, their posterior probability ratios are  $\lambda_a \approx 0.013(1)$  and  $\lambda_b \approx 0.987(1)$ .

ratio

$$\lambda_i = \frac{P(\delta \mathbf{A}_i | \mathbf{A}^O)}{\sum_j P(\delta \mathbf{A}_j | \mathbf{A}^O)}, \quad (84)$$

this normalization constant plays no role. The above still depends on our chosen model for the production of missing and spurious edges, given by Eq. 75. In the absence of domain-specific information about the source of noise, we must consider all alternative choices  $\{\delta \mathbf{A}_i\}$  to be equally likely a priori, so that we can simply replace  $P_{\delta \mathbf{A}}(\delta \mathbf{A} | \mathbf{A}^O + \delta \mathbf{A}) \propto 1$  in Eq. 83 — although more realistic choices can also be included.

In Fig. 15 we show the relative probabilities of two hypothetical missing edges for the American college football network, obtained with the approach above. We see that a particular missing edge between teams of the same conference is almost a hundred times more likely than one between teams of different conference.

The use of the SBM to predict missing and spurious edges has been employed in a variety of applications, such as the prediction of novel interactions between drugs [98], conflicts in social networks [99], as well to provide user recommendations [100, 101], and in many cases has outperformed a variety of competing methods.

## XI. FUNDAMENTAL LIMITS OF INFERENCE: THE DETECTABILITY-INDETECTABILITY PHASE TRANSITION

Besides defining useful models and investigating their behavior in data, there is another line of questioning which deals with how far it is possible to go when we try to infer the structure of networks. Naturally, the quality of the inference depends on the statistical evidence available in the data, and we may therefore ask if it is possible at all to uncover *planted* structures — i.e. structures that we impose ourselves — with our inference methods, and if so, what is the best performance we can expect. Research in this area has exploded in recent years [92, 93], after it was shown by Decelle et al [102, 103] that not only it may be impossible to uncover planted structures with the SBM, but the inference undergoes a “phase transition” where it becomes possible only if the structure is strong enough to cross a non-trivial threshold. This result was obtained using methods from statistical physics, which we now describe.

The situation we will consider is a “best case scenario,” where all parameters of the model are known, with the exception of the partition  $\mathbf{b}$  — this in contrast to our overall approach so far, where we considered all parameters to be unknown random variables. In particular, we will consider only the prior

$$P(\mathbf{b}|\boldsymbol{\gamma}) = \prod_i \gamma_{b_i}. \quad (85)$$

where  $\gamma_r$  is the probability of a node belonging in group  $r$ . Given this, we wish to obtain the posterior distribution of the node partition, using the SBM of Eq. 8,

$$P(\mathbf{b}|\mathbf{A}, \boldsymbol{\lambda}, \boldsymbol{\gamma}) = \frac{P(\mathbf{A}|\mathbf{b}, \boldsymbol{\lambda})P(\mathbf{b}|\boldsymbol{\gamma})}{P(\mathbf{A}|\boldsymbol{\lambda}, \boldsymbol{\gamma})} = \frac{e^{-\mathcal{H}(\mathbf{b})}}{Z} \quad (86)$$

which was written above in terms of the “Hamiltonian”

$$\mathcal{H}(\mathbf{b}) = - \sum_{i < j} (A_{ij} \ln \lambda_{b_i, b_j} - \lambda_{b_i, b_j}) - \sum_i \ln \gamma_{b_i}, \quad (87)$$

drawing an analogy with Potts-like models in statistical physics [104]. The normalization constant, called the “partition function,” is given by

$$Z = \sum_{\mathbf{b}} e^{-\mathcal{H}(\mathbf{b})}. \quad (88)$$

Far from being an unimportant detail, the partition function can be used to determine all statistical properties of our inference procedure. For example, if we wish to obtain the marginal posterior distribution of node  $i$ , we can do so by introducing the perturbation  $\mathcal{H}'(\mathbf{b}) = \mathcal{H}(\mathbf{b}) - \mu \delta_{b_i, r}$  and computing the derivative

$$P(b_i = r|\mathbf{A}, \boldsymbol{\lambda}, \boldsymbol{\gamma}) = \left. \frac{\partial \ln Z}{\partial \mu} \right|_{\mu=0} = \sum_{\mathbf{b}} \delta_{b_i, r} \frac{e^{-\mathcal{H}(\mathbf{b})}}{Z}. \quad (89)$$

Unfortunately, it does not seem possible to compute the partition function  $Z$  in closed form for an arbitrary graph  $\mathbf{A}$ . However, there is a special case for which we *can* compute the partition function, namely when  $\mathbf{A}$  is a *tree*. This is useful for us, because graphs sampled from the SBM will be “locally tree-like” if they are sparse (i.e. the degrees are small compared to the size of

the network  $k_i \ll N$ , and the group sizes scale with the size of the system, i.e.  $n_r = O(N)$  (which implies  $B \ll N$ ). Locally tree-like means that typical loops will have length  $O(N)$ , and hence at the immediate neighborhood of any given node the graph will look like a tree. Although being locally tree-like is not quite the same as being a tree, the graph will become increasing *closer* to being a tree in the “thermodynamic limit”  $N \rightarrow \infty$ . Because of this, many properties of locally tree-like graphs will become asymptotically identical to trees in this limit. If we assume that this limit holds, we can compute the partition function by pretending that the graph is close enough to being a tree, in which case we can write the so-called Bethe free energy (we refer to Refs. [103, 105] for a detailed derivation)

$$\mathcal{F} = -\ln Z = -\sum_i \ln Z^i + \sum_{i<j} A_{ij} \ln Z^{ij} - E \quad (90)$$

with the auxiliary quantities given by

$$Z^{ij} = N \sum_{r<s} \lambda_{rs} (\psi_r^{i \rightarrow j} \psi_s^{j \rightarrow i} + \psi_s^{i \rightarrow j} \psi_r^{j \rightarrow i}) + N \sum_r \lambda_{rr} \psi_r^{i \rightarrow j} \psi_r^{j \rightarrow i} \quad (91)$$

$$Z^i = \sum_r n_r e^{-h_r} \prod_{j \in \partial i} \sum_r N \lambda_{rb_i} \psi_r^{j \rightarrow i}, \quad (92)$$

where  $\partial i$  means the neighbors of node  $i$ . In the above equations, the values  $\psi_r^{i \rightarrow j}$  are called “messages,” and they must fulfill the self-consistency equations

$$\psi_r^{i \rightarrow j} = \frac{1}{Z^{i \rightarrow j}} \gamma_r e^{-h_r} \prod_{k \in \partial i \setminus j} \left( \sum_s N \lambda_{rs} \psi_s^{k \rightarrow i} \right) \quad (93)$$

where  $k \in \partial i \setminus j$  means all neighbors  $k$  of  $i$  excluding  $j$ , the value  $Z^{i \rightarrow j}$  is a normalization constant enforcing  $\sum_r \psi_r^{i \rightarrow j} = 1$ , and  $h_r = \sum_i \sum_r \lambda_{rb_i} \psi_r^i$  is a local auxiliary field. Eqs. 93 are called the **belief-propagation** (BP) equations [105], and the entire approach is also known under the name “cavity method” [106]. The values of the messages are typically obtained by iteration, where we start from some initial configuration (e.g. a random one), and compute new values from the right-hand side of Eq. 93, until they converge asymptotically. Note that the messages are only defined on edges of the network, and an update involves inspecting the values at the neighborhood of the nodes, where the messages can be interpreted as carrying information about the marginal distribution of a given node, if the same is removed from the network (hence the names “belief propagation” and “cavity method”). Each iteration of the BP equations can be done in time  $O(EB^2)$ , and the convergence is often obtained only after a few iterations, rendering the whole computation fairly efficient, provided  $B$  is reasonably small. After the messages have been obtained, they can be used to compute the node marginals,

$$P(b_i = r | \mathbf{A}, \boldsymbol{\lambda}, \boldsymbol{\gamma}) = \psi_r^i = \frac{1}{Z^i} \gamma_r \prod_{j \in \partial i} \left[ \sum_s (N \lambda_{rs})^{A_{ij}} e^{-\lambda_{rs}} \psi_s^{j \rightarrow i} \right], \quad (94)$$

where  $Z^i$  is a normalization constant.

This whole procedure gives a way of computing the marginal distribution  $P(b_i = r | \mathbf{A}, \boldsymbol{\lambda}, \boldsymbol{\gamma})$  in a manner that is asymptotically exact — if  $\mathbf{A}$  is sufficiently large and locally tree-like. Since networks that are sampled from the SBM fulfill this property<sup>21</sup>, we may proceed with our original

<sup>21</sup>Real networks, however, should not be expected to be locally tree-like. This does not invalidate the results of this

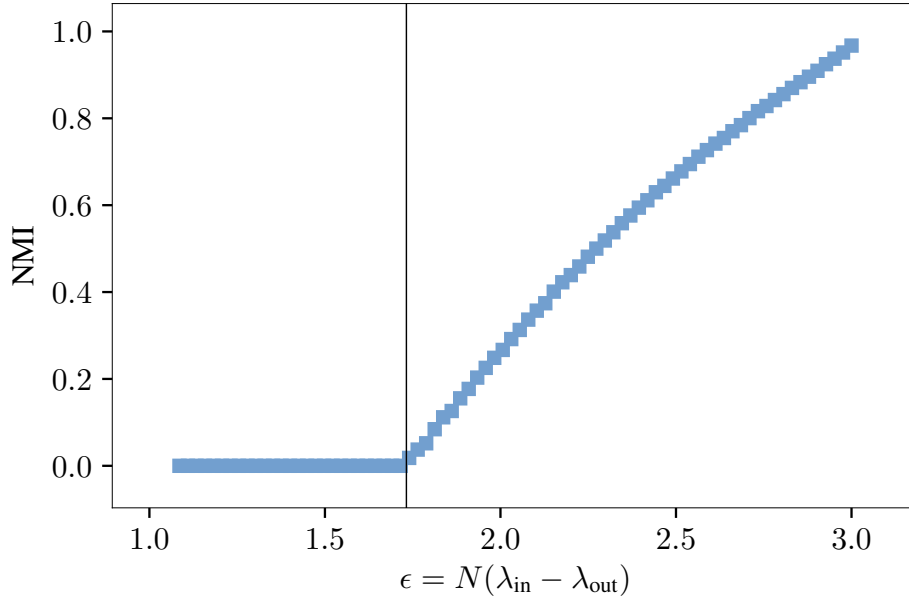


Figure 16. Normalized mutual information (NMI) between the planted and inferred partitions of a PP model with  $N = 10^5$ ,  $B = 3$  and  $\langle k \rangle = 3$  and  $\varepsilon = N(\lambda_{\text{in}} - \lambda_{\text{out}})$ . The vertical line marks the detectability threshold  $\varepsilon^* = B\sqrt{\langle k \rangle}$ .

question, and test if we can recover the true value of  $\mathbf{b}$  we used to generate a network. For the test, we use a simple parametrization named the planted partition model (PP) [10, 107], where  $\gamma_r = 1/B$  and

$$\lambda_{rs} = \lambda_{\text{in}} \delta_{rs} + \lambda_{\text{out}} (1 - \delta_{rs}), \quad (95)$$

with  $\lambda_{\text{in}}$  and  $\lambda_{\text{out}}$  specifying the expected number of edges between nodes of the same groups and of different groups, respectively. If we generate networks from this ensemble, use the BP equations to compute the posterior marginal distribution of Eq. 94 and compare its maximum values with the planted partition, we observe, as shown in Fig. 16, that it is recoverable only up to a certain value of  $\varepsilon = N(\lambda_{\text{in}} - \lambda_{\text{out}})$ , above which the posterior distribution is fully uniform. By inspecting the stability of the fully uniform solution of the BP equations, the exact threshold can be determined [103],

$$\varepsilon^* = B\sqrt{\langle k \rangle}, \quad (96)$$

where  $\langle k \rangle = N \sum_{rs} \lambda_{rs} / B^2$  is the average degree of the network. The existence of this threshold is remarkable, because the ensemble is only equivalent to a completely random one if  $\varepsilon = 0$ ; yet there is a non-negligible range of values  $\varepsilon \in [0, \varepsilon^*]$  for which *the planted structure cannot be recovered even though the model is not random*. This might seem counter-intuitive, if we argue that making  $N$  sufficiently large should at some point give us enough data to infer the model with arbitrary precision. The hole in logic lies in the fact that the number of parameters — the node partition  $\mathbf{b}$  — also grows with  $N$ , and that we would need the effective sample size, i.e. the number of edges  $E$ , to grow *faster* than  $N$  to guarantee that the data is sufficient. Since for sparse graphs we have  $E = O(N)$ , we are never able to reach the limit of sufficient data. Thus, we should be able to achieve asymptotically perfect inference only for dense graphs (e.g. with  $E = O(N^2)$ ) or by

---

section, which pertain strictly to data sampled from the SBM. However, despite not being exact, the BP algorithm yields surprisingly accurate results for real networks, even when the tree-like property is violated [103].

inferring simultaneously from many graphs independently sampled from the same model. Neither situation, however, is representative of what we typically encounter when we study networks.

The above result carries important implications into the overall field of network clustering. The existence of the “detectable” phase for  $\varepsilon > \varepsilon^*$  means that, in this regime, it is possible for algorithms to discover the planted partition in polynomial time, with the BP algorithm doing so optimally. Furthermore, for  $B > 4$  (or  $B > 3$  for the dissortative case with  $\lambda_{\text{in}} < \lambda_{\text{out}}$ ) there is another regime in a range  $\varepsilon^* < \varepsilon < \varepsilon^\dagger$ , where BP converges to the planted partition only if the messages are initialized close enough to the corresponding fixed point. In this regime, the posterior landscape exhibits a “glassy” structure, with exponentially many maxima that are almost as likely as the planted partition, but are completely uncorrelated with it. The problem of finding the planted partition in this case is possible, but conjectured to be NP-hard.

Many systematic comparisons of different community detection algorithms were done in a manner that was oblivious to these fundamental facts regarding detectability and hardness [108, 109], even though their existence had been conjectured before [110, 111], and hence should be re-framed with it in mind. Furthermore, we point out that although the analysis based on the BP equations is mature and widely accepted in statistical physics, they are not completely rigorous from a mathematical point of view. Because of this, the result of Decelle et al [103] leading to the threshold of Eq. 96 has initiated intense activity from mathematicians in search of rigorous proofs, which have subsequently been found for a variety of relaxations of the original statement (see Ref. [112] for a review), and remains an active area of research.

## XII. CONCLUSION

In this chapter we gave a description of the basic variants of the stochastic blockmodel (SBM), and a consistent Bayesian formulation that allows us to infer them from data. The focus has been on developing a framework to extract the large-scale structure of networks while avoiding both overfitting (mistaking randomness for structure) and underfitting (mistaking structure for randomness), and doing so in a manner that is analytically tractable and computationally efficient.

The Bayesian inference approach provides a methodologically correct answer to the very central question in network analysis of whether patterns of large-scale structure can in fact be supported by statistical evidence. Besides this practical aspect, it also opens a window into the fundamental limits of network analysis itself, giving us a theoretical underpinning we can use to understand more about the nature of network systems.

Although the methods described here go a long way into allowing us to understand the structure of networks, some important open problems remain. From a modeling perspective, we know that for most systems the SBM is quite simplistic, and falls very short of giving us a mechanistic explanation for them. We can interpret the SBM as being to network data what a histogram is to spatial data [113], and thus while it fulfills the formal requirements of being a generative model, it will never deplete the modeling requirements of any particular real system. Although it is naive to expect to achieve such a level of success with a general model like the SBM, it is yet still unclear how far we can go. For example, it remains to be seen how tractable it is to incorporate local structures — like densities of subgraphs — together with the large-scale structure that the SBM prescribes.

From a methodological perspective, although we can select between the various SBM flavors given the statistical evidence available, we still lack good methods to assess the quality of fit of the SBM at an absolute level. In particular, we do not yet have a systematic understanding of how well the SBM is able to reproduce properties of empirical systems, and what would be the most

important sources of deficiencies, and how these could be overcome.

In addition to these outstanding challenges, there are areas of development that are more likely to undergo continuous progress. Generalizations and extensions of the SBM to cover specific cases are essentially open ended, such as the case of dynamic networks, and we can perhaps expect more realistic models to appear. Furthermore, since the inference of the SBM is in general a NP-hard problem, and thus most probably lacks a general solution, the search for more efficient algorithmic strategies that work in particular cases is also a long term goal that is likely to attract further attention.

- 
- [1] Paul Erdős and Alfréd Rényi, “On random graphs, I,” *Publicationes Mathematicae (Debrecen)* **6**, 290–297 (1959).
  - [2] Roger Guimerà, Marta Sales-Pardo, and Luís A. Nunes Amaral, “Modularity from fluctuations in random graphs and complex networks,” *Physical Review E* **70**, 025101 (2004).
  - [3] E. T. Jaynes, *Probability Theory: The Logic of Science*, edited by G. Larry Bretthorst (Cambridge University Press, Cambridge, UK ; New York, NY, 2003).
  - [4] Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt, “Stochastic blockmodels: First steps,” *Social Networks* **5**, 109–137 (1983).
  - [5] Yuchung J. Wang and George Y. Wong, “Stochastic Blockmodels for Directed Graphs,” *Journal of the American Statistical Association* **82**, 8–19 (1987).
  - [6] Tom A. B. Snijders and Krzysztof Nowicki, “Estimation and Prediction for Stochastic Blockmodels for Graphs with Latent Block Structure,” *Journal of Classification* **14**, 75–100 (1997).
  - [7] Krzysztof Nowicki and Tom A. B. Snijders, “Estimation and Prediction for Stochastic Blockstructures,” *Journal of the American Statistical Association* **96**, 1077–1087 (2001).
  - [8] Bo Söderberg, “General formalism for inhomogeneous random graphs,” *Physical Review E* **66**, 066121 (2002).
  - [9] Béla Bollobás, Svante Janson, and Oliver Riordan, “The phase transition in inhomogeneous random graphs,” *Random Structures & Algorithms* **31**, 3–122 (2007).
  - [10] Anne Condon and Richard M. Karp, “Algorithms for graph partitioning on the planted partition model,” *Random Structures & Algorithms* **18**, 116–140 (2001).
  - [11] Marián Boguñá and Romualdo Pastor-Satorras, “Class of correlated random networks with hidden variables,” *Physical Review E* **68**, 036112 (2003).
  - [12] J.-J. Daudin, F. Picard, and S. Robin, “A mixture model for random graphs,” *Statistics and Computing* **18**, 173–183 (2008).
  - [13] Ginestra Bianconi, Paolo Pin, and Matteo Marsili, “Assessing the relevance of node features for network structure,” *Proceedings of the National Academy of Sciences* **106**, 11433–11438 (2009).
  - [14] Santo Fortunato, “Community detection in graphs,” *Physics Reports* **486**, 75–174 (2010).
  - [15] Mark Newman, *Networks: An Introduction* (Oxford University Press, 2010).
  - [16] Brian Karrer and M. E. J. Newman, “Stochastic blockmodels and community structure in networks,” *Physical Review E* **83**, 016107 (2011).
  - [17] Tiago P. Peixoto, “Entropy of stochastic blockmodel ensembles,” *Physical Review E* **85**, 056122 (2012).
  - [18] Tiago P. Peixoto, “Model Selection and Hypothesis Testing for Large-Scale Network Models with Overlapping Groups,” *Physical Review X* **5**, 011033 (2015).
  - [19] Tiago P. Peixoto, “Nonparametric Bayesian inference of the microcanonical stochastic block model,”

- Physical Review E **95**, 012317 (2017).
- [20] M. B. Hastings, “Community detection as an inference problem,” *Physical Review E* **74**, 035102 (2006).
- [21] Charles Kemp and Joshua B. Tenenbaum, “Learning systems of concepts with an infinite relational model,” in *In Proceedings of the 21st National Conference on Artificial Intelligence* (2006).
- [22] Martin Rosvall and Carl T. Bergstrom, “An information-theoretic framework for resolving community structure in complex networks,” *Proceedings of the National Academy of Sciences* **104**, 7327–7331 (2007).
- [23] Jake M. Hofman and Chris H. Wiggins, “Bayesian Approach to Network Modularity,” *Physical Review Letters* **100**, 258701 (2008).
- [24] Edoardo M. Airolidi, David M. Blei, Stephen E. Fienberg, and Eric P. Xing, “Mixed Membership Stochastic Blockmodels,” *J. Mach. Learn. Res.* **9**, 1981–2014 (2008).
- [25] 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).
- [26] M. Mørup, M. N. Schmidt, and Lars Kai Hansen, “Infinite multiple membership relational modeling for complex networks,” in *2011 IEEE International Workshop on Machine Learning for Signal Processing* (2011) pp. 1–6.
- [27] Jörg Reichardt, Roberto Alamino, and David Saad, “The Interplay between Microscopic and Mesoscopic Structures in Complex Networks,” *PLoS ONE* **6**, e21282 (2011).
- [28] Morten Mørup and Mikkel N. Schmidt, “Bayesian Community Detection,” *Neural Computation* **24**, 2434–2456 (2012).
- [29] Tiago P. Peixoto, “Parsimonious Module Inference in Large Networks,” *Physical Review Letters* **110**, 148701 (2013).
- [30] M.N. Schmidt and M. Morup, “Nonparametric Bayesian Modeling of Complex Networks: An Introduction,” *IEEE Signal Processing Magazine* **30**, 110–128 (2013).
- [31] 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).
- [32] Xiaoran Yan, “Bayesian Model Selection of Stochastic Block Models,” arXiv:1605.07057 [cs, stat] (2016), arXiv: 1605.07057.
- [33] M. E. J. Newman and Gesine Reinert, “Estimating the Number of Communities in a Network,” *Physical Review Letters* **117**, 078301 (2016).
- [34] Wenjie Fu, Le Song, and Eric P. Xing, “Dynamic Mixed Membership Blockmodel for Evolving Networks,” in *Proceedings of the 26th Annual International Conference on Machine Learning, ICML ’09* (ACM, New York, NY, USA, 2009) pp. 329–336.
- [35] K.S. Xu and A.O. Hero, “Dynamic Stochastic Blockmodels for Time-Evolving Social Networks,” *IEEE Journal of Selected Topics in Signal Processing* **8**, 552–562 (2014).
- [36] Tiago P. Peixoto and Martin Rosvall, “Modelling sequences and temporal networks with dynamic community structures,” *Nature Communications* **8**, 582 (2017).
- [37] Leto Peel and Aaron Clauset, “Detecting Change Points in the Large-Scale Structure of Evolving Networks,” in *Twenty-Ninth AAAI Conference on Artificial Intelligence* (2015).
- [38] Amir Ghasemian, Pan Zhang, Aaron Clauset, Cristopher Moore, and Leto Peel, “Detectability Thresholds and Optimal Algorithms for Community Structure in Dynamic Networks,” *Physical Review X* **6**, 031005 (2016).
- [39] Xiao Zhang, Cristopher Moore, and Mark E. J. Newman, “Random graph models for dynamic networks,” *The European Physical Journal B* **90**, 200 (2017).
- [40] Marco Corneli, Pierre Latouche, and Fabrice Rossi, “Exact ICL maximization in a non-stationary



- temporal extension of the stochastic block model for dynamic networks,” *Neurocomputing Advances in artificial neural networks, machine learning and computational intelligence Selected papers from the 23rd European Symposium on Artificial Neural Networks (ESANN 2015)*, **192**, 81–91 (2016).
- [41] Matias Catherine and Miele Vincent, “Statistical clustering of temporal networks through a dynamic stochastic block model,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79**, 1119–1141 (2016).
- [42] N. J. A. Sloane, *The on-line encyclopedia of integer sequences: A000670* (2003).
- [43] N. J. A. Sloane, *The on-line encyclopedia of integer sequences: A008277* (2003).
- [44] David J. C. MacKay, *Information Theory, Inference and Learning Algorithms*, first edition ed. (Cambridge University Press, 2003).
- [45] J. Rissanen, “Modeling by shortest data description,” *Automatica* **14**, 465–471 (1978).
- [46] Peter D. Grünwald, *The Minimum Description Length Principle* (The MIT Press, 2007).
- [47] Yurii Mikhailovich Shtar’kov, “Universal sequential coding of single messages,” *Problemy Peredachi Informatsii* **23**, 3–17 (1987).
- [48] Peter Grünwald, “A tutorial introduction to the minimum description length principle,” arXiv:math/0406077 (2004).
- [49] Gideon Schwarz, “Estimating the Dimension of a Model,” *The Annals of Statistics* **6**, 461–464 (1978), mathematical Reviews number (MathSciNet): MR468014; Zentralblatt MATH identifier: 0379.62005.
- [50] H. Akaike, “A new look at the statistical model identification,” *IEEE Transactions on Automatic Control* **19**, 716–723 (1974).
- [51] 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,” *Journal of Statistical Mechanics: Theory and Experiment* **2014**, P05007 (2014).
- [52] Thomas M. Cover and Joy A. Thomas, *Elements of Information Theory*, 99th ed. (Wiley-Interscience, 1991).
- [53] 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).
- [54] T S Evans, “Clique graphs and overlapping communities,” *Journal of Statistical Mechanics: Theory and Experiment* **2010**, P12037 (2010).
- [55] T S Evans, “American College Football Network Files,” FigShare (2012), 10.6084/m9.figshare.93179.
- [56] Leto Peel, Daniel B. Larremore, and Aaron Clauset, “The ground truth about metadata and community detection in networks,” *Science Advances* **3**, e1602548 (2017).
- [57] M. E. J. Newman and Aaron Clauset, “Structure and inference in annotated networks,” *Nature Communications* **7**, 11863 (2016).
- [58] Darko Hric, Tiago P. Peixoto, and Santo Fortunato, “Network Structure, Metadata, and the Prediction of Missing Nodes and Annotations,” *Physical Review X* **6**, 031038 (2016).
- [59] Y. X. Rachel Wang and Peter J. Bickel, “Likelihood-based model selection for stochastic block models,” *The Annals of Statistics* **45**, 500–528 (2017).
- [60] Santo Fortunato and Marc Barthélemy, “Resolution limit in community detection,” *Proceedings of the National Academy of Sciences* **104**, 36–41 (2007).
- [61] Tiago P. Peixoto, “Hierarchical Block Structures and High-Resolution Model Selection in Large Networks,” *Physical Review X* **4**, 011047 (2014).
- [62] D. Holten, “Hierarchical Edge Bundles: Visualization of Adjacency Relations in Hierarchical Data,” *IEEE Transactions on Visualization and Computer Graphics* **12**, 741–748 (2006).

- [63] Benjamin H. Good, Yves-Alexandre de Montjoye, and Aaron Clauset, “Performance of modularity maximization in practical contexts,” *Physical Review E* **81**, 046106 (2010).
- [64] Darko Hric, Richard K. Darst, and Santo Fortunato, “Community detection in networks: Structural communities versus ground truth,” *Physical Review E* **90**, 062805 (2014).
- [65] Harold Jeffreys, *Theory of Probability*, Auflage: third. ed. (Oxford University Press, Oxford Oxfordshire : New York, 2000).
- [66] 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.
- [67] Roger Guimerà and Luís A. Nunes Amaral, “Functional cartography of complex metabolic networks,” *Nature* **433**, 895–900 (2005).
- [68] Aaron Clauset, Cristopher Moore, and M. E. J. Newman, “Hierarchical structure and the prediction of missing links in networks,” *Nature* **453**, 98–101 (2008).
- [69] Diego Garlaschelli, Frank den Hollander, and Andrea Roccaverde, “Ensemble nonequivalence in random graphs with modular structure,” *Journal of Physics A: Mathematical and Theoretical* **50**, 015001 (2017).
- [70] Brian Ball, Brian Karrer, and M. E. J. Newman, “Efficient and principled method for detecting communities in networks,” *Physical Review E* **84**, 036103 (2011).
- [71] V Krebs, “Political Books Network,” unpublished, retrieved from Mark Newman’s website: <http://www-personal.umich.edu/~mejn/netdata/>.
- [72] Yong-Yeol Ahn, James P. Bagrow, and Sune Lehmann, “Link communities reveal multiscale complexity in networks,” *Nature* **466**, 761–764 (2010).
- [73] Tiago P. Peixoto, “Inferring the mesoscale structure of layered, edge-valued, and time-varying networks,” *Physical Review E* **92**, 042807 (2015).
- [74] Christopher Aicher, Abigail Z. Jacobs, and Aaron Clauset, “Learning latent block structure in weighted networks,” *Journal of Complex Networks*, cnu026 (2014).
- [75] Tiago P. Peixoto, “Nonparametric weighted stochastic block models,” *Physical Review E* **97**, 012306 (2018).
- [76] N. Stanley, S. Shai, D. Taylor, and P. J. Mucha, “Clustering Network Layers with the Strata Multi-layer Stochastic Block Model,” *IEEE Transactions on Network Science and Engineering* **3**, 95–105 (2016).
- [77] Subhadeep Paul and Yuguo Chen, “Consistent community detection in multi-relational data through restricted multi-layer stochastic blockmodel,” *Electronic Journal of Statistics* **10**, 3807–3870 (2016).
- [78] Toni Vallès-Català, Francesco A. Massucci, Roger Guimerà, and Marta Sales-Pardo, “Multilayer Stochastic Block Models Reveal the Multilayer Structure of Complex Networks,” *Physical Review X* **6**, 011036 (2016).
- [79] Caterina De Bacco, Eleanor A. Power, Daniel B. Larremore, and Cristopher Moore, “Community detection, link prediction, and layer interdependence in multilayer networks,” *Physical Review E* **95**, 042317 (2017).
- [80] Leto Peel, “Active discovery of network roles for predicting the classes of network nodes,” *Journal of Complex Networks* **3**, 431–449 (2015).
- [81] Travis Martin, Brian Ball, and M. E. J. Newman, “Structural inference for uncertain networks,” *Physical Review E* **93**, 012306 (2016).
- [82] M. E. J. Newman and Tiago P. Peixoto, “Generalized Communities in Networks,” *Physical Review Letters* **115**, 088701 (2015).
- [83] M. E. J. Newman and G. T. Barkema, *Monte Carlo Methods in Statistical Physics* (Oxford University

- Press, U.S.A., Oxford : New York, 1999).
- [84] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller, “Equation of State Calculations by Fast Computing Machines,” *The Journal of Chemical Physics* **21**, 1087 (1953).
  - [85] W. K. Hastings, “Monte Carlo sampling methods using Markov chains and their applications,” *Biometrika* **57**, 97–109 (1970).
  - [86] S. Kirkpatrick, C. D. Gelatt Jr, and M. P. Vecchi, “Optimization by simulated annealing,” *Science* **220**, 671 (1983).
  - [87] Prem K. Gopalan and David M. Blei, “Efficient discovery of overlapping communities in massive networks,” *Proceedings of the National Academy of Sciences* **110**, 14534–14539 (2013).
  - [88] Maria A. Riolo, George T. Cantwell, Gesine Reinert, and M. E. J. Newman, “Efficient method for estimating the number of communities in a network,” *Physical Review E* **96**, 032310 (2017).
  - [89] Tiago P. Peixoto, “Efficient Monte Carlo and greedy heuristic for the inference of stochastic block models,” *Physical Review E* **89**, 012804 (2014).
  - [90] J. Kiefer, “Sequential Minimax Search for a Maximum,” *Proceedings of the American Mathematical Society* **4**, 502 (1953).
  - [91] Tiago P. Peixoto, “The graph-tool python library,” figshare (2014), 10.6084/m9.figshare.1164194, available at <https://graph-tool.skewed.de>.
  - [92] Christopher Moore, “The Computer Science and Physics of Community Detection: Landscapes, Phase Transitions, and Hardness,” arXiv:1702.00467 [cond-mat, physics:physics] (2017), arXiv: 1702.00467.
  - [93] Lenka Zdeborová and Florent Krzakala, “Statistical physics of inference: thresholds and algorithms,” *Advances in Physics* **65**, 453–552 (2016).
  - [94] Trevor F. Cox and M. A. A. Cox, *Multidimensional Scaling, Second Edition*, 2nd ed. (Chapman and Hall/CRC, Boca Raton, 2000).
  - [95] Wayne W. Zachary, “An Information Flow Model for Conflict and Fission in Small Groups,” *Journal of Anthropological Research* **33**, 452–473 (1977).
  - [96] M. E. J. Newman, “Finding community structure in networks using the eigenvectors of matrices,” *Physical Review E* **74**, 036104 (2006).
  - [97] Toni Vallès-Català, Tiago P. Peixoto, Roger Guimerà, and Marta Sales-Pardo, “On the consistency between model selection and link prediction in networks,” arXiv:1705.07967 [cond-mat, stat] (2017), arXiv: 1705.07967.
  - [98] Roger Guimerà and Marta Sales-Pardo, “A Network Inference Method for Large-Scale Unsupervised Identification of Novel Drug-Drug Interactions,” *PLoS Comput Biol* **9**, e1003374 (2013).
  - [99] Núria Rovira-Asenjo, Tània Gumí, Marta Sales-Pardo, and Roger Guimerà, “Predicting future conflict between team-members with parameter-free models of social networks,” *Scientific Reports* **3** (2013), 10.1038/srep01999.
  - [100] Roger Guimerà, Alejandro Llorente, Esteban Moro, and Marta Sales-Pardo, “Predicting Human Preferences Using the Block Structure of Complex Social Networks,” *PLoS ONE* **7**, e44620 (2012).
  - [101] Antonia Godoy-Lorite, Roger Guimerà, Christopher Moore, and Marta Sales-Pardo, “Accurate and scalable social recommendation using mixed-membership stochastic block models,” *Proceedings of the National Academy of Sciences* **113**, 14207–14212 (2016).
  - [102] Aurelien Decelle, Florent Krzakala, Christopher Moore, and Lenka Zdeborová, “Inference and Phase Transitions in the Detection of Modules in Sparse Networks,” *Physical Review Letters* **107**, 065701 (2011).
  - [103] Aurelien Decelle, Florent Krzakala, Christopher Moore, and Lenka Zdeborová, “Asymptotic analy-

- sis of the stochastic block model for modular networks and its algorithmic applications,” *Physical Review E* **84**, 066106 (2011).
- [104] F. Y. Wu, “The Potts model,” *Reviews of Modern Physics* **54**, 235–268 (1982).
- [105] Marc Mezard and Andrea Montanari, *Information, Physics, and Computation* (Oxford University Press, 2009).
- [106] M. Mezard, *Spin Glass Theory And Beyond: An Introduction To The Replica Method And Its Applications* (Wspc, Singapore ; New Jersey, 1986).
- [107] M. E Dyer and A. M Frieze, “The solution of some random NP-hard problems in polynomial expected time,” *Journal of Algorithms* **10**, 451–489 (1989).
- [108] Andrea Lancichinetti, Santo Fortunato, and Filippo Radicchi, “Benchmark graphs for testing community detection algorithms,” *Physical Review E* **78**, 046110 (2008).
- [109] Andrea Lancichinetti and Santo Fortunato, “Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities,” *Physical Review E* **80**, 016118 (2009).
- [110] Jörg Reichardt and Michele Leone, “(Un)detectable Cluster Structure in Sparse Networks,” *Physical Review Letters* **101**, 078701 (2008).
- [111] Peter Ronhovde and Zohar Nussinov, “Multiresolution community detection for megascale networks by information-based replica correlations,” *Physical Review E* **80**, 016109 (2009).
- [112] Emmanuel Abbe, “Community detection and stochastic block models: recent developments,” arXiv:1703.10146 [cs, math, stat] (2017), arXiv: 1703.10146.
- [113] Sofia C. Olhede and Patrick J. Wolfe, “Network histograms and universality of blockmodel approximation,” *Proceedings of the National Academy of Sciences* **111**, 14722–14727 (2014).