Nonparametric weighted stochastic block models

Tiago P. Peixoto*

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

We present a Bayesian formulation of weighted stochastic block models that can be used to infer the large-scale modular structure of weighted networks, including their hierarchical organization. Our method is nonparametric, and thus does not require the prior knowledge of the number of groups or other dimensions of the model, which are instead inferred from data. We give a comprehensive treatment of different kinds of edge weights (i.e. continuous or discrete, signed or unsigned, bounded or unbounded), as well as arbitrary weight transformations, and describe an unsupervised model selection approach to choose the best network description. We illustrate the application of our method to a variety of empirical weighted networks, such as global migrations, voting patterns in congress, and neural connections in the human brain.

I. INTRODUCTION

Many network systems lack a natural low-dimensional embedding from which we can readily extract their most prominent large-scale features. Instead, we have to infer this information from data, typically by decomposing the observed network into modules [1]. A principled approach to perform this task is to formulate generative models that allow this modular decomposition to be found via statistical inference [2]. The most fundamental model used for this purpose is the stochastic block model (SBM) [3], which groups nodes according to their probabilities of connection to the rest of the network. However, a central limitation of most SBM implementations is that they are defined strictly for simple or multigraphs. This means that they do not incorporate extra information on the edges, which are typically present in a variety of systems, and are required for an accurate representation of their structure. For example, to the existence of a route between two airports is associated a distance, to the biomass flow between two species in a food web is associated a flow magnitude, etc. In this work, we develop variations of the SBM that allow for this type of information on the edges to be incorporated into the network model and guide the partition of the nodes into groups in a statistically meaningful way.

We follow the same basic idea put forth by Aicher et al. [4], who adapted the SBM to weighted networks by including edge values as additional covariates. However, our approach diverges from Ref. [4] in key aspects. First, here we develop a nonparametric Bayesian approach, based on exact integrated likelihoods, that is capable of inferring the dimension of the model — e.g. the number of groups — from the data itself, without requiring it to be known *a priori*. This is achieved by departing from the canonical exponential family of distributions, and using instead *microcanonical* formulations that are easier to compute exactly and approach the canonical distributions asymptotically. Second, our approach also infers the hierarchical modular structure of the network, extending the nested SBM of Refs. [5, 6] to the weighted case. The hierarchical nature of the model is implemented via structured Bayesian priors that have been shown to significantly decrease the tendency of the nonparametric approach to underfit [7] and are capable of uncovering small but statistically significant modules in large networks [5, 6]. And third, our approach is efficient, making use of MCMC sampling that requires only $O(E)$ operations per sweep, where E is the number of edges in the network, independently of the number of groups.

This paper is organized as follows. In Sec. II we present our general approach, and in Sec. III we illustrate its use in a variety of empirical weighted network datasets. In Sec. IV we elaborate on the diverse models for edge weights based on basic properties (such as whether they are discrete or continuous, signed or unsigned, bounded or unbounded), show how these models can be extended via weight transformations, and how different models can be chosen via Bayesian model selection. We finalize in Sec. V with a discussion.

II. WEIGHTED SBMS VIA EDGE COVARIATES

We consider generative models for networks that, in addition to the adjacency matrix $\boldsymbol{A} = \{A_{ij}\}$, also possess real or discrete edge covariates $\boldsymbol{x} = \{\boldsymbol{x}_{ij}\}$ on the edges. Without loss of generality, here we assume that the networks are multigraphs, i.e. $A_{ij} \in \mathbb{N}_0$, such that $\boldsymbol{x}_{ij}$ is a vector containing one weight for each parallel edge between nodes i and j, and no weights if $A_{ij} = 0$. Furthermore, we assume that the edge existence is decoupled from its weight, i.e. the non-existence of an edge is different from an edge with zero weight (the special case where the zeros of the adjacency matrix are considered values of the edge covariates can be recovered by using a complete graph in place of $\boldsymbol{A}$, and adapting $\boldsymbol{x}$ accordingly). As done in Ref. [4], we follow the underlying assumption of the SBM that the nodes are divided into B groups, with $b_i \in \{1, \dots, B\}$ specifying the group membership of node

* t.peixoto@bath.ac.uk

i, and where in addition to the edge placement, the edge weights are sampled only according to the group memberships of their endpoints. Concretely, this means they are both sampled from parametric distributions that are conditioned only on the group memberships of the nodes i.e.

$$P(\boldsymbol{A},\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{b}) = P(\boldsymbol{x}|\boldsymbol{A},\boldsymbol{\gamma},\boldsymbol{b})P(\boldsymbol{A}|\boldsymbol{\theta},\boldsymbol{b}) \tag{1}$$

with the covariates being sampled only on existing edges,

$$P(\boldsymbol{x}|\boldsymbol{A},\boldsymbol{\gamma},\boldsymbol{b}) = \prod_{r\le s} P(\boldsymbol{x}_{rs}|\boldsymbol{\gamma}_{rs}) \tag{2}$$

with $\boldsymbol{x}_{rs} = \{\boldsymbol{x}_{ij}|A_{ij} > 0 \wedge (b_i, b_j) = (r,s)\}$ being the covariates between groups r and s, and where $\boldsymbol{\gamma}_{rs}$ is a set of parameters that govern the sampling of the weights between groups r and s. The placement of the edges is done independently of the weights by choosing any SBM flavor with parameters $\boldsymbol{\theta}$; For example, with the degree-corrected SBM [8] we would have

$$P(\boldsymbol{A}|\boldsymbol{\theta}=\{\boldsymbol{\lambda},\boldsymbol{\kappa}\},\boldsymbol{b}) = \prod_{i<j}\frac{\mathrm{e}^{-\lambda_{b_i,b_j}\kappa_i\kappa_j}(\lambda_{b_i,b_j}\kappa_i\kappa_j)^{A_{ij}}}{A_{ij}!}, \tag{3}$$

where λ_{rs} controls the number of edges that are placed between groups, and κ_i the expected degree of node i.

Given the generative model above, we could proceeded by estimating the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ via maximum likelihood. However, doing so would be subject to overfitting, as the likelihood would increase monotonically with the complexity of the model. Instead, here we are interested in solving a more general and arguably more well-posed problem, namely to obtain the Bayesian posterior probability of partitions, in a *nonparametric* manner, taking into account only the weighted network,

$$P(\boldsymbol{b}|\boldsymbol{A},\boldsymbol{x}) = \frac{P(\boldsymbol{A},\boldsymbol{x}|\boldsymbol{b})P(\boldsymbol{b})}{P(\boldsymbol{A},\boldsymbol{x})}, \tag{4}$$

where the numerator contains the marginal likelihood integrated over the model parameters

$$\begin{aligned}P(\boldsymbol{A},\boldsymbol{x}|\boldsymbol{b}) &= \int P(\boldsymbol{A},\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{b})P(\boldsymbol{\theta})P(\boldsymbol{\gamma})\,\mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{\gamma}\\ &= P(\boldsymbol{A}|\boldsymbol{b})P(\boldsymbol{x}|\boldsymbol{A},\boldsymbol{b}),\end{aligned} \tag{5}$$

where

$$P(\boldsymbol{A}|\boldsymbol{b}) = \int P(\boldsymbol{A}|\boldsymbol{\theta},\boldsymbol{b})P(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} \tag{6}$$

is the marginal likelihood of the unweighted network integrated over the relevant parameters. Integrated marginal likelihoods of this kind were considered in numerous works for several unweighted model variants [5–7, 9–12]. In this work, our approach is fully independent of any particular choice made for this part of the model. However, in our experiments we will use the nested microcanonical degree-corrected SBM described in Ref. [5, 6], due to its efficient and multi-scale nature, as well as a much reduced tendency to underfit when used with large networks. Furthermore, its hierarchical nature will allow us to describe summaries of the network — taking into accounts its edge covariates — at multiple scales, allowing a bird's-eye view of large datasets. We use this model without sacrificing generality, since the usual nonhierarchical SBM amounts exactly to using the nested version with just one hierarchical level.

The crucial part in Eq. 5 that completes our nonparametric approach is the marginal likelihood of the edge weights, which is integrated over the weight parameters $\boldsymbol{\gamma}$ according to their prior distribution $P(\boldsymbol{\gamma}_{rs})$, which is the same for every pair of groups r and s,

$$\begin{aligned}P(\boldsymbol{x}|\boldsymbol{A},\boldsymbol{b}) &= \int P(\boldsymbol{x}|\boldsymbol{A},\boldsymbol{\gamma},\boldsymbol{b})P(\boldsymbol{\gamma})\,\mathrm{d}\boldsymbol{\gamma}\\ &= \prod_{r\le s}\int P(\boldsymbol{x}_{rs}|\boldsymbol{\gamma}_{rs})P(\boldsymbol{\gamma}_{rs})\,\mathrm{d}\boldsymbol{\gamma}_{rs}.\end{aligned} \tag{7}$$

The form of the prior distribution $P(\boldsymbol{\gamma})$ is usually conditioned on hyperparameters $\boldsymbol{\eta}$, which represent our *a priori* assumptions about the data. In order for our inference approach to retain its nonparametric character, we need these hyperparameters to take a single global value, i.e. $P(\boldsymbol{\gamma}_{rs}) = P(\boldsymbol{\gamma}_{rs}|\boldsymbol{\eta})$ for all groups r and s. Alternatively, we may treat $\boldsymbol{\eta}$ as latent variables, and sample them from their own distribution, $P(\boldsymbol{\eta})$, thereby reducing the sensitivity to our *a priori* assumptions. This idea fits well with the nested version of the SBM we will be using [5, 6], which, as part of its prior probabilities, considers the groups themselves as nodes of a smaller multigraph that is also generated by the SBM, with its nodes put in their own groups, forming an even smaller multigraph, and so on recursively, following a nested hierarchy $\{\boldsymbol{b}^l\} = \{\{b_r^{(l)}\}_l\}$, so that $b_r^{(l)} \in \{1,\dots,B_l\}$ is the group membership of group/node r at the hierarchy level $l \in \{1,\dots,L\}$, with the boundary condition that the number of groups at the topmost level $l = L$ is $B_L = 1$. Therefore, the adjacency of the multigraph at level l is

$$m^l_{rs} = \sum_{tu}\frac{m^{l-1}_{tu}\delta_{b_t^{(l)},r}\delta_{b_u^{(l)},s}}{\delta_{rs}+1}, \tag{8}$$

where we assume $m^0_{ij} = A_{ij}$. Following the same logic, we may consider the parameters $\boldsymbol{\gamma}$ as edge covariates in the multigraph of groups, which themselves are generated by another model in a level above, and so on. We may thus let $\boldsymbol{\gamma}^1 \equiv \boldsymbol{\gamma}$ and $\boldsymbol{\gamma}^2 \equiv \boldsymbol{\eta}$ be the first two levels of a hierarchical model, given recursively by

$$P(\boldsymbol{\gamma}^l|\boldsymbol{A},\boldsymbol{b}^{l+1},\boldsymbol{\gamma}^{l+1}) = \prod_{t\le u}P(\boldsymbol{\gamma}^l_{tu}|\boldsymbol{A},\boldsymbol{b}^{l+1},\boldsymbol{\gamma}^{l+1}_{b_t^{(l+1)},b_u^{(l+1)}}), \tag{9}$$

where $\boldsymbol{\gamma}^{l+1}_{tu} = \{\boldsymbol{\gamma}^l_{rs}|m^l_{rs} > 0 \wedge (b_r^{(l+1)}, b_s^{(l+1)}) = (t,u)\}$ are the hyperparameters between groups (t,u) at level $l+1$, with m^l_{rs} given by Eq. 8. The final model is then obtained by integrating over the entire hierarchy,

$$P(\boldsymbol{x}|\boldsymbol{A},\{\boldsymbol{b}^l\}) = \int P(\boldsymbol{x}|\boldsymbol{A},\boldsymbol{\gamma}^1,\boldsymbol{b}^1)\prod_{l=1}^{L}P(\boldsymbol{\gamma}^l|\boldsymbol{A},\boldsymbol{\gamma}^{l+1},\boldsymbol{b}^{l+1})\,\mathrm{d}\boldsymbol{\gamma}^l, \quad (10)$$

assuming the boundary condition $\boldsymbol{\gamma}^{L+1} = \{\hat{\gamma}\}$, such that $\hat{\gamma}$ is a single set of hyperparameters that are left out of the integration at the topmost level, reflecting only global aspects of the covariates, without a significant effect on the model structure and dimension. Instead of defining a unique model, we will consider a variety of elementary choices for $P(\boldsymbol{x}|\boldsymbol{\gamma})$ and $P(\boldsymbol{\gamma})$ that reflect the precise nature of the covariates (e.g. continuous or discrete, signed or unsigned, bounded or unbounded), and for which Eq. 10 can be computed exactly. In particular, we will make use of microcanonical formulations of the weight distributions that permit the straightforward computation of the integrals, without sacrificing descriptive power. We leave the derivations of the likelihood expressions for Sec. IV, and we proceed with a general outline, and an analysis of this approach for empirical networks.

When using the nested model, we have a posterior distribution over hierarchical partitions,

$$P(\{\boldsymbol{b}^l\}|\boldsymbol{A},\boldsymbol{x}) = \frac{P(\boldsymbol{A},\boldsymbol{x}|\{\boldsymbol{b}^l\})P(\{\boldsymbol{b}^l\})}{P(\boldsymbol{A},\boldsymbol{x})}, \quad (11)$$

which can be marginalized, if we so desire, to obtain only the partition at the bottom level $\boldsymbol{b} \equiv \boldsymbol{b}^1$,

$$P(\boldsymbol{b}|\boldsymbol{A},\boldsymbol{x}) = \sum_{\{\boldsymbol{b}^l\,|\,l>1\}} P(\{\boldsymbol{b}^l\}|\boldsymbol{A},\boldsymbol{x}). \quad (12)$$

However, most typically we will want to obtain the entire hierarchical partition, as it is useful for a multilevel description of the data. Since the posterior in Eq. 11 involves a prior probability of the partition $P(\{\boldsymbol{b}^l\})$ (described in detail in Ref. [6]), and is integrated over all remaining model parameters, it possesses an inherent *regularization* property, where overly complicated models are penalized with a lower posterior probability [13]. This means that, differently from maximum likelihood approaches, we can infer properties related to the size of the model, such as the number of groups and hierarchy depth, without danger of overfitting. Furthermore, as we detail further in Sec. IV G, the posterior distribution gives us a principled means of model selection according to statistical significance, which allows us to choose the most appropriate weight model.

Given a choice for the parametric model for weights, we compute Eq. 10, which allows us to determine the posterior distribution of the partitions in Eq. 11 up to the normalizing constant $P(\boldsymbol{A},\boldsymbol{x})$ in the denominator, which is generally intractable. But since we cannot sample from the posterior distribution directly even if we *could* somehow compute this constant, we must resort to MCMC importance sampling methods, for which this normalizing constant is luckily not needed. This is generally implemented by making move proposals $\{\boldsymbol{b}^l\} \to \{\boldsymbol{b}^l\}'$ with probability $P(\{\boldsymbol{b}^l\}'|\{\boldsymbol{b}^l\})$, and rejecting the proposal with probability $1-a$, where a is the Metropolis-Hastings [14, 15] criterion

$$a = \min\left(1, \frac{P(\{\boldsymbol{b}^l\}'|\boldsymbol{A},\boldsymbol{x})}{P(\{\boldsymbol{b}^l\}|\boldsymbol{A},\boldsymbol{x})}\frac{P(\{\boldsymbol{b}^l\}|\{\boldsymbol{b}^l\}')}{P(\{\boldsymbol{b}^l\}'|\{\boldsymbol{b}^l\})}\right). \quad (13)$$

Since the ratio in Eq. 13 does not depend on the normalization constant $P(\boldsymbol{A},\boldsymbol{x})$, the value of a can be computed exactly, and — as long as the move proposals are ergodic — the algorithm above will eventually sample partitions from the desired posterior distribution asymptotically. We can also obtain the *most likely* hierarchical partition,

$$\{\boldsymbol{b}^l\}^* = \underset{\{\boldsymbol{b}^l\}}{\operatorname{argmax}}\, P(\{\boldsymbol{b}^l\}|\boldsymbol{A},\boldsymbol{x}) \quad (14)$$

by replacing $P(\{\boldsymbol{b}^l\}|\boldsymbol{A},\boldsymbol{x}) \to P(\{\boldsymbol{b}^l\}|\boldsymbol{A},\boldsymbol{x})^\beta$ in Eq. 13 and making $\beta \to \infty$ in slow increments. Therefore, we can both maximize and sample from the posterior distribution, using the same algorithm. In this work we use the move proposals defined in Ref. [16] that use the local information of a node's neighbourhood to improve equilibration speed, as well as the agglomerative initialization heuristic described in the same reference and extended to the nested model in Ref. [5]. The combination of these move proposals with the likelihood of the microcanonical SBM of Ref. [6], as well as any of the weight likelihoods defined in Sec. IV, yields an algorithm where each MCMC sweep (i.e. for every node one move is attempted) is performed in time $O(E)$, independently of how many groups are occupied with nodes. For more details of the algorithm we defer to Ref. [6] and to the freely available C++ implementation in the `graph-tool` Python library [17].

III. EMPIRICAL NETWORKS

A. Migrations between countries

We begin with an illustration of how incorporating edge weights with our method can have a significant effect on the analysis of network data. We use for this purpose a dataset of global migrations between $N = 232$ countries, assembled in 2015 by the United Nations [19]. This dataset can be represented as a directed network (see Appendix A), where for a pair of countries (i,j) there is a net migrant stock $x_{ij} \in \mathbb{Z}$ which is defined as the number of migrants that moved from i to j minus the number that moved from j to i. If we only had an unweighted SBM at our disposal, a common approach would be to threshold this data, yielding a directed edge $A_{ij} = 1$ if $x_{ij} > 0$ and $A_{ij} = 0$ otherwise. As argued by Aicher et al [4], this type of data manipulation should be avoided whenever possible, since not only it destroys potentially valuable

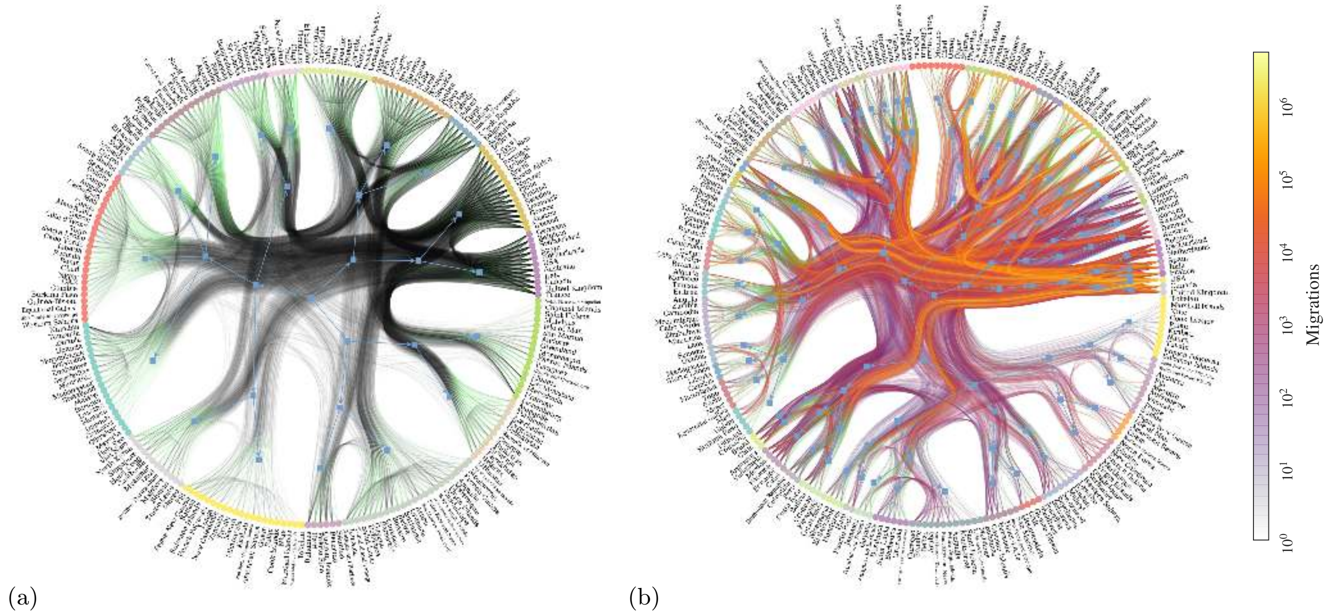

Figure 1. (a) Fit of the unweighted SBM for UN migration data, using the threshold approach described in the text. The edges are routed according the inferred hierarchy (shown in blue), using an edge-bundling algorithm by Holten [18], and the edge sources are marked with a green color. (b) Fit of the weighted SBM for the same data with the migrant stocks included, as shown by the edge colors and in the legend.

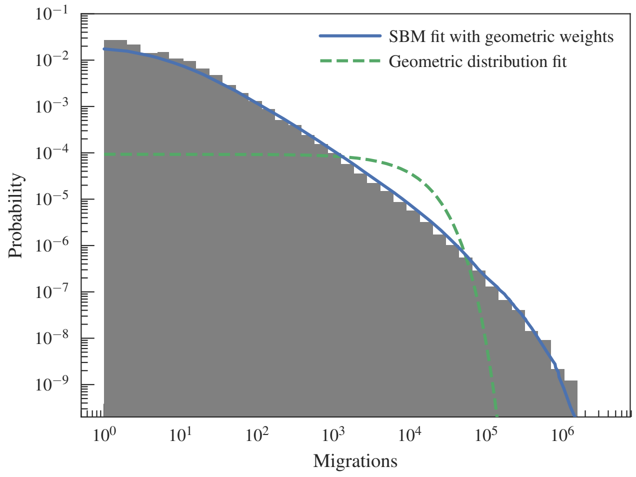

Figure 2. Overall distribution of the number of migrations for the UN data. The solid line shows the inferred distribution according to the weighted SBM using geometric distributions. The dashed line shows the best fit of a single geometric distribution.

information, but also it is possible to construct examples where no single threshold can accurately reproduce the large-scale structure in the data. In this particular case, this approach actually does seem to yield usable information at first, as can be seen in Fig. 1a, which shows a fit of the unweighted SBM. We can see that the network division obtained in this manner essentially categorizes countries on whether they are net sources or targets of migration, as well as the typical regions people migrate to and from. However, a closer inspection reveals that it is not able to distinguish between countries like Costa Rica, South Africa and Finland (which end up clustered in the same group as Austria and Ireland), which not only are geographically far apart, but do have, in fact, very distinct migration volumes and patterns. Since migration volumes between countries can vary by several orders of magnitude (see Fig. 2), any analysis that ignores this aspect must be woefully incomplete. Indeed, if we include the values x_{ij} of the migrant stock, in addition to the same adjacency matrix obtained with the threshold approach, and we use the weighted SBM defined previously, with geometric distributions for the weights as described in Sec IV C, we obtain a much more detailed representation of the data, as shown in Fig. 1b. Not only we find a larger number of groups, but now countries like France, Canada and United Kingdom appear as members of very specific groups. The United States of America gets placed in its own group, due its unique volume and pattern of (mostly incoming) migrations. The remaining countries end up divided in geographically meaningful categories, with regions like South America, Middle East, Africa and Asia being easily recognizable. However, there are exceptions to this, where geographically separated countries get clustered together. Examples of this include Germany and India, as well as China and South Africa. These countries are either sources or targets of global migration which goes well beyond their immediate neighborhoods, and they possess similar overall patterns despite geographical distance (we emphasize that due to the degree-corrected nature of our model, countries with

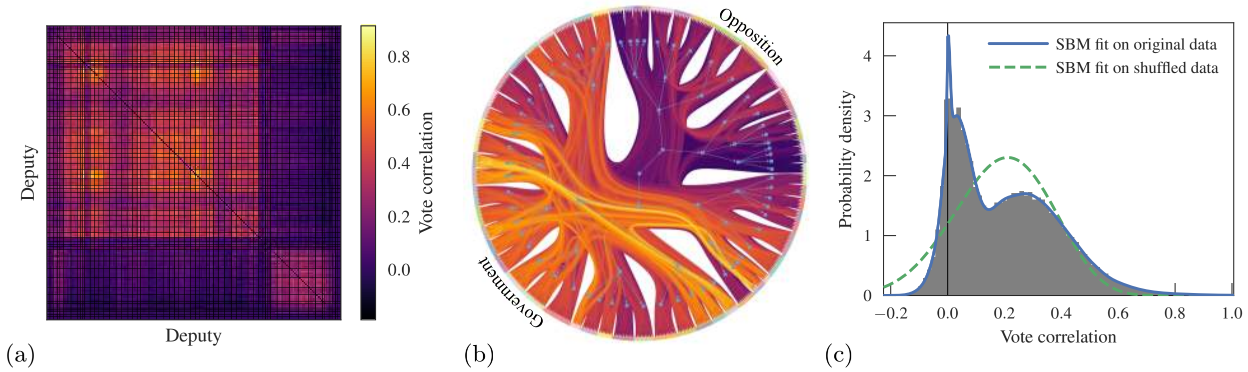

Figure 3. (a) Fit of the weighted SBM for a matrix of vote correlations between deputies of the lower house of the Brazilian congress. The group boundaries are shown by horizontal and vertical lines. (b) Same as in (a) but using the layout of Fig. 1 that shows the entire hierarchical division. (c) Overall distribution of vote correlations. The solid line shows the inferred distribution according to the weighted SBM using transformed normal distributions. The dashed line shows the best fit using the same model, but on the shuffled data with the same empirical distribution.

distinct migration balances can be put in the same category, if their group affinities are the same).

We can also assess the quality of the SBM in capturing the overall weight distribution, computed from the model as

$$P(x|\boldsymbol{A},\boldsymbol{x},\{\boldsymbol{b}^l\}) = \frac{1}{E}\sum_{r\le s} m^1_{rs} P(x|\bar{\gamma}^1_{rs}), \qquad (15)$$

with $E=\sum_{r\le s} m^1_{rs}$ being the total number of edges, and $P(x|\bar{\gamma}^1_{rs})$ is the marginal covariate distribution between groups r and s, which in this case is given by Eq. 68. From Fig. 2, we see that the overall inferred distribution — which is a particular mixture of geometric distributions — is capable of providing a very good fit of the empirical data, despite the fact it is much broader than any single geometric distribution (a best fit of which is shown for reference).

B. Vote correlations in congress

We move now to another example where methods for unweighted graphs are ill suited. We consider the voting patterns of $N=475$ members of the lower house of the Brazilian congress during 2009 [20]: Each deputy voted "yes" or "no" on proposed laws during the legislative year, and based on this, we computed the normalized correlation between the votes $x_{ij}\in[-1,1]$ of deputies i and j. Note that in this case we have an adjacency matrix which is a complete graph, i.e. $A_{ij}=1$ for all i,j, and any pairs with zero correlation are considered particular values of the covariates.

This time we skip any attempt at thresholding the data, and we move directly to the analysis using the weighted SBM. For this, we use the version with normal distributions described in Sec. IV B, adapted to bounded weights via the variable transformation $y_{ij}=2\,\mathrm{arctanh}(x_{ij})$ that maps the intervals $[-1,1]\to[-\infty,\infty]$, as described in Sec. IV F. As shown in Fig. 3a, the method uncovers many groups of deputies, which collectively can be divided into two overall groups at the highest hierarchical level. These two large groups are more correlated with their own members than with nonmembers. An inspection of the known party affiliations of the deputies reveals that these two overall groups correspond to the government and opposition, which tend to vote together either against or in favor of bills. If we again inspect the overall distribution of vote correlations, we see that the weighted SBM provides a very good fit, as seen in Fig. 3c. The model captures the bimodal nature of the vote correlations — with higher values corresponding to pairs of deputies belonging both to either the government or opposition, and lower values to pairs belonging to different factions. It should be emphasized that the quality of the fit is not merely an outcome of using a sufficiently large mixture of normal distributions, as we are not modelling the overall distribution directly. Instead, the distributions are tied to the division of the nodes into groups, and the quality of the overall fit shows that the distribution of weights is well correlated with this categorization. For comparison, we show in Fig. 3c the outcome of the same analysis where the exact same weights are used, but they are randomly shuffled across pairs of deputies, thereby destroying any group organization but preserving the overall weight distribution. In this case, the best SBM fit is composed of only one group, $B=1$, and the corresponding normal fit cannot capture the bimodal structure of the weights — although it is still present in the shuffled data, albeit in a manner which is completely uncorrelated with any partition of the deputies. Therefore a close match between the empirical weight distribution and the SBM fit like the one in Fig. 3c — as well as the one in Fig. 2 for the UN migration data — is a testament to the quality of the SBM ansatz in explaining the data, rather than of an arbitrary mix of elementary unimodal distributions.

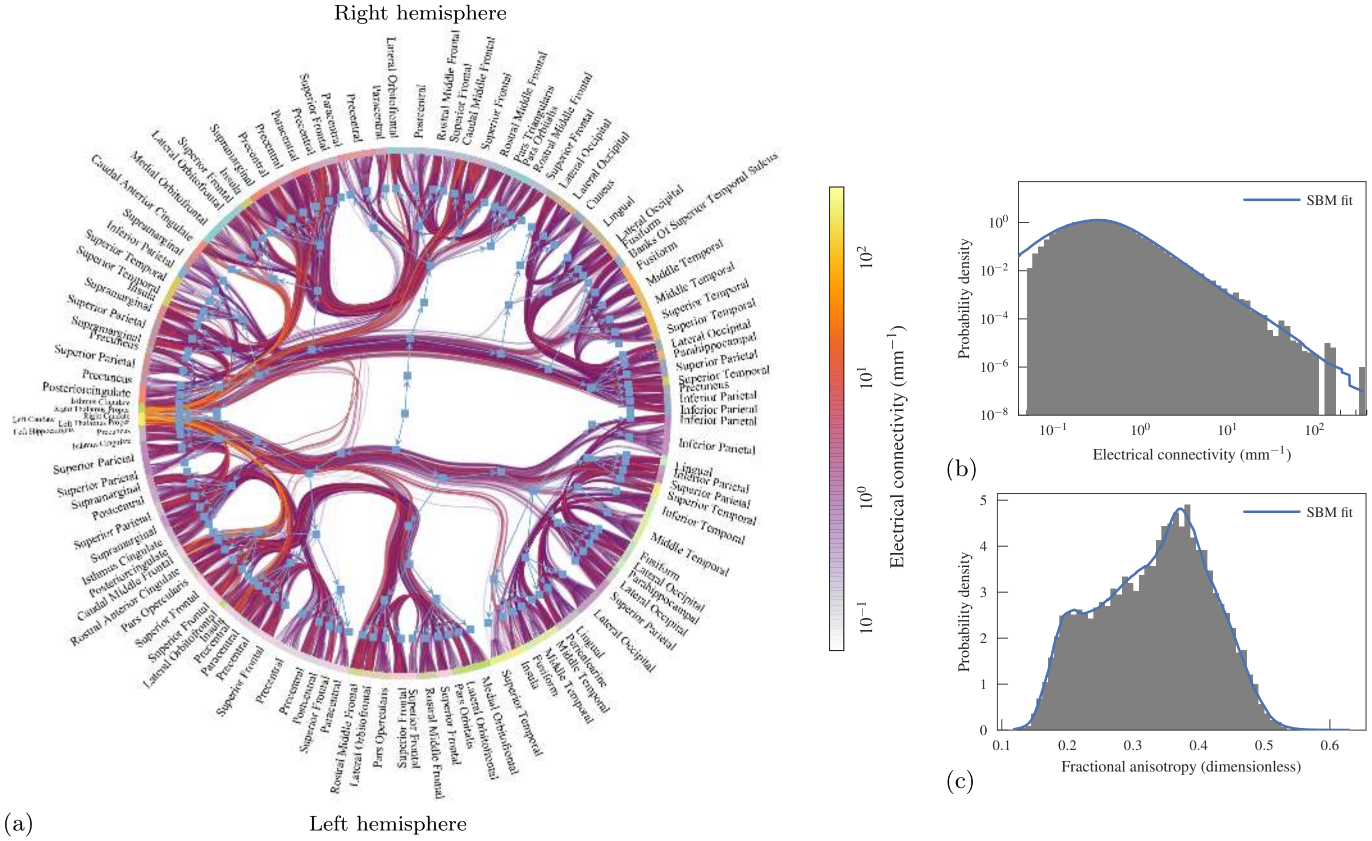

Figure 4. (a) Inferred SBM for the human connectome, using electrical connectivity and fractional anisotropy as edge covariates. The text labels show the most frequent anatomical annotation inside each group at the lowest hierarchical level; (b) Empirical and fitted distribution of electrical connectivity of the edges; (c) Empirical and fitted distribution of fractional anisotropy of the edges.

C. The human brain

We now analyze empirical networks of interactions between parts of human the brain, using data from the Budapest Reference Connectome [21] (which itself is based on primary data from the Human Connectome Project [22]). This dataset corresponds to a consensus between 477 people, where an edge between two of $N = 1,006$ pre-defined anatomical regions is considered to exist, i.e. $A_{ij} = 1$, if neuronal fibers connecting these two regions have been detected in at least 20% of the individuals. In addition to this basic connectivity, we consider two edge covariates, averaged over individuals: The "electrical connectivity" $x_{ij} \in [0, \infty]$, defined as the number of recorded fibers divided by their length, and the fractional anisotropy [23], $y_{ij} \in [0, 1]$, which is maximal if all fibers in the affected region go in the same direction in 3D space, or minimal if they all go in different directions. Indeed, we use this dataset as an opportunity to highlight that our method can also be used when there are multiple covariates available. This can be done in an intuitive manner by assuming that their generation is conditioned on the same network partition, but otherwise are independent, i.e.

$$P(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{A}, \{\boldsymbol{b}^l\}) = P(\boldsymbol{x}|\boldsymbol{A}, \{\boldsymbol{b}^l\})P(\boldsymbol{y}|\boldsymbol{A}, \{\boldsymbol{b}^l\}). \qquad (16)$$

We can then use the exact same algorithm to obtain the posterior $P(\{\boldsymbol{b}^l\}|\boldsymbol{A}, \boldsymbol{x}, \boldsymbol{y})$ by simply combining both terms for $\boldsymbol{x}$ and $\boldsymbol{y}$. This approach will use the information in both covariates simultaneously to inform the partition of the network. (This is easily extended for an arbitrary number of covariates, and hence yields a method that is also suitable for vector-valued covariates, which is supported in our reference implementation [17].) In the following, we use normal models for the transformed covariates $\ln x_{ij}$ and $\text{logit}(y_{ij})$.

When applied to the brain dataset, our method reveals the structure shown in Fig. 4a. It decomposes the network into left and right hemispheres at the topmost hierarchical level, and proceeds to subdivide it into smaller regions. The subdivisions in both hemispheres are similar but not quite identical, indicating imperfect bilateral symmetry. The divisions at the bottom level are well correlated with known anatomical divisions, as shown by the labels in Fig. 4a. Most often, our method finds *subdivisions* of anatomical regions — i.e. a single anatomical region is divided in one or more groups — which are

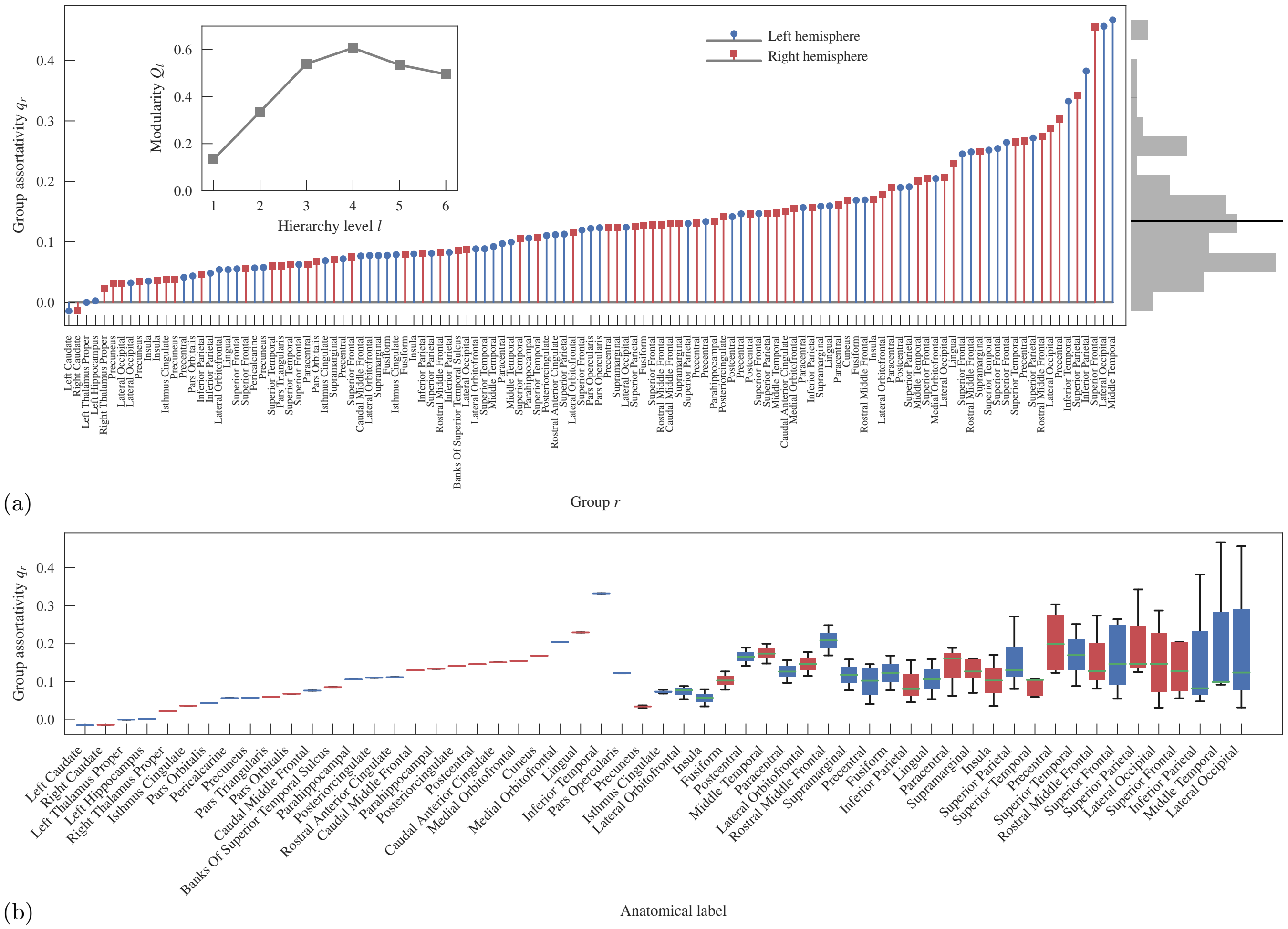

Figure 5. (a) Group assortativity q_r (Eq. 19) for the lowest level of the hierarchy in Fig 4a, with groups labelled using the most frequent anatomical annotation. Blue circle (red square) markers correspond to the left (right) hemisphere. On the right axis is shown a histogram of the q_r values, with a horizontal line marking the average $Q = \sum_r q_r/B \approx 0.13$. The inset shows the modularity value Q_l as a function of the hierarchy level l. (b) Dispersion of q_r values for groups that share the same anatomical annotation, as labelled in the x-axis.

then grouped together higher in the hierarchy. But we also find some regions that belong to the same anatomical group that end up classified in significantly different hierarchy branches, pointing to a further degree of heterogeneity inside anatomical regions. Since the various traditional approaches to determine such anatomical classification do not always take into account the local and global connectivity patterns (i.e. the actual connectome), our approach suggests an alternative or complementary method to perform such a task.

Like in the previous examples, the fit of the overall distributions of edge covariates provided by the SBM is reasonably convincing, as we see in Figs. 4b and c, indicating that these nontrivial distributions — which deviate significantly from the basic distributions used in the model — can be well explained by group-to-group mixtures.

1. *Community structure?*

The modular structure of the brain has been studied numerous times before, using a variety of methods (e.g. [24–26]). Most often, however, this is done by searching for *assortative* modules [27], i.e. groups of nodes more connected to themselves than with the rest of the network — a pattern commonly called *community structure* [1, 28]. In contrast, the approach developed here seeks to find groups of nodes that have similar probabilities of connection with the rest of the network (and to generate edge covariates), regardless if they form a community or not. Naturally, community structure is a special case of the general class of patterns that we consider, but our approach is capable of accommodating many others, such as core-peripheries and bipartiteness — in fact, any arbitrary kind of group affinities. This means that if the formation of assortative communities is the main driving mechanism responsible for the network structure, we should be able to detect it with our method, but otherwise it will prefer a non-assortative division. This makes it a more flexible and potentially more informative approach in comparison to typical community detection methods, which, by construction, will tend to omit non-assortative divisions, however important they may be, in favor of assortative ones. In the case of brain networks, very often the community detection approach used is the maximization of modularity [27], defined as

$$Q = \frac{1}{2E}\sum_r e_{rr} - \frac{e_r^2}{2E}, \qquad (17)$$

where $e_{rs} = m_{rs}(1+\delta_{rs})$ is the number of edges between groups r and s (or twice that if $r = s$), and $e_r = \sum_s e_{rs}$. As has been known for a long time [29], and since then has become well understood [30–34], the direct maximization of Q to detect communities will generically *overfit*, as it will misleadingly find many spurious communities and produce large Q values for completely random networks, as well as arguably non-modular networks such as trees. Somewhat paradoxically, the same approach will also generically *underfit*, as it is incapable of detecting a number of communities larger than $\sqrt{E}$ [35], even if their presence is statistically significant. Because of these and other limitations, as well as its non-statistical nature, the unsupervised maximization of Q to find communities is ill-advised in most contexts [36]. In contrast, the approach presented here is free of both these problems: When applied to completely random networks, it will not uncover spurious groups not sufficiently backed by statistical evidence [7]; and it is capable of detecting up to $\propto N/\log N$ groups [5, 6], whenever they are present. Since we have principled guarantees that the modules uncovered with our method are statistically significant, we can then use the value of Q to characterize the degree of assortativity of the modules found (rather than the quality of the partition). For the result shown in Fig. 4, we obtain $Q \approx 0.13$, which is typically considered a low value indicating weak community structure. We may understand this value in more detail by decomposing it as

$$Q = \frac{1}{B}\sum_r q_r \qquad (18)$$

where

$$q_r = \frac{B}{2E}\left(e_{rr} - \frac{e_r^2}{2E}\right) \qquad (19)$$

is the local assortativity of group r, with $q_r \in [-1, 1]$. In Fig. 5a we show the values of q_r for the modules inferred with our method, labelled according to most prominent anatomical classification. We see that while most values are positive, $q_r > 0$, strictly indicating a degree of assortativity, they are distributed across a broad range — with regions like the *Caudate nucleus* (associated with motor functions) even showing dissortativity with $q_r < 0$ — indicating that assortativity, although it is present, is not an overwhelmingly dominant descriptor of the large-scale structure (a similar point has been made recently [37] using the method of Ref. [4]). We note also that inferred groups that are associated with the same anatomical region sometimes possess very different assortativity, as shown in Fig. 5b. This gives us an insight as to why they were classified in different groups in the first place, and further corroborating the idea that specific anatomical regions have noticeable internal heterogeneity.

One might speculate that assortativity is just one of a diverse set of driving forces behind the network formation, and that inspecting a detailed model of the network might dilute its importance. Here we can can further exploit the multilevel nature of our inferred model to assess if assortativity becomes more relevant at higher levels of coarse-graining. We can do so by computing a different value Q_l for each hierarchical level l, defined by replacing $m_{rs} \to m_{rs}^l$ in Eq. 17, where m_{rs}^l is the number of edges between groups r and s at level l. As we can see in the inset of Fig. 5a, the values of Q_l do significantly increase at higher levels, suggesting that assortativity might be an important mechanism for the most global structures of the network, but not as much for its sub-structures at a smaller scale.

IV. ELEMENTARY MODELS FOR EDGE WEIGHTS

In this section we derive models for edge covariates based on basic properties, such as whether they are signed or unsigned, continuous or discrete, bounded or unbounded. In particular, we focus on formulations that allow the integrated marginal likelihood of the SBM to be computed exactly. For some of the derivations, we will assume — for convenience of notation — that the graphs are simple, i.e. $A_{ij} \in \{0, 1\}$. We do so without loss of generality, as the final expressions will also be valid for multigraphs.

A. Continuous unsigned weights

If all we know about the edge weights is that they are continuous and positive, i.e. $x_{ij} > 0$, a reasonable model is a maximum-entropy distribution with a fixed average, i.e. the exponential distribution

$$P(x|\lambda) = \lambda \mathrm{e}^{-\lambda x}. \qquad (20)$$

Using this as the basis of our weighted SBM yields,

$$P(\boldsymbol{x}_{rs}|\boldsymbol{A}, \boldsymbol{\lambda}, \boldsymbol{b}) = \prod_{ij} P(x_{ij}|\lambda_{rs})^{\frac{A_{ij}\delta_{b_i,r}\delta_{b_i,s}}{1+\delta_{rs}}} \qquad (21)$$

$$= \lambda_{rs}^{m_{rs}} \mathrm{e}^{-\lambda_{rs}\mu_{rs}}, \qquad (22)$$

with

$$\mu_{rs} = \sum_{ij} \frac{A_{ij}x_{ij}\delta_{b_i,r}\delta_{b_j,s}}{1+\delta_{rs}} \tag{23}$$

being the sum of the weights between groups r and s. Before computing the integrated marginal likelihood of Eq. 7, we need to select a prior for $\boldsymbol{\lambda}$. A natural choice that makes the computation feasible is known as a conjugate prior, which in this case is the gamma distribution

$$P(\lambda|\alpha,\beta) = \frac{\beta^\alpha\lambda^{\alpha-1}}{\Gamma(\alpha)}\mathrm{e}^{-\lambda\beta}, \tag{24}$$

where α and β are hyperparameters controlling its shape. Using this prior, we can write the marginal likelihood for the network weights by integrating over all λ_{rs}, yielding

$$P(\boldsymbol{x}|\boldsymbol{A},\boldsymbol{b},\alpha,\beta) = \prod_{r\le s}\frac{\Gamma(m_{rs}+\alpha)}{\Gamma(\alpha)}\frac{\beta^\alpha}{(\mu_{rs}+\beta)^{m_{rs}+\alpha}}. \tag{25}$$

Doing so, we have reduced the initially high number of parameters from $B(B+1)/2$ to only two, corresponding to the hyperparameters α and β. Being global parameters, independent of the internal dimension of the model, they can be chosen via maximum likelihood, without significant risk of overfitting,

$$\hat{\alpha},\hat{\beta} = \underset{\alpha,\beta}{\operatorname{argmax}}\; P(\boldsymbol{x}|\boldsymbol{A},\boldsymbol{b},\alpha,\beta), \tag{26}$$

which can be done efficiently with any standard optimization method. Alternatively, we may consider the choice $\alpha = 1$, for which $P(\lambda|\alpha,\beta)$ becomes the maximum-entropy distribution with a fixed mean, and hence has the same shape as $P(x|\lambda)$. Even with this choice, however, is difficult to incorporate this prior in the nested SBM via Eq. 10, as the integration over the remaining hierarchical levels is cumbersome. Instead, we now describe a microcanonical formulation which generates covariates in an asymptotically identical manner, but permits the exact integration of Eq. 10.

1. *Microcanonical distribution*

Instead of generating each covariate independently, we consider the uniform joint distribution of N positive real values $\boldsymbol{x} = \{x_1,\dots,x_N\}$ conditioned on their total sum $\mu = \sum_i x_i$,

$$P(\boldsymbol{x}|\mu) = \begin{cases} \dfrac{(N-1)!}{\mu^{N-1}}\delta(\mu - \sum_i x_i) & \text{if } \mu > 0 \\ \prod_i \delta(x_i) & \text{if } \mu = 0, \end{cases} \tag{27}$$

where the normalization constant $(N-1)!/\mu^{N-1}$ above accounts for the volume of a scaled simplex of dimension $N-1$. Although the covariates are not generated independently in Eq. 27, the marginal distribution of the individual values x_i can be obtained as

$$P(x_i|N,\mu) = \frac{P(\boldsymbol{x}|\mu)}{P(\boldsymbol{x}\setminus x_i|\mu - x_i)}, \tag{28}$$

$$= \frac{(N-1)(\mu-x_i)^{N-2}}{\mu^{N-1}}\Theta(\mu - x_i) \tag{29}$$

using Eq. 27 both in the numerator and denominator of Eq. 28, and $\Theta(x)$ is the Heaviside step function. Taking the limit $N\to\infty$ while keeping the mean $\bar{x} = \mu/N$ fixed, $P(x_i|N,\mu)$ becomes Eq. 20 with $\lambda = 1/\bar{x}$ (see Fig. 6). Since in the limit of sufficient data both models become identical, the microcanonical model enables us to have an exact hierarchical SBM, as we will now show, without sacrificing descriptive power.

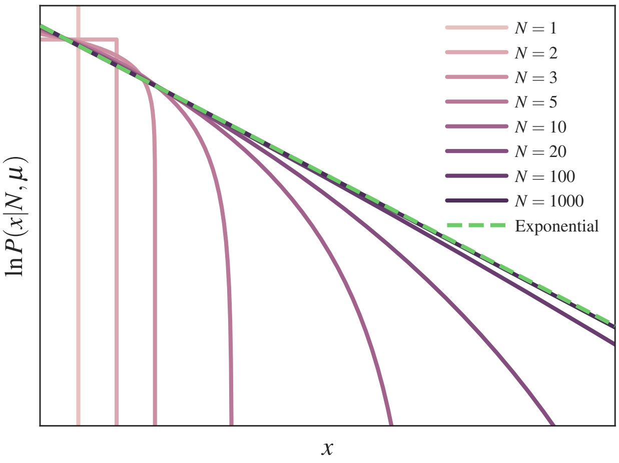

Figure 6. The marginal distribution of each individual covariate x in the unsigned microcanonical model, given by Eq. 29, approaches asymptotically the exponential distribution as the number of values N increases, and if the mean $\bar{x} = \mu/N$ is kept fixed.

Incorporating the microcanonical model of Eq. 27 in the SBM amounts simply to

$$P(\boldsymbol{x}_{rs}|\boldsymbol{A},\boldsymbol{\lambda},\boldsymbol{b}) = P(\boldsymbol{x}_{rs}|\mu_{rs}), \tag{30}$$

where, as before, μ_{rs} is the sum of covariates between groups r and s. To generate the parameters μ_{rs} — which are also non-negative real numbers — we can use the exact same distribution again at a higher hierarchical level, by treating them as edge covariates of the graph of groups, as described in Eq. 10. The microcanonical nature of this model makes the integration over all parameters $\{\boldsymbol{\mu}^l_{rs}\}$ trivial due to the hard constraints, i.e.

$$P(\boldsymbol{x}|\boldsymbol{A},\{\boldsymbol{b}^l\}) = \int P(\boldsymbol{x}|\boldsymbol{A},\boldsymbol{\mu}^1,\boldsymbol{b}^1)\prod_{l=1}^{L}\prod_{r\le s}\left[P(\boldsymbol{\mu}^l_{rs}|\mu^{l+1}_{b_r^{(l+1)},b_s^{(l+1)}})\,\mathrm{d}\boldsymbol{\mu}^l_{rs}\right]^{1-\delta_{m^l_{rs},0}} \tag{31}$$

$$= \prod_{l=1}^{L}\prod_{r\le s}\left[\frac{(m^l_{rs}-1)!}{(\bar\mu^l_{rs})^{m^l_{rs}-1}}\right]^{1-\delta_{\mu^l_{rs},0}}, \tag{32}$$

where

$$\bar\mu^l_{rs} = \sum_{tu}\left(\frac{\bar\mu^{l-1}_{tu}\delta_{b^l_t,r}\delta_{b^l_u,s}}{1+\delta_{rs}}\right)^{1-\delta_{m^l_{tu},0}} \tag{33}$$

is the sum of covariates between groups r and s at level $l > 1$, and with $\bar\mu^1_{rs} = \mu_{rs}$ given by Eq. 23 at the lowest level. Recall that the boundary condition used in Eq. 10 is that at the topmost level there is only one group, and hence $m^L_{rs} = E\delta_{r,1}\delta_{s,1}$ and $\bar\mu^L_{rs} = \hat\mu\delta_{r,1}\delta_{s,1}$, where $\hat\mu = \sum_{i<j}A_{ij}x_{ij}$ is the total sum of edge weights, and the sole remaining parameter of the model. The marginal likelihood of Eq. 32 is a simple term that can be computed easily by obtaining the covariate summaries at each level, and amounts to a straightforward modification of the algorithm of Ref. [6] to obtain the posterior distribution of hierarchical partitions. In particular, this additional term does not affect its algorithm complexity, since changes in a lower hierarchical level that are compatible with the partition at higher level do not alter the likelihoods in the upper levels, as the covariate sums remain unchanged.

B. Continuous signed weights

For weights that can be either positive or negative, we require a maximum entropy distribution with fixed average and variance, which is the normal distribution

$$P(x|\bar x,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\mathrm{e}^{-\frac{(x-\bar x)^2}{2\sigma^2}}. \tag{34}$$

Incorporating this in the SBM, we obtain

$$P(\boldsymbol{x}_{rs}|\boldsymbol{A},\bar x_{rs},\sigma^2_{rs}) = \prod_{ij}P(x_{ij}|\bar x_{rs},\sigma^2_{rs})^{\frac{A_{ij}\delta_{b_i,r}\delta_{b_i,s}}{1+\delta_{rs}}} \tag{35}$$

$$= \frac{\mathrm{e}^{-\frac{\nu_{rs}-2\mu_{rs}\bar x_{rs}+m_{rs}\bar x^2_{rs}}{2\sigma^2_{rs}}}}{(2\pi\sigma^2_{rs})^{m_{rs}/2}}, \tag{36}$$

with

$$\nu_{rs} = \sum_{ij}\frac{A_{ij}x^2_{ij}\delta_{b_i,r}\delta_{b_j,s}}{1+\delta_{rs}} \tag{37}$$

being the sum of squares of covariates between groups r and s. The conjugate prior for $\bar x$ and σ^2 is the normal-inverse-chi-squared distribution [38]

$$P(\bar x,\sigma^2|\mu_0,\kappa_0,\nu_0,\sigma^2_0) = \mathcal{N}(\bar x|\mu_0,\sigma^2/\kappa_0)\chi^{-2}(\sigma^2|\nu_0,\sigma^2_0), \tag{38}$$

where $\mathcal{N}(\bar x|a,b)$ is a normal distribution with mean a and variance b, and the variance is sampled from an inverse-chi-squared distribution

$$\chi^{-2}(\sigma^2|\upsilon,\tau^2) = \frac{(\tau^2\upsilon/2)^{\upsilon/2}}{\Gamma(\upsilon/2)}\,\frac{\mathrm{e}^{\frac{-\upsilon\tau^2}{2\sigma^2}}}{\sigma^{2+\upsilon}}. \tag{39}$$

Using this prior, after the integration over $\bar x$ and σ^2, the marginal likelihood becomes

$$P(\boldsymbol{x}_{rs}|\boldsymbol{A},\mu_0,\kappa_0,\nu_0,\sigma^2_0) = \frac{\Gamma(\nu'_{rs}/2)}{\Gamma(\nu_0/2)}\sqrt{\frac{\kappa_0}{\kappa_{rs}}}\frac{(\nu_0\sigma^2_0)^{\nu_0/2}}{(\nu'_{rs}S_{rs})^{\nu_{rs}/2}}\frac{1}{\pi^{m_{rs}/2}}, \tag{40}$$

with auxiliary quantities

$$\kappa_{rs} = \kappa_0 + m_{rs}, \quad \nu'_{rs} = \nu_0 + m_{rs}, \tag{41}$$

$$z_{rs} = \nu_{rs} - \mu^2_{rs}/m_{rs}, \tag{42}$$

$$S_{rs} = \frac{1}{\nu'_{rs}}\left[\nu_0\sigma^2_0 + z_{rs} + \frac{m_{rs}\kappa_0}{\kappa_0+m_{rs}}\left(\mu_0 - \frac{\mu_{rs}}{m_{rs}}\right)^2\right]. \tag{43}$$

This leaves us with four global parameters, μ_0, κ_0, ν_0 and σ^2_0, that we have to determine either with maximum likelihood, or maximum entropy arguments. However, like the unsigned case previously, the shape of the marginal likelihood leaves little chance of building a hierarchical model in closed form. Luckily, we can once more construct a microcanonical model that allows us do precisely that.

1. *Microcanonical distribution*

The corresponding microcanonical maximum-entropy formulation for signed covariates is the uniform distribution of N values $\boldsymbol{x}$ conditioned in the total sum μ and sum of squares ν,

$$P(\boldsymbol{x}|\mu,\nu) = \frac{\delta(\mu-\sum_i x_i)\delta(\nu-\sum_i x_i^2)}{\Omega}. \tag{44}$$

The normalization constant is computed as

$$\Omega = \int \delta(\mu - \textstyle\sum_i x_i)\delta(\nu - \sum_i x_i^2)\,\mathrm{d}\boldsymbol{x} \tag{45}$$

$$= \int_H \frac{\delta(\nu - \sum_i x_i(\boldsymbol{y})^2)}{\sqrt{N}}\,\mathrm{d}\boldsymbol{y}(\boldsymbol{x}) \tag{46}$$

$$= \int_S \frac{\mathrm{d}\boldsymbol{\sigma}(\boldsymbol{x})}{2\sqrt{N\nu - \mu^2}} \tag{47}$$

$$= \frac{\pi^{(N-1)/2}\left(\nu - \mu^2/N\right)^{(N-3)/2}}{\Gamma(N/2 - 1/2)\sqrt{N}}, \tag{48}$$

where H in Eq. 46 is the hyperplane given by $\sum_i x_i = \mu$, parametrized by $N-1$ coordinates $\boldsymbol{y}(\boldsymbol{x})$, and S in Eq. 47 is the intersection of a N-sphere of radius $\sqrt{\nu}$ and the hyperplane H, which corresponds to the surface of a $(N-1)$-sphere of radius $\sqrt{\nu - \mu^2/N}$, with surface element $\mathrm{d}\boldsymbol{\sigma}(\boldsymbol{x})$, leading to Eq. 48. Therefore, we have for the complete microcanonical distribution

$$P(\boldsymbol{x}|\mu,\nu) = \begin{cases} \dfrac{\Gamma(N/2-1/2)\sqrt{N}}{\pi^{(N-1)/2}(\nu - \mu^2/N)^{(N-3)/2}}\delta(\mu - \sum_i x_i)\delta(\nu - \sum_i x_i^2), & \text{if } \nu > \mu^2/N,\\ \prod_i \delta(\mu/N - x_i), & \text{if } \nu = \mu^2/N. \end{cases} \tag{49}$$

The marginal distribution of the individual covariates x_i can be obtained as

$$P(x_i|N,\mu,\nu) = \frac{P(\boldsymbol{x}|\mu,\nu)}{P(\boldsymbol{x}\setminus x_i|\mu - x_i, \nu - x_i^2)}, \tag{50}$$

$$= \frac{\Gamma(N/2-1/2)}{\Gamma(N/2-1)}\sqrt{\frac{N}{\pi(N-1)}}\frac{[\nu - x_i^2 - (\mu - x_i)^2/(N-1)]^{N/2-2}}{(\nu - \mu^2/N)^{(N-3)/2}}\Theta(\mu - x_i)\Theta(\nu - x_i^2), \tag{51}$$

using Eq. 49 both in the numerator and denominator of Eq. 50. Taking the limit $N \to \infty$ while keeping both the mean $\bar{x} = \mu/N$ and variance $\sigma^2 = \nu/N - \bar{x}^2$ fixed, $P(x_i|N,\mu,\nu)$ becomes the normal distribution of Eq. 44 (see Fig. 7). Therefore, like with the unsigned case, the microcanonical model yields an easy-to-integrate model, without sacrificing descriptive power.

Incorporating the above distribution into the SBM yields

$$P(\boldsymbol{x}_{rs}|\boldsymbol{A},\boldsymbol{\mu},\boldsymbol{\nu},\boldsymbol{b}) = P(\boldsymbol{x}_{rs}|\mu_{rs},\nu_{rs}), \tag{52}$$

where, as before, μ_{rs} is the sum of covariates between groups r and s, and ν_{rs} is the sum of squares of the same covariates. To generate the parameters μ_{rs}, we can use the exact same distribution again at a higher hierarchical level. The parameters ν_{rs}, however, are strictly positive, and hence require a different model. Furthermore, μ_{rs} and ν_{rs} are not independent parameters, as they must satisfy the inequality $\nu_{rs} \ge \mu_{rs}^2/m_{rs}$. Therefore, we reparametrize the model using the auxiliary quantity of Eq. 42

$$z_{rs} = \nu_{rs} - \mu_{rs}^2/m_{rs}, \tag{53}$$

which is simply the scaled variance of the covariates, and thus is strictly non-negative and can be chosen independently from μ_{rs}. We can then generate z_{rs} from the unsigned microcanonical model of Eq. 27. Although we can easily write the final marginal likelihood of the model if we propagate the hyperpriors of z_{rs} upwards in the hierarchy of the nested SBM, we would have the following problems: Not only this would increase the total number of edge covariates at the highest levels (and hence it is unclear *a priori* if it is the most parsimonious approach), since each signed parameter requires two hyperparameters, but also it leads to a model that is cumbersome computationally, as changes in a lower level would always propagate through the whole hierarchy. Instead, here we opt to propagate only μ_{rs} upwards in the hierarchy, whereas we generate all z_{rs} from the same distribution at each level. More concretely, we write

$$P(\boldsymbol{x}|\boldsymbol{A},\{\boldsymbol{b}^l\}) = \int P(\boldsymbol{x}|\boldsymbol{A},\boldsymbol{\mu}^1,\boldsymbol{z}^1,\boldsymbol{b}^1)P(\boldsymbol{\mu}_z)\prod_{l=1}^{L}P(\boldsymbol{z}^l|\mu_z^l)\prod_{r\le s}\left[P(\boldsymbol{\mu}_{rs}^l|\mu^{l+1}_{b_r^{(l+1)},b_s^{(l+1)}},\boldsymbol{z}^{l+1}_{b_r^{(l+1)},b_s^{(l+1)}})\right]^{1-\delta_{m_{rs}^l,0}}\,\mathrm{d}\boldsymbol{\mu}^l\mathrm{d}\boldsymbol{z}^l\mathrm{d}\mu_z^l, \tag{54}$$

$$= \frac{(\bar{L}-1)!}{(\sum_{l=1}^{L} m_z^l\bar{\mu}_z^l)^{\bar{L}-1}}\prod_{l=1}^{L}(m_z^l)^{1-\delta_{m_z^l,0}}\left[\frac{(m_z^l-1)!}{(\bar{\mu}_z^l)^{m_z^l-1}}\right]^{1-\delta_{\bar{\mu}_z^l,0}}\prod_{r\le s}\left[\frac{\Gamma(m_{rs}^l/2-1/2)\sqrt{m_{rs}^l}}{\pi^{(m_{rs}^l-1)/2}(\bar{z}_{rs}^l)^{(m_{rs}^l-3)/2}}\right]^{1-\delta_{\bar{z}_{rs}^l,0}}, \tag{55}$$

where $\bar{z}^l_{rs} = \bar{\nu}^l_{rs} - (\bar{\mu}^l_{rs})^2/m^l_{rs}$, with $\bar{\mu}^l_{rs}$ given by Eq. 33, and

$$\bar{\nu}^l_{rs} = \sum_{tu}\left[\frac{(\bar{\mu}^{l-1}_{tu})^2\delta_{b^l_t,r}\delta_{b^l_u,s}}{1+\delta_{rs}}\right]^{1-\delta_{m^l_{tu},0}}, \tag{56}$$

corresponds to the scaled variance of the values of $\bar{\mu}^{l-1}_{rs}$ at a lower level (assuming the boundary conditions $\bar{\mu}^1_{rs} = \mu_{rs}$ and $\bar{\nu}^1_{rs} = \nu_{rs}$ given by Eqs. 23 and 37, respectively), and where

$$m^l_z = \sum_{r\le s} H(m^l_{rs} - 1), \tag{57}$$

$$\bar{\mu}^l_z = \sum_{r\le s} \bar{z}^l_{rs} H(m^l_{rs} - 1), \tag{58}$$

are the sum and scaled average $\bar{z}^l_{rs}$ of $m^l_{rs} > 1$ entries at level l, with $H(x) = 1$ if $x > 0$, otherwise $H(x) = 0$, and $P(\boldsymbol{z}^l|\mu^l_z)$ is given by Eq. 27. The above means that when computing $P(\boldsymbol{z}^l|\mu^l_z)$ we must only consider values of z^l_{rs} for which $m^l_{rs} > 1$. Otherwise, if $m^l_{rs} = 1$, the corresponding parameter must always be $\nu^l_{rs} = (\mu^l_{rs})^2$, and hence $z^l_{rs} = 0$, which does not need to be sampled from a prior. Finally, the values of $\boldsymbol{\mu}_z = \{\mu^l_z\}$ across all levels are also sampled from their own model as

$$P(\boldsymbol{\mu}_z) = P(\{m^l_z\mu^l_z\}|\textstyle\sum_{l=1}^L m^l_z\mu^l_z)\displaystyle\prod_{l=1}^L (m^l_z)^{1-\delta_{m^l_z,0}}, \tag{59}$$

using again Eq. 27, and where the trailing product is a derivative term that accounts for the scaling of the variables in the argument of the first term, and with

$$\bar{L} = \sum_{l=1}^L H(m^l_z) \tag{60}$$

being the number of levels with non-zero values of m^l_z. The boundary condition in Eqs. 10 and 54, i.e. that the last level of the hierarchy has only one group, means that the two remaining parameters are $\bar{\mu}^L_{rs} = \hat{\mu}\delta_{r,1}\delta_{s,1}$, where $\hat{\mu} = \sum_{i<j} A_{ij}x_{ij}$ is the total sum of edge weights, and $\hat{\mu}_z = \sum_l m^l_z\bar{\mu}^l_z$ which is the sum of scaled variances across the hierarchy levels.

Like with the unsigned model, Eq. 55 amounts to a straightforward modification of the algorithm of Ref. [6], requiring only an additional book-keeping of the values of z^l_{rs} for which m^l_{rs} is larger than one, and their respective sums, which can be done without altering the overall algorithmic complexity. We remark also that, unlike maximum likelihood approaches applied directly to Eq. 44, the resulting marginal likelihood of the microcanonical model is well defined and yields non-degenerate results for any possible set of covariates, even those yielding zero variance or populations with single elements.

C. Geometric discrete weights

For discrete non-negative weights, i.e. $x \in \mathbb{N}_0$, the maximum entropy distribution with a fixed average is

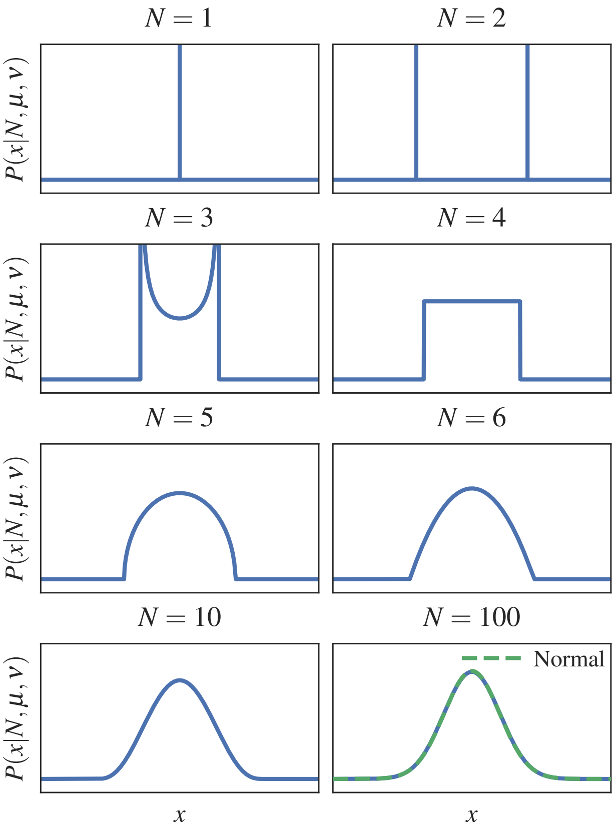

Figure 7. The marginal distribution of each individual covariate x in the signed microcanonical model, given by Eq. 51, approaches asymptotically the normal distribution as the number of values N increases, and if the mean $\bar{x} = \mu/N$ and variance $\sigma^2 = \nu/N - \bar{x}^2$ are kept fixed.

the geometric distribution

$$P(x|p) = (1-p)^x p. \tag{61}$$

Using it for the SBM, we have

$$P(\boldsymbol{x}_{rs}|\boldsymbol{A},\boldsymbol{b},p_{rs}) = (1-p_{rs})^{\mu_{rs}}p_{rs}^{m_{rs}}. \tag{62}$$

The conjugate prior for p is the beta distribution

$$P(p|\alpha,\beta) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}, \tag{63}$$

with $B(x,y) = \Gamma(x)\Gamma(y)/\Gamma(x+y)$, which yields the marginal distribution

$$P(\boldsymbol{x}_{rs}|\boldsymbol{A},\boldsymbol{b},\alpha,\beta) = \frac{B(m_{rs}+\alpha,\mu_{rs}+\beta)}{B(\alpha,\beta)}. \tag{64}$$

Unlike the continuous case, we can make a fully "uninformative" choice $\alpha = \beta = 1$ that reflects our maximum ignorance about the parameter p, as in this case it is uniformly sampled in the interval $p \in [0,1]$. This yields simply

$$P(\boldsymbol{x}_{rs}|\boldsymbol{A},\boldsymbol{b}) = \frac{m_{rs}!\,\mu_{rs}!}{(m_{rs}+\mu_{rs}+1)!}. \tag{65}$$

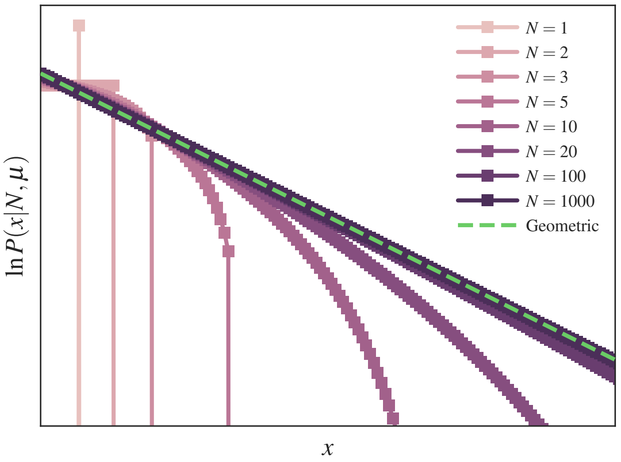

Figure 8. The marginal distribution of each individual covariate x in the discrete microcanonical model, given by Eq. 68, approaches asymptotically the geometric distribution as the number of values N increases, and the mean value μ/N is kept fixed.

However, this kind of uninformative assumption rarely matches what we end up finding in the data, which tends to be significantly more structured. A more robust approach is to construct a hierarchical model, which can be more easily done with a microcanonical description.

1. *Microcanonical distribution*

The microcanonical analogue of the geometric distribution is the uniform distribution of N non-negative discrete real values $\boldsymbol{x}$ conditioned on their total sum μ, given by

$$P(\boldsymbol{x}|\mu) = \left(\!\!\binom{N}{\mu}\!\!\right)^{-1} \delta_{\mu, \sum_i x_i}, \tag{66}$$

where $\left(\!\binom{N}{\mu}\!\right) = \binom{N+\mu-1}{\mu}$ counts the number of ways to distribute a total of μ values into N distinguishable parts. The marginal distribution of the individual covariates x_i can be obtained as

$$\begin{aligned} P(x_i|N, \mu) &= \frac{P(\boldsymbol{x}|\mu)}{P(\boldsymbol{x} \setminus x_i|\mu - x_i)}, &(67)\\ &= \frac{(N+\mu-x_i-1)!\mu!}{(N+\mu-1)!(\mu-x_i)!} H(\mu - x_i). &(68)\end{aligned}$$

Like with the continuous model, for sufficiently large N and with the mean μ/N fixed, the marginal distribution of individual values x_i will follow asymptotically a geometric distribution with $p = N/(\mu + N)$ (see Fig. 8).

Since the value of the parameter μ is also non-negative, we can sample it from the same distribution as a prior. Putting this in the SBM yields

$$P(\boldsymbol{x}|\boldsymbol{A}, \boldsymbol{\mu}^1, \boldsymbol{b}^1) = \prod_{r \le s} P(\boldsymbol{x}^1_{rs}|\mu^1_{rs}) \tag{69}$$

and the final marginal distribution

$$P(\boldsymbol{x}|\boldsymbol{A}, \{\boldsymbol{b}^l\}) = \sum_{\{\mu^l_{rs}\}} P(\boldsymbol{x}|\boldsymbol{A}, \boldsymbol{\mu}^1, \boldsymbol{b}^1) \prod_{l=1}^{L}\prod_{r\le s}\left[P(\boldsymbol{\mu}^l_{rs}|\mu^{l+1}_{b_r^{(l+1)}, b_s^{(l+1)}})\right]^{1-\delta_{m^l_{rs},0}} = \prod_{l=1}^{L}\prod_{r\le s}\left[\left(\!\!\binom{m^l_{rs}}{\bar{\mu}^l_{rs}}\!\!\right)^{-1}\right]^{1-\delta_{\bar{\mu}^l_{rs},0}}, \tag{70}$$

with $\bar{\mu}^l_{rs}$ given by Eq. 33. Like with the continuous models, the use of Eq. 70 requires only a simple modification of the algorithm of Ref. [6], that does not alter its algorithmic complexity.

D. Binomial discrete weights

Often, discrete covariates are bounded in a finite range $x \in \{0, \ldots, M\}$ (a common example are ratings in recommendation systems [39]). In this case, the appropriate distribution is the binomial,

$$P(x|p, M) = \binom{M}{x} p^x (1-p)^{M-x}, \tag{71}$$

where the value x is commonly interpreted as the sum of M independent Bernoulli outcomes with a probability $p \in [0, 1]$ of success. Incorporating it in the SBM yields

$$\begin{aligned} &P(\boldsymbol{x}_{rs}|\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{p}, N)\\ &\quad= \prod_{ij} P(x_{ij}|p_{rs}, N)^{\frac{A_{ij}\delta_{b_i,r}\delta_{b_i,r}}{(1+\delta_{rs})}}\\ &\quad= \left[\prod_{ij}\binom{M}{x_{ij}}^{\frac{A_{ij}\delta_{b_i,r}\delta_{b_i,r}}{(1+\delta_{rs})}}\right] p_{rs}^{\mu_{rs}}(1-p_{rs})^{Mm_{rs}-\mu_{rs}}. \end{aligned} \tag{72}$$

The conjugate prior is the beta distribution of Eq. 63 again, yielding the marginal after integration over all p_{rs},

$$P(\boldsymbol{x}|\boldsymbol{A},\boldsymbol{b},\alpha,\beta) = \left[\prod_{i<j}\binom{M}{x_{ij}}\right]\prod_{r\le s}\frac{B(\mu_{rs}+\alpha, Mm_{rs}-\mu_{rs}+\beta)}{B(\alpha,\beta)}. \quad (73)$$

Once more, we can make the uninformative choice $\alpha = \beta = 1$, which yields

$$P(\boldsymbol{x}|\boldsymbol{A},\boldsymbol{b}) = \left[\prod_{i<j}\binom{M}{x_{ij}}\right]\prod_{r\le s}\frac{\mu_{rs}!(Mm_{rs}-\mu_{rs})!}{(Mm_{rs}+1)!}. \quad (74)$$

But as for the other cases, the best path for a hierarchical model is through a microcanonical model, as described in the following.

1. Microcanonical distribution

A microcanonical version of the Binomial distribution — i.e. the uniform distribution of N non-negative discrete values $\boldsymbol{x}$, where each value is bounded in the range $x_i \in \{0,\dots,M\}$, conditioned in the total sum μ — can be obtained by randomly sampling exactly μ positive outcomes from a total of NM trials. The joint probability for $\boldsymbol{x} = \{x_1,\dots,x_N\}$ is therefore

$$P(\boldsymbol{x}|\mu, M) = \left[\prod_i\binom{M}{x_i}\right]\binom{MN}{\mu}^{-1}\delta_{\mu,\sum_i x_i}, \quad (75)$$

where $\binom{MN}{\mu}$ counts the possible distributions of μ positive outcomes of MN distinguishable trials, and the remaining terms discount all outcomes that lead to the same value of $\boldsymbol{x}$. The marginal distribution of the individual covariates x_i can be obtained as

$$\begin{aligned}P(x_i|N,\mu,M) &= \frac{P(\boldsymbol{x}|\mu,M)}{P(\boldsymbol{x}\setminus x_i|\mu - x_i, M)},\\ &= \binom{M}{x_i}\frac{[M(N-1)]![M(N-1)-\mu+x_i]!\mu!}{(MN)!(\mu-x_i)!(MN-\mu)!}.\end{aligned} \quad (76)$$

Like with the previous models, for sufficiently large N and with μ/N fixed, the marginal distribution of individual values x_i will follow asymptotically a binomial distribution with $p = \mu/(NM)$ (see Fig. 9).

The parameter μ is a non-negative integer that can be chosen arbitrarily, as long as the inequality $M \ge \mu/N$ is satisfied. Therefore, we may sample μ from the distribution of Eq. 66 in an unconstrained manner, and then sample the parameter M from a constrained distribution $P(M|\mu,N)$. Incorporating this into the SBM yields,

$$P(\boldsymbol{x}|\boldsymbol{A},\boldsymbol{\mu}^1,\boldsymbol{b}^1) = \prod_{r\le s}P(\boldsymbol{x}^1_{rs}|\mu^1_{rs},M), \quad (77)$$

and the overall marginal distribution

$$P(\boldsymbol{x},M|\boldsymbol{A},\{\boldsymbol{b}^l\}) = \sum_{\{\mu^l_{rs}\}}P(\boldsymbol{x}|\boldsymbol{A},\boldsymbol{\mu}^1,\boldsymbol{b}^1)P(M|\boldsymbol{\mu}^1,\boldsymbol{A},\boldsymbol{b}^1)\prod_{l=1}^{L}\prod_{r\le s}\left[P(\boldsymbol{\mu}^l_{rs}|\mu^{l+1}_{b_r^{(l+1)},b_s^{(l+1)}})\right]^{1-\delta_{m^l_{rs},0}} \quad (78)$$

$$= P(M|\bar{\boldsymbol{\mu}}^1,\boldsymbol{A},\boldsymbol{b}^1)\left[\prod_{i\le j}\binom{M}{x_{ij}}\right]\left[\prod_{r\le s}\binom{Mm^1_{rs}}{\bar{\mu}^1_{rs}}^{-1}\right]\prod_{l=2}^{L}\prod_{r\le s}\left[\left(\!\!\binom{m^l_{rs}}{\bar{\mu}^l_{rs}}\!\!\right)^{-1}\right]^{1-\delta_{\bar{\mu}^l_{rs},0}}, \quad (79)$$

where $P(M|\bar{\boldsymbol{\mu}}^1,\boldsymbol{A},\boldsymbol{b}^1)$ is a prior distribution for M that respects the constraint $M \ge \bar{\mu}^1_{rs}/m^1_{rs}$. Thus, given any arbitrary value M^*, we can choose

$$P(M|\bar{\boldsymbol{\mu}}^1,\boldsymbol{A},\boldsymbol{b}^1) = \begin{cases}1 & \text{if } M = \max\left(M^*, \lceil\max_{rs}\bar{\mu}^1_{rs}/m^1_{rs}\rceil\right),\\ 0 & \text{otherwise,}\end{cases} \quad (80)$$

such that if M^* is compatible with the observed covariates, i.e. $M^* \ge x_{ij}$, we have $P(M^*|\bar{\boldsymbol{\mu}}^1,\boldsymbol{A},\boldsymbol{b}^1) = 1$ for any possible value of $\bar{\boldsymbol{\mu}}^1$ and $\boldsymbol{b}^1$ encountered in the posterior, as long as $M = M^*$, thereby effectively removing it from Eq. 79. A completely nonparametric approach would require us to include a prior $P(M^*)$, but since it is a single global number, we can safely omit it, as it cannot influence the posterior distribution of partitions. In most practical scenarios, the bound M^* is known *a priori*; otherwise it can be chosen as $M^* = \max_{ij} x_{ij}$.

E. Poisson discrete weights

A natural extension of the binomial weights is the situation where $M \to \infty$ with the mean $\lambda = pM$ kept fixed,

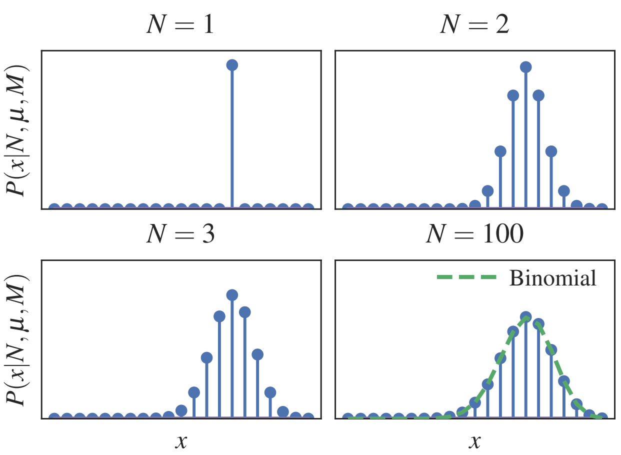

Figure 9. The marginal distribution of each individual covariate x in the discrete microcanonical model, given by Eq. 76, approaches asymptotically the binomial distribution as the number of values N increases, and the mean $\bar{x} = \mu/N$ is kept fixed.

which yields the Poisson distribution

$$P(x|\lambda) = \frac{\lambda^x \mathrm{e}^{-\lambda}}{x!}. \tag{81}$$

Using this in the SBM gives us

$$\begin{aligned} P(\boldsymbol{x}|\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{\lambda}) &= \prod_{i<j} P(x_{ij}|\lambda_{b_i,b_j})^{A_{ij}} & (82) \\ &= \left[\prod_{i<j} x_{ij}!^{A_{ij}}\right]^{-1} \prod_{r\le s} \lambda_{rs}^{\mu_{rs}} \mathrm{e}^{-m_{rs}\lambda_{rs}}. & (83) \end{aligned}$$

Once more, the conjugate prior is the gamma distribution of Eq. 24, which after integrating over λ_{rs} yields the marginal distribution

$$P(\boldsymbol{x}|\boldsymbol{A}, \boldsymbol{b}, \alpha, \beta) = \\ \left[\prod_{i<j} x_{ij}!^{A_{ij}}\right]^{-1} \prod_{r\le s} \frac{\beta^\alpha \Gamma(\mu_{rs} + \alpha)}{\Gamma(\alpha)(m_{rs} + \beta)^{\mu_{rs}+\alpha}}. \tag{84}$$

The uninformative maximum entropy choice is $\alpha = 1$, yielding

$$P(\boldsymbol{x}|\boldsymbol{A}, \boldsymbol{b}, \beta) = \left[\prod_{i<j} x_{ij}!^{A_{ij}}\right]^{-1} \prod_{r\le s} \frac{\beta \mu_{rs}!}{(m_{rs} + \beta)^{\mu_{rs}+1}}. \tag{85}$$

But once more, we can obtain a deeper hierarchical model by formulating an asymptotically equivalent microcanonical model.

1. *Microcanonical distribution*

The joint distribution of N Poisson variables $\boldsymbol{x} = \{x_1, \dots, x_N\}$ can be decomposed into a Poisson distribution for the total sum μ with mean $N\lambda$ and a uniform multinomial distribution for $\boldsymbol{x}$ conditioned on the total sum, i.e.

$$P(\boldsymbol{x}|\lambda) = P(\boldsymbol{x}|\mu)P(\mu|N\lambda). \tag{86}$$

The microcanonical version, therefore, is given simply by replacing $P(\mu|N\lambda) \to \delta_{\mu,\sum_i x_i}$, yielding

$$P(\boldsymbol{x}|\mu) = \frac{\mu!}{\prod_i x_i!} \frac{1}{N^\mu} \delta_{\mu,\sum_i x_i}. \tag{87}$$

The marginal distribution of the individual covariates x_i can be obtained again as

$$\begin{aligned} P(x_i|N, \mu) &= \frac{P(\boldsymbol{x}|\mu)}{P(\boldsymbol{x} \setminus x_i|\mu - x_i)}, & (88) \\ &= \frac{\mu!(N-1)^{\mu - x_i}}{(\mu - x)!N^\mu x_i!}, & (89) \end{aligned}$$

The global constraint on the total sum has a vanishing effect for sufficiently large N, as long as the mean μ/N is kept fixed, as the marginal distribution of individual values x_i will follow asymptotically a Poisson distribution with $\lambda = \mu/N$ (see Fig. 10).

The parameter μ is a non-negative integer that can be chosen arbitrarily. Therefore, we may sample μ from the distribution of Eq. 66. Incorporating this into the SBM yields,

$$P(\boldsymbol{x}|\boldsymbol{A}, \boldsymbol{\mu}^1, \boldsymbol{b}^1) = \prod_{r\le s} P(\boldsymbol{x}_{rs}^1|\mu_{rs}^1), \tag{90}$$

and the overall marginal distribution

$$\begin{aligned} P(\boldsymbol{x}|\boldsymbol{A}, \{\boldsymbol{b}^l\}) &= \sum_{\{\mu_{rs}^l\}} P(\boldsymbol{x}|\boldsymbol{A}, \boldsymbol{\mu}^1, \boldsymbol{b}^1) \prod_{l=1}^{L} \prod_{r\le s} \left[P(\boldsymbol{\mu}_{rs}^l|\mu^{l+1}_{b_r^{(l+1)},b_s^{(l+1)}})\right]^{1-\delta_{m_{rs}^l,0}} & (91) \\ &= \left[\prod_{i\le j} x_{ij}!^{A_{ij}}\right]^{-1} \left[\prod_{r\le s} \left(\frac{\bar{\mu}_{rs}^1!}{(m_{rs}^1)^{\bar{\mu}_{rs}^1}}\right)^{1-\delta_{m_{rs}^1,0}}\right] \prod_{l=2}^{L} \prod_{r\le s} \left[\left(\!\!\binom{m_{rs}^l}{\bar{\mu}_{rs}^l}\!\!\right)^{-1}\right]^{1-\delta_{\bar{\mu}_{rs}^l,0}}. & (92) \end{aligned}$$

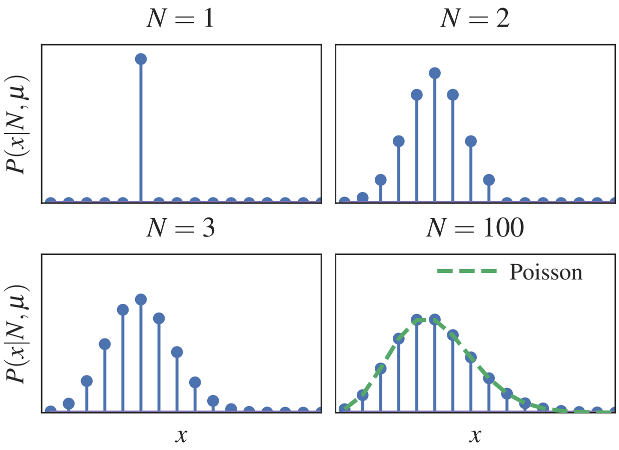

Figure 10. The marginal distribution of each individual covariate x in the discrete microcanonical model, given by Eq. 89, approaches asymptotically the Poisson distribution as the number of values N increases, and the mean $\bar{x} = \mu/N$ is kept fixed.

F. Transformed weights

The models above can be easily modified to accommodate a much wider class of covariates, without any substantial change to the likelihoods, via variable transformations of the type $y_{ij} = f(x_{ij})$, according to some function $f(x)$. For the continuous models in particular, such variable transformations yield the scaled marginal likelihoods

$$P(\boldsymbol{x}|\boldsymbol{A},\{\boldsymbol{b}^l\}) = P(\boldsymbol{y}(\boldsymbol{x})|\boldsymbol{A},\{\boldsymbol{b}^l\}) \prod_{i<j} \left[\frac{\mathrm{d}f}{\mathrm{d}x}(x_{ij})\right]^{A_{ij}}. \tag{93}$$

The product of derivatives in the equation above is a multiplicative constant that does not depend on the hierarchical partition $\{\boldsymbol{b}^l\}$, and hence does not affect the posterior distribution (although it is relevant for model selection; see Sec. IV G below). We are thus free to choose any weight transformation $f(x)$, and use the previously defined distributions and associated algorithms on the transformed weights, without any other alteration. This gives us a wider class of covariate models that may be better suitable for specific datasets, and can be developed in an *ad hoc* manner. In the following, we cover some typical examples, non-exhaustively.

1. *Broadly distributed weights*

If the observed weights are positive and broadly distributed, a possibly better model is the Pareto distribution,

$$P(x|\alpha,x_m) = \begin{cases} \dfrac{\alpha x_m^\alpha}{x^{\alpha+1}}, & \text{if } x > x_m,\\ 0, & \text{otherwise.}\end{cases} \tag{94}$$

Instead of computing the integrated likelihood from scratch, we use the fact that the variable transformation $y = \ln(x/x_m)$ yields

$$P(y|\alpha) = \alpha \mathrm{e}^{-\alpha y}, \tag{95}$$

which is the exponential distribution we used before. So when dealing with broad weights, we can just make this transformation on the weights and use the exponential model.

Alternatively, we may use the normal model for $y = \ln x$, which assumes that x is distributed according to a log-normal. In our experience, we found that this choice also typically yields better results when the positive weights are peaked around a typical value, in a manner that is difficult to represent with a mixture of exponential distributions.

2. *Bounded weights*

If the weights are bounded in an interval $x \in [a,b]$, we can adapt it to an unbounded distribution by first uniformly mapping the weights to the unit interval $x' \in [0,1]$, via

$$x' = \frac{x-a}{b-a}, \tag{96}$$

and then using a logit transformation

$$y = \ln\left(\frac{x'}{1-x'}\right), \tag{97}$$

or, equivalently, first mapping to the symmetric interval $x' \in [-1,1]$, via

$$x' = 2\frac{x-a}{b-a} - 1, \tag{98}$$

and using the inverse hyperbolic tangent

$$y = 2\,\mathrm{arctanh}(x') = \ln\left(\frac{1+x'}{1-x'}\right), \tag{99}$$

both of which yield the same signed unbounded weight $y \in [-\infty,\infty]$, which can be fit using the normal distribution. Alternatively, the negative logarithm can be used with Eq. 96

$$z = -\ln x', \tag{100}$$

which yields a positive unbounded weight $z \in [0,\infty]$ that can be used with the exponential distribution. Which approach is most suitable depends on the actual shape of the data, and can be determined *a posteriori* via model selection, as described in Sec. IV G.

3. *Decomposing covariates*

We can also obtain more elaborate models by decomposing a single covariate into multiple ones. Consider, for example, the case of signed discrete weights $x \in [\dots, -2, -1, 0, 1, 2, \dots]$, which was not considered directly by any of the models so far. This can be done in a straightforward manner by decomposing the numbers into a sign and magnitude, i.e.

$$x_{ij} = (2s_{ij} - 1)y_{ij} \tag{101}$$

where

$$s_{ij} = (\mathrm{sgn}(x_{ij}) + 1)/2, \tag{102}$$

$$y_{ij} = \mathrm{abs}(x_{ij}) \tag{103}$$

is a reversible transformation that extracts the sign and absolute values of x_{ij}. We may then use a Binomial distribution with $M = 1$ (i.e. Bernoulli) for $s_{ij} \in \{0, 1\}$, and any non-negative distribution for $y_{ij} \in \{0, 1, 2, \dots\}$, and obtain the posterior using the joint marginal likelihood

$$P(\boldsymbol{x}|\boldsymbol{A}, \{\boldsymbol{b}^l\}) = P(\boldsymbol{y}, \boldsymbol{s}|\boldsymbol{A}, \{\boldsymbol{b}^l\}) \tag{104}$$

$$= P(\boldsymbol{y}|\boldsymbol{A}, \{\boldsymbol{b}^l\})P(\boldsymbol{s}|\boldsymbol{A}, \{\boldsymbol{b}^l\}). \tag{105}$$

G. Model selection

Given any two models $\mathcal{M}_1$ and $\mathcal{M}_2$ for the same weighted network with edge covariates $\boldsymbol{x}$, for which we obtain the partitions $\{\boldsymbol{b}^l\}_1$ and $\{\boldsymbol{b}^l\}_2$ from their respective posterior distributions, we can perform model selection as described in Ref. [6], by computing the posterior odds ratio

$$\Lambda = \frac{P(\{\boldsymbol{b}^l\}_1, \mathcal{M}_1|\boldsymbol{A}, \boldsymbol{x})}{P(\{\boldsymbol{b}^l\}_2, \mathcal{M}_2|\boldsymbol{A}, \boldsymbol{x})} \tag{106}$$

$$= \frac{P(\boldsymbol{A}|\{\boldsymbol{b}^l\}_1, \mathcal{M}_1)P(\boldsymbol{x}|\boldsymbol{A}, \{\boldsymbol{b}^l\}_1, \mathcal{M}_1)P(\{\boldsymbol{b}^l\}_1)P(\mathcal{M}_1)}{P(\boldsymbol{A}|\{\boldsymbol{b}^l\}_2, \mathcal{M}_2)P(\boldsymbol{x}|\boldsymbol{A}, \{\boldsymbol{b}^l\}_2, \mathcal{M}_2)P(\{\boldsymbol{b}^l\}_2)P(\mathcal{M}_2)}, \tag{107}$$

where $P(\mathcal{M})$ is the prior preference for either model [typically, we are agnostic with $P(\mathcal{M}_1) = P(\mathcal{M}_2)$]. For values of $\Lambda > 1$, the choice $(\{\boldsymbol{b}^l\}_1, \mathcal{M}_1)$ is preferred over $(\{\boldsymbol{b}^l\}_2, \mathcal{M}_2)$ according to the data, and the magnitude of Λ yields the degree of statistical significance.

Using this criterion we can select between unweighted variations of the SBM (e.g. degree-corrected or not) [6], but also between different models of the weights. This is particularly useful when using weight transformations as described in Sec. IV F. For example, when considering two different transformations $y_{ij} = f(x_{ij})$ and $z_{ij} = g(x_{ij})$, using different models $\mathcal{M}_y$ and $\mathcal{M}_z$ for the transformed covariates, the posterior odds ratio [with agnostic priors $P(\mathcal{M}_y) = P(\mathcal{M}_z)$] becomes

$$\Lambda = \frac{P(\boldsymbol{A}, \boldsymbol{y}(\boldsymbol{x})|\boldsymbol{A}, \{\boldsymbol{b}^l\}_1, \mathcal{M}_y)P(\{\boldsymbol{b}^l\}_1)\prod_{i<j} f'(x_{ij})^{A_{ij}}}{P(\boldsymbol{A}, \boldsymbol{z}(\boldsymbol{x})|\boldsymbol{A}, \{\boldsymbol{b}^l\}_2, \mathcal{M}_z)P(\{\boldsymbol{b}^l\}_2)\prod_{i<j} g'(x_{ij})^{A_{ij}}}. \tag{108}$$

Transformation	Derivatives	Weight model	$\ln P(\boldsymbol{A}, \boldsymbol{x}, \{\boldsymbol{b}^l\})$
$y_{ij} = x_{ij}$	1	Exponential	$-56,512$
$y_{ij} = \ln x_{ij}$	$\prod_{i<j} 1/x_{ij}^{A_{ij}}$	Normal	$-52,054$

Table I. Joint log-likelihood $\ln P(\boldsymbol{A}, \boldsymbol{x}, \{\boldsymbol{b}^l\})$ for the human brain data in Sec. III C using the electrical connectivity as edge covariate, for two model variations according to weight transformations.

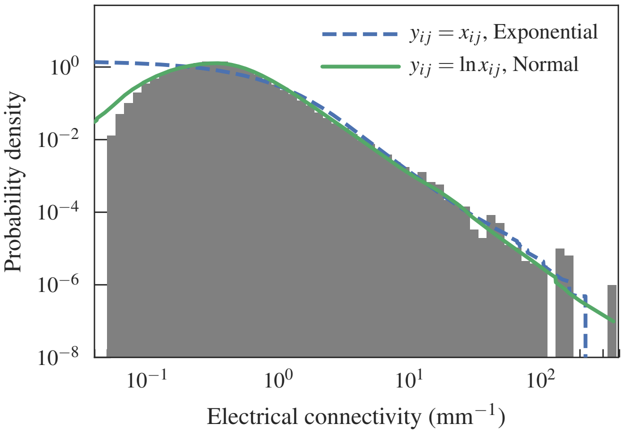

Figure 11. Overall distribution of the electrical connectivity of the human brain data. The solid lines shows the inferred distribution according to the weighted SBM using two models for the edge covariates, as shown in the legend.

The approach is entirely analogous for transformations on discrete weights, where one simple omits the derivative terms.

We illustrate the use of this criterion on the human brain data analyzed in Sec. III C. We consider here only the electric connectivity covariate, which is non-negative and unbounded in the range $[0, \infty]$. We consider two models for the weights: The first is the exponential model of Sec. IV A applied directly to the original covariates, and the second is the normal model of Sec. IV B, applied to the transformed weights $y_{ij} = \ln x_{ij}$, which results in a log-normal model for x_{ij}. As the results of Table I show, we obtain for this dataset a posterior odds ratio of $\ln \Lambda \approx 4,458$ favoring the log-normal model, despite the fact that it contains more internal parameters. As we see in Fig. 11, indeed the log-normal model is better suited to capture the peaked nature of the overall distribution. It should be noted that while it is a trivial feat to obtain better fits with more complicated models, the Bayesian criterion above takes into account the complexity of the model, and will point towards a more complicated one only if the statistical evidence in the data supports it.

V. CONCLUSION

The weighted extensions of the SBM presented in this work allow for a principled inference of large-scale modular structure of weighted networks, in a manner that is fully nonparametric, and algorithmically efficient. As they include a hierarchical description of the network — taking into account both the node adjacency as well as the edge weights — our SBM implementations enable the detection of modular structures at multiple scales, without being biased towards any specific kind of mixing pattern (such as assortativity) in any of them.

The nonparametric nature of our approach means that it can be used to detect the most appropriate model dimension, including the number of groups as well as size and shape of the hierarchical division, directly from data, in a parsimonious way, without requiring any prior input. This comes with the guarantee that the inferred hierarchy is statistically significant, and hence is not the result of statistical fluctuations of a simpler model (such as a completely random graph).

The edge weights are included in the model description as additional covariates, and thus require specific models that reflect their nature. The explicit variations presented in this work cover a broad range of possible types of covariates, that can be either continuous or discrete, signed or unsigned, bounded or unbounded. Furthermore, all these particular variations can be arbitrarily extended to accommodate a much wider class of weight models via variable transformations, which incur no modification to the algorithms. Such transformations can be performed in an *ad hoc* manner, reflecting the specificity of the data at hand, and the best choice can be evaluated *a posteriori* using Bayesian model selection, simultaneously taking into account the quality of fit, the model complexity and the statistical significance available from the data.

Although we do not describe this in detail here, it is easy to see that the exact same approach we present can be used for other variations of the SBM, such as with overlapping groups [40, 41], edge layers [42–44] and dynamic networks [42, 45, 46].

Despite its advantages, our approach inherits the limitations of the underlying SBM ansatz. In particular, it assumes that the weights are distributed on the edges in a manner that is (asymptotically, in the microcanonical case) conditionally independent. Hence, in the same manner that the unweighted SBM does not include the often realistic propensity of the network to form triangles and other local structures, the weighted extensions preclude the existence of certain kinds of weight correlations that are known to exist in key cases [47]. The development of tractable and versatile models that incorporate such higher-order aspects remains an open challenge.

Appendix A: Directed networks

Although we focused on undirected networks in the main text, our methods can be easily adapted to directed networks. The models for directed adjacency matrices $P(\boldsymbol{A}|\{\boldsymbol{b}^l\})$ are described in detail in Ref [6]. For the edge covariates, the modifications are straightforward yielding expressions for $P(\boldsymbol{x}|\boldsymbol{A},\{\boldsymbol{b}^l\})$ that are identical, but with products going over directed pairs of groups and nodes, i.e. $\prod_{r\le s}\to\prod_{rs}$ and $\prod_{i\le j}\to\prod_{ij}$. Our reference implementation supports these variations [17].

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

[2] Tiago P. Peixoto, “Bayesian stochastic blockmodeling,” arXiv:1705.10225 [cond-mat, physics:physics, stat] (2017), arXiv: 1705.10225.

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

[4] Christopher Aicher, Abigail Z. Jacobs, and Aaron Clauset, “Learning latent block structure in weighted networks,” Journal of Complex Networks **3**, 221–248 (2015).

[5] Tiago P. Peixoto, “Hierarchical Block Structures and High-Resolution Model Selection in Large Networks,” Physical Review X **4**, 011047 (2014).

[6] Tiago P. Peixoto, “Nonparametric Bayesian inference of the microcanonical stochastic block model,” Physical Review E **95**, 012317 (2017).

[7] Tiago P. Peixoto, “Parsimonious Module Inference in Large Networks,” Physical Review Letters **110**, 148701 (2013).

[8] Brian Karrer and M. E. J. Newman, “Stochastic blockmodels and community structure in networks,” Physical Review E **83**, 016107 (2011).

[9] 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).

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

[11] 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).

[12] M. E. J. Newman and Gesine Reinert, “Estimating the Number of Communities in a Network,” Physical Review Letters **117**, 078301 (2016).

[13] E. T. Jaynes, *Probability Theory: The Logic of Science*, edited by G. Larry Bretthorst (Cambridge University Press, Cambridge, UK ; New York, NY, 2003).

[14] 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).

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

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

[17] Tiago P. Peixoto, "The `graph-tool` python library," figshare (2014), 10.6084/m9.figshare.1164194, available at `https://graph-tool.skewed.de`.

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

[19] Data available at `https://www.un.org/en/development/desa/population/migration/data/estimates2/estimates15.shtml`.

[20] Data available from the official website `http://www2.camara.leg.br/`.

[21] Balázs Szalkai, Csaba Kerepesi, Bálint Varga, and Vince Grolmusz, "Parameterizable consensus connectomes from the Human Connectome Project: the Budapest Reference Connectome Server v3.0," Cognitive Neurodynamics **11**, 113–116 (2017).

[22] Jennifer A. McNab, Brian L. Edlow, Thomas Witzel, Susie Y. Huang, Himanshu Bhat, Keith Heberlein, Thorsten Feiweier, Kecheng Liu, Boris Keil, Julien Cohen-Adad, M. Dylan Tisdall, Rebecca D. Folkerth, Hannah C. Kinney, and Lawrence L. Wald, "The Human Connectome Project and beyond: Initial applications of 300mt/m gradients," NeuroImage Mapping the Connectome, **80**, 234–245 (2013).

[23] Peter J. Basser and Carlo Pierpaoli, "Microstructural and physiological features of tissues elucidated by quantitative-diffusion-tensor MRI," Journal of Magnetic Resonance Magnetic Moments, **213**, 560–570 (1996).

[24] David Meunier, Renaud Lambiotte, and Edward T. Bullmore, "Modular and Hierarchically Modular Organization of Brain Networks," Frontiers in Neuroscience **4** (2010), 10.3389/fnins.2010.00200.

[25] Danielle S. Bassett, Mason A. Porter, Nicholas F. Wymbs, Scott T. Grafton, Jean M. Carlson, and Peter J. Mucha, "Robust detection of dynamic community structure in networks," Chaos: An Interdisciplinary Journal of Nonlinear Science **23**, 013142 (2013).

[26] Richard F. Betzel, Alessandra Griffa, Andrea Avena-Koenigsberger, Joaquín Goñi, Jean-Philippe Thiran, Patric Hagmann, and Olaf Sporns, "Multi-scale community organization of the human structural connectome and its relationship with resting-state functional connectivity," Network Science **1**, 353–373 (2013).

[27] M. E. J. Newman, "Mixing patterns in networks," Phys. Rev. E **67**, 026126 (2003).

[28] 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).

[29] 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).

[30] Jörg Reichardt and Stefan Bornholdt, "When are networks truly modular?" Physica D: Nonlinear Phenomena **224**, 20–26 (2006).

[31] James P. Bagrow, "Communities and bottlenecks: Trees and treelike networks have high modularity," Physical Review E **85**, 066118 (2012).

[32] Colin McDiarmid and Fiona Skerman, "Modularity in random regular graphs and lattices," Electronic Notes in Discrete Mathematics **43**, 431–437 (2013).

[33] Colin McDiarmid and Fiona Skerman, "Modularity of tree-like and random regular graphs," ArXiv e-prints **1606**, arXiv:1606.09101 (2016).

[34] Liudmila Ostroumova Prokhorenkova, Paweł Prałat, and Andrei Raigorodskii, "Modularity of Complex Networks Models," in *Algorithms and Models for the Web Graph*, Lecture Notes in Computer Science (Springer, Cham, 2016) pp. 115–126.

[35] Santo Fortunato and Marc Barthélemy, "Resolution limit in community detection," Proceedings of the National Academy of Sciences **104**, 36–41 (2007).

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

[37] Richard F. Betzel, John D. Medaglia, and Danielle S. Bassett, "Diversity of meso-scale architecture in human and non-human connectomes," ArXiv e-prints **1702**, arXiv:1702.02807 (2017).

[38] George E. P. Box and George C. Tiao, *Bayesian Inference in Statistical Analysis* (John Wiley & Sons, 2011) google-Books-ID: T8Askeyk1k4C.

[39] Antonia Godoy-Lorite, Roger Guimerà, Cristopher 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).

[40] Brian Ball, Brian Karrer, and M. E. J. Newman, "Efficient and principled method for detecting communities in networks," Physical Review E **84**, 036103 (2011).

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

[42] Tiago P. Peixoto, "Inferring the mesoscale structure of layered, edge-valued, and time-varying networks," Physical Review E **92**, 042807 (2015).

[43] N. Stanley, S. Shai, D. Taylor, and P. J. Mucha, "Clustering Network Layers with the Strata Multilayer Stochastic Block Model," IEEE Transactions on Network Science and Engineering **3**, 95–105 (2016).

[44] Simon Heimlicher, Marc Lelarge, and Laurent Massoulié, "Community Detection in the Labelled Stochastic Block Model," arXiv:1209.2910 (2012).

[45] Kevin S. Xu and Alfred O. Hero Iii, "Dynamic Stochastic Blockmodels: Statistical Models for Time-Evolving Networks," in *Social Computing, Behavioral-Cultural Modeling and Prediction*, Lecture Notes in Computer Science No. 7812, edited by Ariel M. Greenberg, William G. Kennedy, and Nathan D. Bos (Springer Berlin Heidelberg, 2013) pp. 201–210.

[46] Tiago P. Peixoto and Martin Rosvall, "Modelling sequences and temporal networks with dynamic community structures," Nature Communications **8**, 582 (2017).

[47] Mel MacMahon and Diego Garlaschelli, "Community Detection for Correlation Matrices," Physical Review X **5**, 021006 (2015).