# Model selection and hypothesis testing for large-scale network models with overlapping groups

Tiago P. Peixoto*

*Institut für Theoretische Physik, Universität Bremen, Hochschulring 18, D-28359 Bremen, Germany*

The effort to understand network systems in increasing detail has resulted in a diversity of generative models that describe large-scale structure in a variety of ways, and allow its characterization in a principled and powerful manner. Current models include features such as degree correction, where nodes with arbitrary degrees can belong to the same group, and community overlap, where nodes are allowed to belong to more than one group. However, such elaborations invariably result in an increased number of parameters, which makes these model variants prone to overfitting. Without properly accounting for the increased model complexity, one should naturally expect these larger models to "better" fit empirical networks, regardless of the actual statistical evidence supporting them. Here we propose a principled method of model selection based on the minimum description length principle and posterior odds ratios that is capable of fully accounting for the increased degrees of freedom of the larger models, and selects the best model according to the statistical evidence available in the data. Contrary to other alternatives such as likelihood ratios and parametric bootstrapping, this method scales very well, and combined with efficient inference methods recently developed, allows for the analysis of very large networks with an arbitrarily large number of groups. In applying this method to many empirical datasets from different fields, we observed that while degree correction tends to provide better fits for a majority of networks, community overlap does not, and is selected as better model only for a minority of them.

## I. INTRODUCTION

Many networks possess non-trivial large-scale structures such as communities [1, 2], core-peripheries [3, 4], bi-partitions [5] and hierarchies [6, 7]. These structures presumedly reflect the organizational principles behind their formation. Furthermore, their detection can be used to predict missing links [6, 8], detect spurious ones [8], determine the robustness of the system to failure or intentional damage [9], the outcome of the spread of epidemics [10] and functional classification [11], among many other applications. The detail with which such modular features are both represented in generative models, and detected with inference algorithms, reflects directly on the quality of these tasks. Hence, a natural undertaking has been the development of more elaborate models, which include degree correction [12], community overlap [13, 14], hierarchical structure [6, 7], self-similarity [15, 16], bipartiteness [5], edge and node correlates [17, 18], social tiers [19], multilayer structure [20], temporal information [21], to name only a few. Although such developments are essential, they should be made with care, since increasing the complexity of generative models may lead to artificial results caused by overfitting. While this is a well understood phenomenon when dealing with independent data or time series, open problems remain when the empirical data is a network, for which many common assumptions no longer hold and usual methods perform very poorly [22]. This problem is significantly exacerbated when methods are used which make no attempt to assess the statistical significance of the results. Unfortunately, the most widely used methods fall into this class, such as modularity maximization [23], link similarity [24], clique percolation [25], and many others [2]. Although for certain specially constructed examples some direct connections between statistical inference and ad hoc methods can be made [26, 27], and in the case of some spectral methods a much deeper connection seems to exist [28, 29], they still inherently lack the capacity to reliably distinguish signal from noise. Furthermore, what is perhaps even more important, these different methods cannot easily be compared to *each other*. For example, if a non-overlapping partition is found with some spectral method, another overlapping partition is obtained with clique percolation, and yet another with a local method based on link similarity, most of the time they will be very different, and yet there is no obvious way to decide which one is a more faithful representation of the network. We show in this work that this issue can be solved in a consistent and principled manner by restricting oneself to generative models, and by performing model selection based on statistical evidence. In particular, we employ the minimum description length principle (MDL), which seeks to minimize the total information necessary to describe the observed data as well as the model parameters. This can be equivalently formulated as the maximization of a Bayesian posterior likelihood which includes non-informative priors on the parameters, from which a posterior odds ratio between different hypothesis can be computed, yielding a *degree of confidence* for a model to be rejected in favor of another. We focus on the stochastic block model as the underlying generative model, as well as variants which include degree correction and mixed memberships. We show that with these models MDL can be used to produce a very efficient algorithm which scales well for very large networks

* tiago@itp.uni-bremen.de

and with arbitrarily large number of groups. Furthermore we employ the method to a wide variety of network datasets, and we show that degree correction tends to be selected for a significant majority of them, whereas community overlaps are seldom selected. This casts doubt on the claimed pervasiveness of group overlaps [24, 25], obtained predominantly with nonstatistical methods, which — as long as there is a lack of corroborating evidence other than the network structure supporting the overlap — should perhaps be interpreted as an artifact of using methods with more degrees of freedom, instead of an underlying property of many systems.

This paper is divided as follows. In Sec. II we present the generative models considered, and in Sec. III we describe the model selection procedure based on MDL. In Sec. IV we present the results for a variety of empirical networks. In Sec. V we analyze the general identifiability limits of the overlapping models, and in Sec. VI we describe in detail the inference algorithm used. In Sec. VII we finalize with a discussion.

## II. GENERATIVE MODELS FOR NETWORK STRUCTURE

A generative model is one which attributes to each possible graph $G$ a probability $P(G|\{\theta\})$ for it to be observed, conditioned on some set of parameters $\{\theta\}$. Here we will be restricted to discrete models, where specific choices of $\{\theta\}$ prohibit some graphs from occurring, but those which are allowed to occur have the same probability. For these models we can write $P(G|\{\theta\}) = \mathbf{1}_{\{\theta\}}(G)/\Omega(\{\theta\}) = e^{-\mathcal{S}(G|\{\theta\})}$, with $\Omega(\{\theta\})$ being the total number of possible graphs compatible with a given choice of parameters, $\mathbf{1}_{\{\theta\}}(G)$ is the indicator function with value one if the graph $G$ belongs to the ensemble constrained by $\{\theta\}$ or zero otherwise, and $\mathcal{S}(G|\{\theta\}) = \ln\Omega(\{\theta\}) - \ln\mathbf{1}_{\{\theta\}}(G)$ is the entropy of this constrained ensemble (where it should be understood simply that if $\mathbf{1}_{\{\theta\}}(G) = 0$, then $\mathcal{S}(G|\{\theta\})$ is undefined, since the graph has zero probability) [30, 31]. In order to infer the parameters $\{\theta\}$ via maximum likelihood, we need to maximize $P(G|\{\theta\})$, or equivalently, minimize $\mathcal{S}(G|\{\theta\})$. This approach, however, cannot be used if the *order* of the model is unknown, i.e. the number of degrees of freedom in the parameter set $\{\theta\}$, since choices with higher order will almost always increase $P(G|\{\theta\})$, resulting in overfitting. For the same reason, maximum likelihood cannot be used to distinguish between models belonging to different classes, since models with larger degrees of freedom will inherently lead to larger likelihoods. In order to avoid overfitting, one needs to maximize instead the Bayesian posterior probability $P(\{\theta\}|G) = P(G|\{\theta\})P(\{\theta\})/P(G)$, with $P(G)$ being a normalizing constant. The prior probability $P(\{\theta\})$, which encodes our a priori knowledge of the parameters (if any) should inherently become smaller if the number of degrees of freedom increases. We will also be restricted to discrete parameters with constant prior probabilities, so that $P(\{\theta\}) = e^{-\mathcal{L}(\{\theta\})}$, with $\mathcal{L}(\{\theta\})$ being the entropy of the ensemble of possible parameter choices. We can thus write the total posterior likelihood as $P(\{\theta\}|G) = e^{-\Sigma}$, with $\Sigma = \mathcal{L}(\{\theta\}) + \mathcal{S}(G|\{\theta\})$. The value $\Sigma$ is the *description length* of the data [32, 33], i.e. the total amount of information required to describe the observed data conditioned on a set of parameters as well as the parameter set itself [34]. Hence, if we maximize $P(\{\theta\}|G)$ we are automatically finding the parameter choice which most *compresses* the data, since it will also minimize its description length $\Sigma$. Because of this, there is no difference between specifying probabilistic models for both $G$ and $\{\theta\}$, or encoding schemes that quantify the amount of information necessary to describe both. In the following, we will use both terminologies interchangeably.

### A. Model without degree correction

Here we consider a simple variation of the stochastic block model [35–38] with $N$ nodes and $E$ edges, where the nodes can belong to different groups. Hence, to each node we attribute a binary vector $\vec{b}_i$ with $B$ entries, where a given entry $b_i^r \in \{0,1\}$ specifies whether or not the node belongs to block $r \in [1, B]$. In addition to this overlapping partition, we simply define the edge-count matrix $\{e_{rs}\}$, which specifies how many edges are placed between nodes belonging to blocks $r$ and $s$ (or twice that number for $r = s$, for convenience of notation), where we have $\sum_{rs} e_{rs} = 2E$. This simple definition allows one to generate a broad variety of overlapping patterns, which are not confined to purely assortative structures, and the non-overlapping model can be recovered as a special case, simply by putting each node in a single group.

The posterior likelihood of observing a given graph with the above constraints is simply $P(G|\{\vec{b}_i\}, \{e_{rs}\}) = 1/\Omega(\{\vec{b}_i\}, \{e_{rs}\})$, where $\Omega(\{\vec{b}_i\}, \{e_{rs}\})$ is the number of possible graphs. In this construction, the existence of multiple edges is allowed. However, the placement of multiple edges between nodes of blocks $r$ and $s$ should occur with a probability proportional to $O(e_{rs}/n_r n_s)$, where $n_r$ is the number of nodes which belong to block $r$, i.e. $n_r = \sum_i b_i^r$ (note that $\sum_r n_r \geq N$). Since here we are predominantly interested in the sparse situation where $e_{rs} \sim O(N/B^2)$ and $n_r \sim O(N/B)$, the probability of observing parallel edges will decay as $O(1/N)$, and hence can be neglected in the large network limit. Making use of this simplification, we may approximately count all possible graphs generated by the parameters $\{\vec{b}_i\}, \{e_{rs}\}$ as the number of graphs where each distinct membership of a single node is considered to be a different node with a single membership. This corresponds to an augmented graph generated via a non-overlapping block model with $N' = \sum_r n_r$ nodes, where $N' \geq N$, but with the same matrix $\{e_{rs}\}$, for which the entropy is [31]

$$S_t \simeq E - \frac{1}{2}\sum_{rs} e_{rs}\ln\left(\frac{e_{rs}}{n_r n_s}\right), \tag{1}$$

where $S_t = \ln\Omega(\{\vec{b}_i\},\{e_{rs}\})$, and $n_r n_s \gg e_{rs}$ was assumed. Under this formulation, we recover trivially the single-membership case simply by assigning each node to a single group, since Eq. 1 is identical in that special case. It is possible to remove the approximation that no parallel edges occur, by defining the model somewhat differently, as in shown in appendix B 1, in which case the Eq. 1 holds exactly as long as no parallel edges are observed.

Like its non-overlapping counterpart, the block model without degree correction assumes that nodes belonging to the same group will receive approximately the same number of edges of that type. Hence, when applied to empirical data, the modules discovered will also tend to have this property. This means that if the graph possesses large degree variability, the groups inferred will tend to correspond to different degree classes. On a similar vein, if a node belongs to more then one group, it will also tend to have a total degree which is larger than nodes that belong to either group alone, since it will receive edges of each type in an independent fashion. In other words, the group intersections are expected to be strictly *denser* than the non-overlapping portions of each group. Note that in this respect this model differs from other popular ones, such as the mixed membership stochastic block model (MMSBM) [13], where the density at the intersections is the weighted average of the groups (see appendix B 1).

### B. Model with degree correction

In a manner analogous to the non-overlapping model [12], a multiple membership version of the stochastic block model with degree correction can be defined. This can be achieved simply by specifying, in addition to the overlapping partition $\{\vec{b}_i\}$, the number of half-edges incident on a given node $i$ which belong to group $r$, i.e. $k_i^r$. Given this labelled degree sequence, one can simply use the same edge count matrix $\{e_{rs}\}$ as before to generate the graph. If we again make the assumption that the occurrence of parallel edges can be neglected, the total number of graphs fulfilling these constrains is approximately equal to the non-overlapping ensemble where each set of half-edges incident on any given node $i$ that belongs to the same group $r$ is considered as an individual node with degree $k_i^r$, for which the entropy is [31]

$$S_d \simeq -E - \frac{1}{2}\sum_{rs} e_{rs}\ln\left(\frac{e_{rs}}{e_r e_s}\right) - \sum_{ir}\ln k_i^r!, \tag{2}$$

where $e_{rs}(\left\langle k^2\right\rangle_r - \langle k\rangle_r)(\left\langle k^2\right\rangle_s - \langle k\rangle_s)/\langle k\rangle_r^2\langle k\rangle_s^2 \ll n_r n_s$ has been assumed. Similarly to the non-degree-corrected case, it is possible to remove the approximation that no parallel edges occur, by using a "Poisson" version of the model, as is shown in appendix B 2. Under this formulation, it can be shown that this model is equivalent to the one proposed by Ball et al [14], although here we keep track of the individual labels on the half-edges as latent variables, instead of their probabilities.

Since we incorporate the labelled degree sequence as parameters to the model, nodes which belong to the same group can have arbitrary degrees. Furthermore, since the same applies to nodes which belong simultaneously to more than one group, the overlap between groups are neither preferably dense nor preferably sparse; it all depends on the parameters $\{k_i^r\}$.

## III. MODEL SELECTION

As discussed previously, in order to perform model selection, it is necessary to include the information necessary to describe the model parameters, in addition to the data. The parameters which need to be described are the overlapping partition $\{\vec{b}_i\}$, the edge counts $\{e_{rs}\}$, and in the case of the degree-corrected model we also need to the describe the labeled degree sequence $\{k_i^r\}$.

When choosing an encoding for the parameters we need to avoid redundancy, and describe them as parsimoniously as possible. This means that we need to make few prior assumptions, in order to be able to use observed patterns in the data to compress the parameter description as much as possible. In Bayesian language, we need non-informative priors which express maximal uncertainty about the parameters. On the other hand, we need to fully exploit known or intrinsic constraints, since they should not be learned from the data.

In order to specify the partition $\{\vec{b}_i\}$, we assume that all different $2^B - 1$ mixtures are not necessarily equally likely, and furthermore the sizes $d_i = \sum_r b_i^r$ of the mixtures are also not a priori assumed to follow any specific distribution. More specifically, we consider the mixtures to be the outcome of a generative process with two steps. We first generate the local mixture sizes $\{d_i\}$, which are sampled uniformly from the distribution with fixed counts $\{n_d\}$, such that its description length is

$$\mathcal{L}_d = \ln\left(\!\!\binom{D}{N}\!\!\right) + \ln N! - \sum_d \ln n_d!, \tag{3}$$

where $D$ is the maximum value of $d$, and $\left(\!\binom{D}{N}\!\right)$ is the total number of different choices of $\{n_d\}$, with $\left(\!\binom{n}{m}\!\right) = \binom{n+m-1}{m}$ being the total number of $m$-combinations with repetitions from a set of size $n$. Then, for all $n_d$ nodes with the same value of $d$, we sample a sequence of $\{\vec{b}_i\}$ from a distribution with support $|\vec{b}_i|_1 \equiv \sum_r b_i^r = d$ and with fixed counts $n_{\vec{b}}$, which has a description length

$$\mathcal{L}_b = \sum_d \left[ \ln \left(\!\!\binom{\binom{B}{d}}{n_d}\!\!\right) + \ln n_d! - \sum_{|\vec{b}|_1 = d} \ln n_{\vec{b}}! \right], \tag{4}$$

where, similarly, $\left(\!\!\binom{\binom{B}{d}}{n_d}\!\!\right)$ enumerates the total number of $\{n_{\vec{b}}\}$ counts with $|\vec{b}|_1 = d$. The whole description length $\mathcal{L}_p = \mathcal{L}_d + \mathcal{L}_b$ becomes

$$\mathcal{L}_p = \ln \left(\!\!\binom{D}{N}\!\!\right) + \sum_d \ln \left(\!\!\binom{\binom{B}{d}}{n_d}\!\!\right) + \ln N! - \sum_{\vec{b}} \ln n_{\vec{b}}!. \tag{5}$$

Although it is possible to encode the partition in different ways (e.g. by sampling the membership to each group independently [39]), this choice makes no assumptions regarding the types of overlaps which are more likely to occur, either according to the number of groups to which a node may belong, or the actual combination of groups — it is all left to be learned from data. In particular, it is not a priori assumed that if many nodes belong to groups $r$ and $s$ then the overlap between these two groups will also contain many nodes. As desired, if the observed partition deviates from this pattern, this will be used to compress it further. Only if the observed partition falls squarely into this pattern will further compression not be possible, and we would have an overhead describing it using Eq. 5, when compared to an encoding which expects it a priori. However, one can also see that in the limit $n_{\vec{b}} \gg 1$, as the first two terms in Eq. 5 grow asymptotically only with $\ln N$ and $\ln n_d$, respectively, the whole description length becomes $\mathcal{L}_p \simeq NH(\{n_{\vec{b}}/N\})$, where $H(\{p_x\})$ is the entropy of the distribution $\{p_x\}$, which is the optimal limit. Hence if we have a prior which better matches the observed overlap, the difference in description length compared to Eq. 5 will disappear asymptotically for large systems. Another advantage of this encoding is that it incurs no overhead when there are no overlaps at all (i.e. $D = 1$), and in this case the description length is identical to the non-overlapping case,

$$\mathcal{L}_p(D=1) = \ln \left(\!\!\binom{B}{N}\!\!\right) + \ln N! - \sum_r \ln n_r!, \tag{6}$$

as defined in Ref. [7].

### A. Degree correction

For the degree-corrected model, we need to describe the labeled degree sequence $\{\vec{k}_i\}$. We need to do so in a way which is compatible with the partition $\{\vec{b}_i\}$ so far described, and with edge counts $\{e_{rs}\}$, which will restrict the average degrees of each type.

In order to fully utilize the partition $\{\vec{b}_i\}$, we describe for each value of $\vec{b}$ its individual degree sequence $\{\vec{k}_i | \vec{b}_i = \vec{b}\}$, via the counts $n^{\vec{b}}_{\vec{k}}$, i.e. the number of nodes in partition $\vec{b}$ which possess labeled degrees $\vec{k}$. We do so in order to preserve the lack of preference for patterns involving the degrees in the overlaps between groups. Since the model itself is agnostic with respect to the density of the overlaps, not only does this choice remain consistent with this, but also any existing pattern in the degree sequence in the overlaps will be used to construct a shorter description.

In addition, we must also consider the total number of half-edges of a given type $r$ incident on a partition $e^r_{\vec{b}} = \sum_{\vec{k}} k_r n^{\vec{b}}_{\vec{k}}$, which must be compatible with the edge counts $\{e_{rs}\}$ via $e_r = \sum_s e_{rs} = \sum_{\vec{b}} e^r_{\vec{b}}$.

We begin by first distributing all the $e_r$ half-edges of type $r$ in all the $m_r$ bins corresponding to each non-empty $\vec{b}$ partition which contains the label $r$, i.e. $m_r = \sum_{\vec{b}} b_r [n_{\vec{b}} > 0]$. The total number of such partitions is simply $\left(\!\!\binom{m_r}{e_r}\!\!\right)$. Now we need to distribute the labelled half-edges inside each partition to obtain each degree sequence. The logarithm of the total number of degree sequences fulfilling all necessary constraints is

$$\mathcal{L}^{(1)}_{\vec{b}} = \sum_r \ln \left(\!\!\binom{n_{\vec{b}}}{e^r_{\vec{b}}}\!\!\right). \tag{7}$$

However, most degree sequences uniformly sampled from this set will result in nodes with very similar degrees. Since we want to profit from degree variability, it is better to condition the description on the counts $n^{\vec{b}}_{\vec{k}}$, i.e. how many nodes in partition $\vec{b}$ possess labelled degree $\vec{k}$. The description in this case becomes

$$\mathcal{L}^{(2)}_{\vec{b}} = \sum_r b_r \ln \Xi^r_{\vec{b}} + \ln n_{\vec{b}}! - \sum_{\vec{k}} \ln n^{\vec{b}}_{\vec{k}}!, \tag{8}$$

where $\Xi^r_{\vec{b}}$ is the enumeration of all possible $n^{\vec{b}}_{\vec{k}}$ counts which fulfill the constraints $\sum_{\vec{k}} n^{\vec{b}}_{\vec{k}} = n_{\vec{b}}$ and $\sum_{\vec{k}} k_r n^{\vec{b}}_{\vec{k}} = e^r_{\vec{b}}$. Unfortunately, this enumeration cannot be done easily in closed form. However, the maximum entropy ensemble where these constraints are enforced *on average* is analytically tractable, and as we show in appendix C, can be very well approximated by

$$\ln \Xi^r_{\vec{b}} \simeq 2\sqrt{\zeta(2) e^r_{\vec{b}}}, \tag{9}$$

where $\zeta(x)$ is the Riemann zeta function. This approximation with “soft” constraints should become asymptotically exact as the number of nodes become large, but otherwise will deviate from the actual entropy. On the other hand, if the number of nodes is very small, describing the degree sequence via Eq. 8 may not provide a shorter description, even if computed exactly. In this situation, Eq. 7 may actually provide a shorter description of the degree sequence. We therefore compute both Eq. 7 and Eq. 8 and choose whatever is shorter. Putting it all together, the description length for the whole degree sequence becomes

$$\mathcal{L}_\kappa = \sum_r \ln \left(\!\!\binom{m_r}{e_r}\!\!\right) + \sum_{\vec{b}} \min\left(\mathcal{L}^{(1)}_{\vec{b}}, \mathcal{L}^{(2)}_{\vec{b}}\right). \tag{10}$$

In the limit $n^{\vec{b}}_{\vec{k}} \gg 1$, we have that $\mathcal{L}_\kappa \simeq \sum_{\vec{b}} H(\{n^{\vec{b}}_{\vec{k}}/n_{\vec{b}}\})$, and hence the degree sequences in each partition are described close to the optimum.

For the non-overlapping case with $D = 1$, the description length simplifies to

$$\mathcal{L}_\kappa = \sum_r \min\left(\mathcal{L}^{(1)}_r, \mathcal{L}^{(2)}_r\right), \tag{11}$$

with

$$\mathcal{L}^{(1)}_r = \ln \left(\!\!\binom{n_r}{e_r}\!\!\right), \tag{12}$$

$$\mathcal{L}^{(2)}_r = \ln \Xi_r + \ln n_r! - \sum_k \ln n^r_k!, \tag{13}$$

and $\ln \Xi_r \simeq 2\sqrt{\zeta(2) e_r}$. For $n_r \gg 1$ we obtain $\mathcal{L}_\kappa \simeq \sum_r H(\{n^r_k/n_r\})$. This approximation was used a priori in Ref. [7], but Eq. 11 is a more complete description length of the non-overlapping degree sequence, and its use should be preferred. Hence, like the description length of the overlapping partition, the encoding above offers no overhead when the partition is non-overlapping.

### B. Matrix of edge counts

The final piece that needs to be described is the matrix of edge counts $\{e_{rs}\}$. We may view this set as an adjacency matrix of a multigraph with $B$ nodes and $E = \sum_{rs} e_{rs}/2$ edges. The total number of such matrices is $\left(\!\!\binom{\binom{B}{2}}{E}\!\!\right)$, and the logarithm of this number can be used as the description length [40]. However, this implicitly assumes that all matrices are equally likely a priori. Not only this is unlikely to be the case, since most networks still possess structure at the block level, but this assumption also leads to a limit in the detection of small groups, with a maximum detectable number of groups scaling as $B_{\max} \sim \sqrt{N}$ [40]. Similarly to what we did for the node partition and the degree sequence, this can be solved by considering a generative model for the edge counts themselves, with its own set of hyperparameters. Since they correspond to a multigraph, a natural choice is the stochastic block model itself, which has its own set of edge counts, and so on recursively until one has only one group at the top. This nested stochastic block model was proposed in Ref. [7], where it has been shown to reduce the resolution limit to $B_{\max} \sim N/\log N$, making it often virtually non-existent in practice. Furthermore, since the number of levels and the topology at each level is obtained by minimizing the overall description length, it corresponds to a fully non-parametric way of inferring the multilevel structure of networks. If we denote the observed network to be at the level $l = 0$ of the hierarchy, then the total description length is

$$\Sigma = \mathcal{S}_{t/c} + \sum_{l=1}^{L} S_m(\{e^l_{rs}\}, \{n^l_r\}) + \mathcal{L}^{l-1}_t, \tag{14}$$

with $\{e^l_{rs}\}$, $\{n^l_r\}$ describing the block model at level $l$, and

$$\mathcal{S}_m = \sum_{r>s} \ln \left(\!\!\binom{n_r n_s}{e_{rs}}\!\!\right) + \sum_r \ln \left(\!\!\binom{\left(\!\binom{n_r}{2}\!\right)}{e_{rr}/2}\!\!\right) \tag{15}$$

is the entropy of the corresponding multigraph ensemble and

$$\mathcal{L}^l_t = \ln \left(\!\!\binom{B_l}{B_{l-1}}\!\!\right) + \ln B_{l-1}! - \sum_r \ln n^l_r!. \tag{16}$$

is the description length of the node partition at level $l > 0$. For the level $l = 0$ we have $\mathcal{L}^0_t = \mathcal{L}_p$ given by Eq. 5, or $\mathcal{L}^0_t = \mathcal{L}_p + \mathcal{L}_\kappa$ for the degree-corrected model.

Note that here we use the single-membership non-degree-corrected model at the upper layers. This could be modified to include arbitrary mixtures of degree correction and multiple membership, but we stick with this formulation for simplicity.

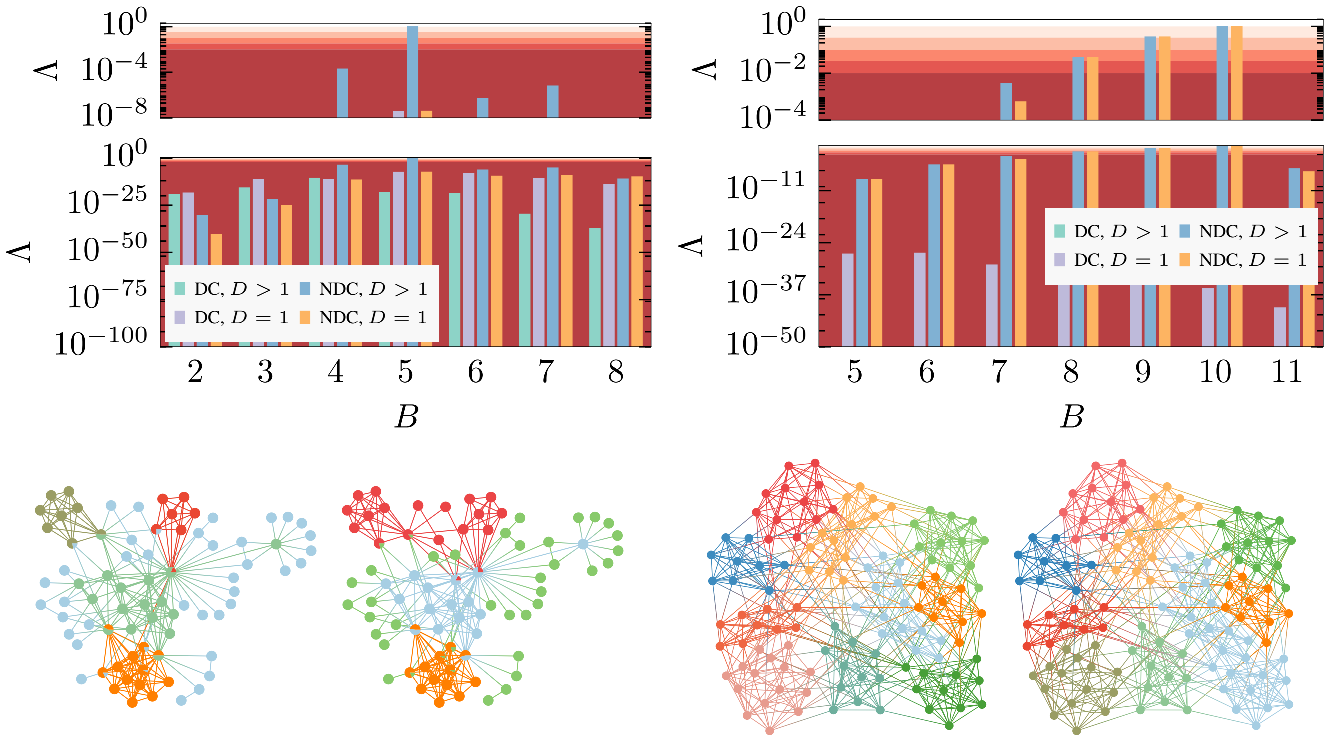
### C. Significance levels

Minimizing the description length will select the model which is most favored given the evidence in the data. But there will be situations where more than one model describes the data almost equally well, and one would like to be able to rule out alternative models with some degree of confidence. This can be done by computing the posterior probability of observing a given partition according to some version $\mathcal{H}$ of the model (e.g. degree-corrected or non-degree-corrected),

$$P(\{\vec{b}_i\}, \mathcal{H}|G) = \frac{P(G|\{\vec{b}_i\}, \mathcal{H}) P(\{\vec{b}_i\}|\mathcal{H}) P(\mathcal{H})}{P(G)}, \tag{17}$$

where $P(\mathcal{H})$ is the prior probability associated with model type $\mathcal{H}$, and $P(G) = \sum_{\{\vec{b}_i\},\mathcal{H}} P(G|\{\vec{b}_i\}, \mathcal{H}) P(\{\vec{b}_i\}|\mathcal{H}) P(\mathcal{H})$ is a normalization constant. The marginal likelihood $P(G|\{\vec{b}_i\}, \mathcal{H})$ is obtained by summing over the remaining model parameters. In the case of the degree corrected model ($\mathcal{H} = \text{DC}$) they are the $\{e_{rs}\}$ matrix and the labelled degree sequence $\{\vec{k}_i\}$ (which is omitted for the non-degree-corrected model, $\mathcal{H} = \text{NDC}$),

$$\begin{aligned} P(G|\{\vec{b}_i\}, \text{DC}) &= \sum_{\substack{\{e'_{rs}\}\\ \{\vec{k}'_i\}}} P(G|\{\vec{b}_i\}, \{e'_{rs}\}, \{\vec{k}'_i\}) P(\{e'_{rs}\}) P(\{\vec{k}'_i\}) \\ &= P(G|\{\vec{b}_i\}, \{e_{rs}\}, \{\vec{k}_i\}) P(\{e_{rs}\}) P(\{\vec{k}_i\}), \end{aligned} \tag{18}$$



$B=5$,
non-degree-corrected,
overlapping, $\Lambda=1$

$B=4$,
non-degree-corrected,
overlapping, $\Lambda\simeq 2\times 10^{-4}$,

$B=10$,
non-degree-corrected,
non-overlapping, $\Lambda=1$

$B=9$,
non-degree-corrected,
non-overlapping, $\Lambda\simeq 0.36$

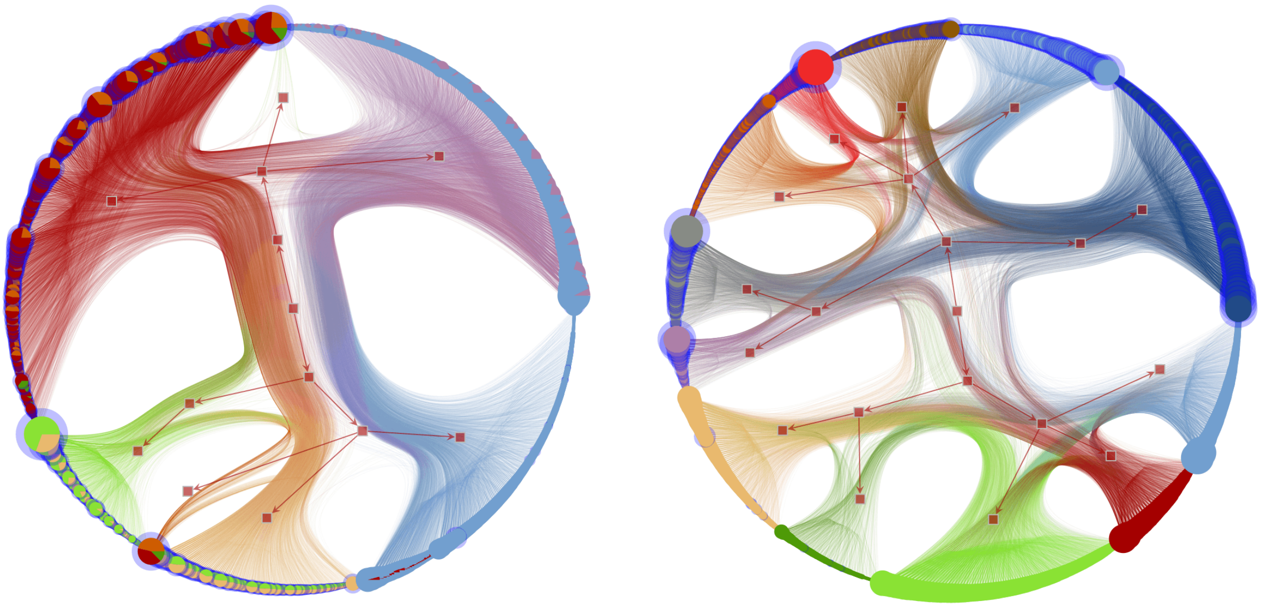
Figure 1. Left: Values for posterior odds ratio $\Lambda$ for the network of co-appearances of characters in the novel "Les Misérables", for all model variations ($D>1$ indicates an overlapping model, "DC" a degree-corrected model and "NDC" a non-degree-corrected one). The models with the best and second-best fits are shown at the bottom. Right: Same as in the left, but for the American college football network.

where the sum contains trivially only one term, since for the same graph $G$ and partition $\{\vec{b}_i\}$, there is only one possible choice for the $\{e_{rs}\}$ matrix and degree sequence $\{\vec{k}_i\}$, which is a convenient feature of the microcanonical model formulation considered here [the same holds for $\mathcal{H} = \text{NDC}$, i.e. $P(G|\{\vec{b}_i\},\text{NDC}) = P(G|\{\vec{b}_i\},\{e_{rs}\})P(\{e_{rs}\})$]. Now if we want to compare two competing partitions $\{\vec{b}_i\}_a$ and $\{\vec{b}_i\}_b$, we must compute the posterior odds ratio $\Lambda$,

$$\begin{aligned}
\Lambda &= \frac{P(\{\vec{b}_i\}_a,\mathcal{H}_a|G)}{P(\{\vec{b}_i\}_b,\mathcal{H}_b|G)} && (19)\\
&= \frac{P(G|\{\vec{b}_i\}_a,\mathcal{H}_a)P(\{\vec{b}_i\}_a|\mathcal{H}_a)P(\mathcal{H}_a)}{P(G|\{\vec{b}_i\}_b,\mathcal{H}_b)P(\{\vec{b}_i\}_b|\mathcal{H}_b)P(\mathcal{H}_b)} && (20)\\
&= \exp\left(-\Delta\Sigma\right), && (21)
\end{aligned}$$

with $\Delta\Sigma = \Sigma_a - \Sigma_b$ being the difference in the description length, and in Eq. 21 it was assumed that $P(\mathcal{H}_a) = P(\mathcal{H}_b) = 1/2$, corresponding to a lack of a priori preference for either model variant (which in fact makes $\Lambda$ identical to the Bayes factor [41]). For a value of $\Lambda = 1$, both models explain the data equally well. For values of $\Lambda < 1$ model $a$ is rejected in favor of $b$ with a confidence increasing as $\Lambda$ diminishes. In order to simplify its interpretation, the values of $\Lambda$ are usually divided into regions corresponding to a subjective assessment of the evidence strength. A common classification is as follows [41]: Values of $\Lambda$ in the intervals $\{[1,1/3],[1/3,1/10],[1/10,1/30],[1/30,1/100],[1/100,0]\}$ are considered to be very weak evidence supporting model $a$, substantial evidence, strong evidence, very strong evidence and decisive evidence, respectively.

Using the posterior odds ratio $\Lambda$ is more practical than some alternative model selection approaches, such as likelihood ratios. As has been recently shown [22], the likelihood distribution for the stochastic block model does not follow a $\chi^2$-distribution asymptotically for sparse networks, and hence the calculation of a $p$-value must be done via an empirical computation of the likelihood distribution which is computationally costly, and prohibitively so for very large networks. In contrast, computing $\Lambda$ can be done easily, and it properly accounts for the increased complexity of models with larger parameters, and protects against overfitting.



$B = 7$, overlapping, degree-corrected, $\Lambda = 1$

$B = 12$, non-overlapping, degree-corrected, $\log_{10}\Lambda \simeq -747$

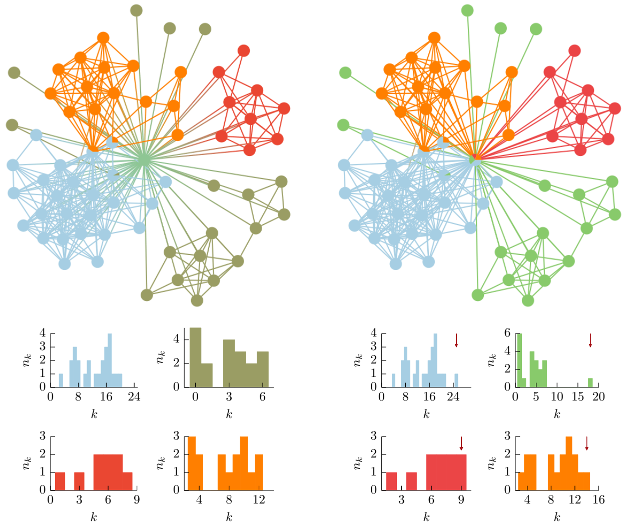
Figure 2. The network of political blogs by Adamic et al [42]. The left panel shows the best model with an overlapping partition, and the right the best non-overlapping one. Nodes with a blue halo belong to the Republican faction, as determined in Ref. [42]. For the visualization, the hierarchical edge bundles algorithm [43] was used.

## IV. EMPIRICAL NETWORKS

The method outlined in the previous section allows one to determine the best model from the various available choices. Here we analyze some empirical examples, and determine the most appropriate model, and examine the consequences of the balance struck between model complexity and quality of fit. We start with two small networks, the co-appearance of characters in the Victor Hugo novel "Les Misérables" [45], and a network of college American football games [46, 47]. For both networks, we obtain the best partition according all model variations and for different number of groups $B$, and compute the value of $\Lambda$ relative to the best model, as shown in Fig. 1. For the "Les Misérables" network, the best fit is a non-degree-corrected overlapping model which puts the most central characters in more than one group. All other partitions for different values of $B$ and model types result in values significantly below the plausibility line of $\Lambda = 10^{-2}$, indicating that the overlapping model offers a better explanation for the data with a large degree of confidence. In particular, it offers a better description than the non-overlapping model with degree correction. For the Football network, on the other hand, the proffered model is non-overlapping and without degree correction with $B = 10$, which matches very well the assumed correct partition into 10 conferences. The groups are relatively homogeneous, with most nodes having similar degrees, such that degree correction becomes an extra burden, with very little explanatory power. For this network, however, there are alternative fits with values of $\Lambda$ within the plausibility region, which means that the communities are not very strongly defined, and they admit alternative partitions with $B = 9$ and $B = 8$ groups which cannot be fully discarded given the evidence in the data.

Degree correction tends to become a better choice for larger data sets, which display stronger degree variability. One example of this is the network of political blogs obtained by Adamic et al [42]. For this network, the best model is a degree-corrected, overlapping partition into $B = 7$ groups, shown in Fig. 2. Compared to this partition, the best alternative model without overlap divides the network into $B = 12$ groups[1], but has a posterior odds ratio significantly below the plausibility region. It should be observed that the non-overlapping version captures well the segregation into two groups (Republicans and Democrats) at the topmost level of the hierarchy. The overlapping version, on the other hand, tends to classify half-edges belonging to different camps into different groups, which is compatible with the accepted division, but the upper layers of the hierarchy do not reflect this, and prefers to merge together groups that belong to different factions, but that have similar roles.

[1] In Ref. [7] using the same non-overlapping model, a value of $B = 15$ was found. This is due the difference in the description length for the degree sequence, where here we use a more complete estimation than in Ref. [7], which results in this slight difference.



$B = 5$, overlapping, non-degree-corrected, $\Lambda = 1$

$B = 4$, overlapping, non-degree-corrected, $\Lambda \simeq 0.053$

Figure 3. Ego network of Facebook contacts [44]. Left: The best model fit across all model variations, which puts the ego node in its own group. Right: The alternative hypothesis where the node is split in several groups. Below each network are shown the degree distributions inside each group. The arrow marks the degree of the ego node.

Overlapping partitions, however, do not always provide better descriptions, even in situations where it might be considered more intuitive. One of the contexts where overlapping communities are often considered to be better explanations are in social networks, where different social circles could be represented as different groups (e.g. family, co-workers, friends, etc.), and one could belong to more than one of these groups. This is well illustrated by so-called "ego networks", where one examines only the immediate neighbours of a node, and their mutual connections. One such network, extracted from the Facebook online social network [44], is shown in Fig. 3. The common interpretation of networks such as these is shown on the right in Fig. 3, and corresponds to a partition of the central "ego" node so that it belongs to all of the different circles. Under this description, the ego node is only special in the sense that it belongs to all groups, but inside each group it is just a common member. However, among all model variants, the best fit turns out to be the one where the ego node is put separately in its own group, as shown in the left in Fig. 3. In this example it is easy to see why this is the case. If we observe the degree distribution inside each group for the network on the left, we see that there is no strong degree variation. On the right, as the ego is included in each group, it becomes systematically the most connected node. This is simply by construction, since the ego must connect to every other node. The only situation where the ego would not stand out inside each group, would be if the communities were cliques. Hence, since the ego is not a typical member of any group, it is simpler to classify it separately in its own group, which is selected by the method as a being a more plausible hypothesis. Note that degree correction is not selected as the most plausible solution, since it is burdened with the individual description of every degree in the network, which are fairly uniform with the exception of the ego. One can imagine a different situation where there would be other very well connected nodes inside each group, so that the ego could be described as a common member of each group, but this not observed in any other network obtained in Ref. [44]. Naturally, if one considers the complete network, of which the ego neighbourhood is only a small part, the situation may change, since there may me members of each group to which the ego does not have a direct connection.

When performing model selection for larger networks, it is often the case that the overlapping models are not chosen. In table I are shown the results for many empirical networks belonging to different domains. For the majority of cases, the non-overlapping degree-corrected models are selected. The are, however, many exceptions which include two social networks (Gowalla and Brightkite [53]), the global airport network of `openflights.com`, the neuronal network of *C. elegans* [54], the political blog network already mentioned, the arXiv co-authorship networks [55] [in the fields of general relativity and quantum cosmology (gr-qc), high-energy physics (hep-th), condensed matter (cond-mat), and astronomy (astro-ph)], co-authorship in network science [56], and the network of genes implicated in diseases [58], for which some version of the overlapping model is chosen. Interestingly, for the arXiv co-authorship network in high-energy physics/phenomenology (hep-ph) a non-overlapping model is selected instead. For only one of the remaining four arXiv networks (astro-ph), the degree-corrected version of the overlapping model is selected, whereas for the other three the non-degree-corrected version is preferred. Hence, for co-authorship networks the model selection procedure seems to correspond to the intuition that they are composed predominantly of overlapping groups [25].

We take the arXiv cond-mat network as a representative example of the differences between the inferred models. As can seen in Fig. 4, although the degree distribution is very broad, the inferred labelled degree distribution is narrower, meaning that many large-degree nodes can be well explained as having a smaller degree of any single type, but belonging simultaneously to many groups (in the specific context of this network, prolific authors tend to be the ones which belong to many different types of collaborations). The distribution of mixture sizes $n_d$ has almost always a maximum at $d = 1$, meaning that most nodes belong to one group, but with a tail which is comparatively broad (this seems to be a general feature which is observed in the majority of networks analyzed). The distribution of group sizes can be very different, depending on which model is used. Non-overlapping models

| No. | $N$ | $\langle k\rangle$ | $\log_{10}\Lambda_{\text{DCO}}$ | $\log_{10}\Lambda_{\text{DC}}$ | $\log_{10}\Lambda_{\text{NDCO}}$ | $\log_{10}\Lambda_{\text{NDC}}$ | $B$ | $\langle d\rangle$ | $\Sigma/E$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 34 | 4.6 | $-2.1$ | $-2.1$ | — | 0 | 2 | 1 | 4 |
| 2 | 62 | 5.1 | $-4.6$ | $-1.4$ | — | 0 | 2 | 1 | 4.8 |
| 3 | 77 | 6.6 | $-17$ | $-7.7$ | 0 | $-7.3$ | 5 | 1.1 | 4 |
| 4 | 105 | 8.4 | $-12$ | $-2.8$ | $-6.6$ | 0 | 5 | 1 | 4.4 |
| 5 | 115 | 10.7 | $-79$ | $-27$ | — | 0 | 10 | 1 | 4.3 |
| 6 | 297 | 15.9 | 0 | $-61$ | $-2.0\times10^2$ | $-2.1\times10^2$ | 5 | 1.8 | 5.1 |
| 7 | 379 | 4.8 | $-47$ | $-6.6$ | 0 | $-8.9$ | 20 | 1.1 | 6.2 |
| 8 | 903 | 15.0 | $-3.8\times10^2$ | $-3.7\times10^2$ | 0 | $-3.7\times10^2$ | 60 | 1.2 | 3.1 |
| 9 | 1, 278 | 2.8 | $-8.1$ | 0 | $-1.5\times10^2$ | $-89$ | 2 | 1 | 7.4 |
| 10 | 1, 490 | 25.6 | 0 | $-5.2\times10^2$ | $-2.3\times10^3$ | $-2.3\times10^3$ | 7 | 1.8 | 4.4 |
| 11 | 1, 536 | 3.8 | $-2.5\times10^2$ | 0 | $-65$ | $-62$ | 38 | 1 | 6.7 |
| 12 | 1, 622 | 11.2 | $-4.3\times10^2$ | 0 | $-12$ | $-82$ | 48 | 1 | 3.3 |
| 13 | 1, 756 | 4.5 | $-43$ | 0 | $-4.0\times10^2$ | $-2.8\times10^2$ | 7 | 1 | 5.9 |
| 14 | 2, 018 | 2.9 | $-9.2$ | 0 | $-2.9\times10^2$ | $-2.1\times10^2$ | 2 | 1 | 8.5 |
| 15 | 4, 039 | 43.7 | $-1.5\times10^3$ | 0 | $-8.1\times10^2$ | $-9.5\times10^2$ | 158 | 1 | 3.2 |
| 16 | 4, 941 | 2.7 | $-2.2\times10^2$ | 0 | $-21$ | $-25$ | 25 | 1 | 11 |
| 17 | 7, 663 | 17.8 | 0 | $-1.1\times10^4$ | $-5.3\times10^3$ | $-1.6\times10^4$ | 85 | 1 | 3.2 |
| 18 | 7, 663 | 5.3 | $-1.8\times10^3$ | 0 | $-9.3\times10^2$ | $-7.3\times10^2$ | 63 | 1 | 5 |
| 19 | 8, 298 | 25.0 | $-9.1\times10^3$ | 0 | $-1.4\times10^4$ | $-1.4\times10^4$ | 34 | 1 | 5.4 |
| 20 | 9, 617 | 7.7 | $-4.2\times10^3$ | 0 | $-2.3\times10^3$ | $-2.5\times10^3$ | 34 | 1 | 9.3 |
| 21 | 26, 197 | 2.2 | $-2.4\times10^3$ | $-1.2\times10^3$ | 0 | $-2.7\times10^3$ | 363 | 1.3 | 4.5 |
| 22 | 36, 692 | 20.0 | $-4.1\times10^4$ | 0 | $-8.5\times10^4$ | $-2.8\times10^4$ | 1812 | 1 | 5.5 |
| 23 | 39, 796 | 15.2 | $-6.1\times10^4$ | 0 | $-8.8\times10^4$ | $-4.5\times10^4$ | 1323 | 1 | 6.3 |
| 24 | 52, 104 | 15.3 | $-1.5\times10^5$ | 0 | $-3.7\times10^4$ | $-4.0\times10^4$ | 172 | 1 | 6.4 |
| 25 | 58, 228 | 14.7 | 0 | $-5.8\times10^4$ | $-1.8\times10^5$ | $-1.4\times10^5$ | 1995 | 3.2 | 7.3 |
| 26 | 65, 888 | 305.2 | $-4.4\times10^4$ | 0 | $-4.6\times10^5$ | $-4.6\times10^5$ | 384 | 1 | 4.1 |
| 27 | 68, 746 | 1.5 | $-4.8\times10^3$ | $-1.4\times10^3$ | 0 | $-7.0\times10^3$ | 719 | 1.4 | 6.4 |
| 28 | 75, 888 | 13.4 | $-1.1\times10^5$ | 0 | $-8.2\times10^4$ | $-9.0\times10^4$ | 143 | 1 | 8.9 |
| 29 | 89, 209 | 5.3 | $-1.0\times10^4$ | 0 | $-9.7\times10^3$ | $-1.1\times10^4$ | 848 | 1 | 3.2 |
| 30 | 108, 300 | 3.5 | $-3.3\times10^3$ | $-5.2\times10^3$ | 0 | $-2.4\times10^4$ | 1660 | 1.8 | 5.7 |
| 31 | 133, 280 | 5.9 | 0 | $-4.4\times10^3$ | $-7.4\times10^4$ | $-3.8\times10^4$ | 1944 | 5.3 | 4.4 |
| 32 | 196, 591 | 19.3 | 0 | $-1.9\times10^5$ | $-7.1\times10^5$ | $-6.6\times10^5$ | 6856 | 3.7 | 7.8 |
| 33 | 265, 214 | 3.2 | $-1.4\times10^4$ | 0 | $-9.2\times10^4$ | $-8.5\times10^4$ | 549 | 1 | 8.6 |
| 34 | 273, 957 | 16.8 | $-5.4\times10^5$ | 0 | $-4.6\times10^4$ | $-7.2\times10^4$ | 727 | 1 | 5.8 |
| 35 | 281, 904 | 16.4 | $-1.2\times10^6$ | 0 | $-2.8\times10^5$ | $-1.5\times10^5$ | 6655 | 1 | 4.3 |
| 36 | 317, 080 | 6.6 | $-1.7\times10^5$ | 0 | $-3.9\times10^5$ | $-4.2\times10^5$ | 8766 | 1 | 11 |
| 37 | 325, 729 | 9.2 | $-5.8\times10^5$ | 0 | $-1.1\times10^6$ | $-2.3\times10^5$ | 4293 | 1 | 5.8 |
| 38 | 325, 729 | 9.2 | $-5.6\times10^5$ | 0 | $-1.2\times10^6$ | $-2.5\times10^5$ | 3995 | 1 | 5.8 |
| 39 | 334, 863 | 5.5 | $-3.3\times10^5$ | 0 | $-3.6\times10^5$ | $-3.4\times10^4$ | 9118 | 1 | 11 |
| 40 | 372, 787 | 9.7 | $-1.0\times10^6$ | 0 | $-1.3\times10^5$ | $-1.4\times10^5$ | 965 | 1 | 11 |
| 41 | 463, 347 | 20.3 | $-6.4\times10^5$ | 0 | $-1.8\times10^6$ | $-1.5\times10^6$ | 9276 | 1 | 9.3 |
| 42 | 1, 134, 890 | 5.3 | — | 0 | $-4.5\times10^5$ | $-4.9\times10^5$ | 264 | 1 | 13 |

| | | | |
|---|---|---|---|
| 1 | Karate Club [48] | 22 | Enron emails [49, 50] |
| 2 | Dolphins [51] | 23 | PGP [52] (directed) |
| 3 | Les Misérables [45] | 24 | Internet AS (Caida)[a] (directed) |
| 4 | Political Books[b] | 25 | Brightkite social network [53] |
| 5 | American Football [46, 47] | 26 | netflix-pruned-smaller-u |
| 6 | *C. elegans* Neurons [54] (directed) | 27 | arXiv Co-Authors (hep-th) [55] |
| 7 | Coauthorships in network science [56] | 28 | Epinions.com trust network [57] (directed) |
| 8 | Disease Genes [58] | 29 | arXiv Co-Authors (hep-ph) [55] |
| 9 | Yeast protein interactions (CCSB-YI11) [59] | 30 | arXiv Co-Authors (cond-mat) [55] |
| 10 | Political Blogs [42] (directed) | 31 | arXiv Co-Authors (astro-ph) [55] |
| 11 | Yeast protein interactions (LC) [60] | 32 | Gowalla social network [53] |
| 12 | Yeast protein interactions (Combined AP/MS) [61] | 33 | EU email [55] (directed) |
| 13 | *E. coli* gene regulation [62] (directed) | 34 | Flickr [63] |
| 14 | Yeast protein interactions (Y2H union) [59] | 35 | Web graph of `stanford.edu`. [64] (directed) |
| 15 | Facebook egos [44] | 36 | DBLP collaboration [65] |
| 16 | Power Grid [54] | 37 | Web graph of `nd.edu`. [64] (directed) |
| 17 | Airport routes [c] (directed) | 38 | WWW [66] (directed) |
| 18 | Airport routes | 39 | Amazon product network [65] |
| 19 | Wikipedia Votes [67, 68] (directed) | 40 | IMDB film-actor[d] [40] |
| 20 | Human protein interactions (HPRD r9) [69] | 41 | APS citations[e] (directed) |
| 21 | arXiv Co-Authors (gr-qc) [55] | 42 | Youtube social network [65] |

[a] Retrieved from `http://www.caida.org`.

[b] V. Krebs, retrieved from `http://www-personal.umich.edu/~mejn/netdata/`

[c] Retrieved from `http://openflights.org/`

[d] Retrieved from `http://www.imdb.com/interfaces`.

[e] Retrieved from `http://publish.aps.org/dataset`.

Table I. Comparison of different models for many empirical networks. The columns at the top table correspond to the data set number (with the name given at the bottom table), the number of nodes $N$, the average degree $\langle k\rangle = 2E/N$, the posterior odds ratios relative to the best model for the degree-corrected overlapping ($\Lambda_{\text{DCO}}$), the degree-corrected non-overlapping ($\Lambda_{\text{DC}}$), non-degree-corrected overlapping ($\Lambda_{\text{NDCO}}$) and non-degree-corrected non-overlapping ($\Lambda_{\text{NDC}}$) models. Missing entries correspond to situations where the best overlapping partition turns out to be non-overlapping. The last three columns show some parameters of the best model: The number of groups $B$, the average mixture size $\langle d\rangle$, and the description length per edge (in bits per edge).

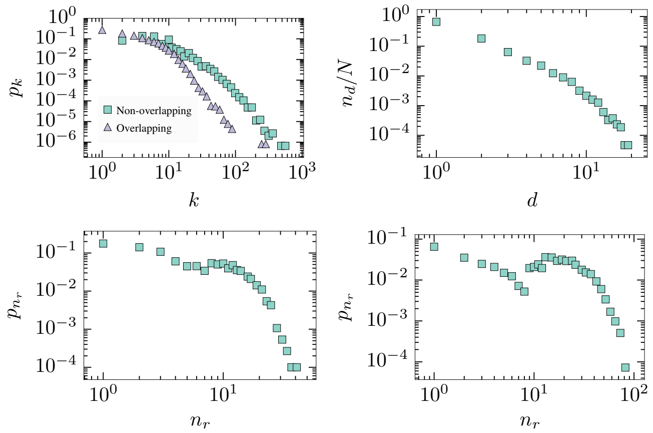


Figure 4. Statistical properties of the best model inferred for the network of arXiv co-authors in the field of condensed matter (cond-mat). Top left: Degree distribution of the original network and of the overlapping model (where the labeled degree sequence $\{\vec{k}_i\}$ is flattened into a single histogram for all labelled degrees $\{k_i^r\}$). Top right: Distribution of mixture sizes, $n_d$. Bottom left: Distribution of group sizes for the best-fitting *non-overlapping*, non-degree-corrected model. Bottom right: Distribution of group sizes for the best-fitting *overlapping*, non-degree-corrected model.

without degree correction tend to find groups which are strongly correlated with degrees [12], and hence lead to a broad distribution of group sizes when the degree distribution is also broad. On the other hand, either degree correction or group overlap tend to change the distribution considerably. In the literature there are often claims of community sizes following power-law distributions [70–73] with figures similar to the lower left panel of Fig. 4. Regardless to the validity of this hypothesis for the various methods used in the literature, this is certainly not the case for the overlapping model as shown in the lower right panel of the same figure. Indeed, for most networks analyzed, the model which best fits the data (which tends to be degree-corrected and non-overlapping) shows no vestige of group sizes following a scale-free distribution. Some further examples of this are shown in Fig. 5, where characteristic size scales can be clearly identified.

## V. MODEL IDENTIFIABILITY: OVERLAPPING VS. NON-OVERLAPPING

A central issue when selecting between non-overlapping and overlapping models is to decide when a group of nodes should belong simultaneously to two or more groups, of if these nodes should be better represented by a single membership to a different unique group. The choice is not always immediately obvious, since we can always generate very similar networks with either model. If we generate a network with the overlapping model, but treat it as if it were generated by the non-overlapping model, with each distinct mixture $\vec{b}$

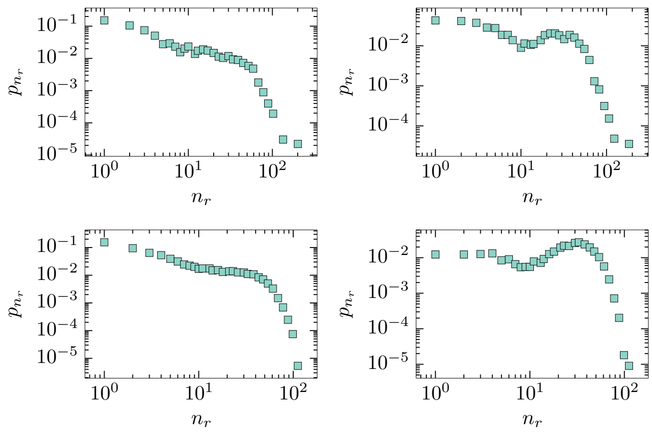


Figure 5. Distribution of group sizes for the best fitting non-degree-corrected non-overlapping model (left) and the degree-corrected non-overlapping model (right), for the PGP [52] (top) and DBLP collaboration [65] (bottom) networks. In both cases the degree-corrected model provides a better fit, as shown in table I.

corresponding to a separate non-overlapping group, the associated entropy will be

$$\mathcal{S}'_t \simeq E - \frac{1}{2}\sum_{\vec{b}_1\vec{b}_2} e_{\vec{b}_1\vec{b}_2}\ln\left(\frac{e_{\vec{b}_1\vec{b}_2}}{n_{\vec{b}_1}n_{\vec{b}_2}}\right), \tag{22}$$

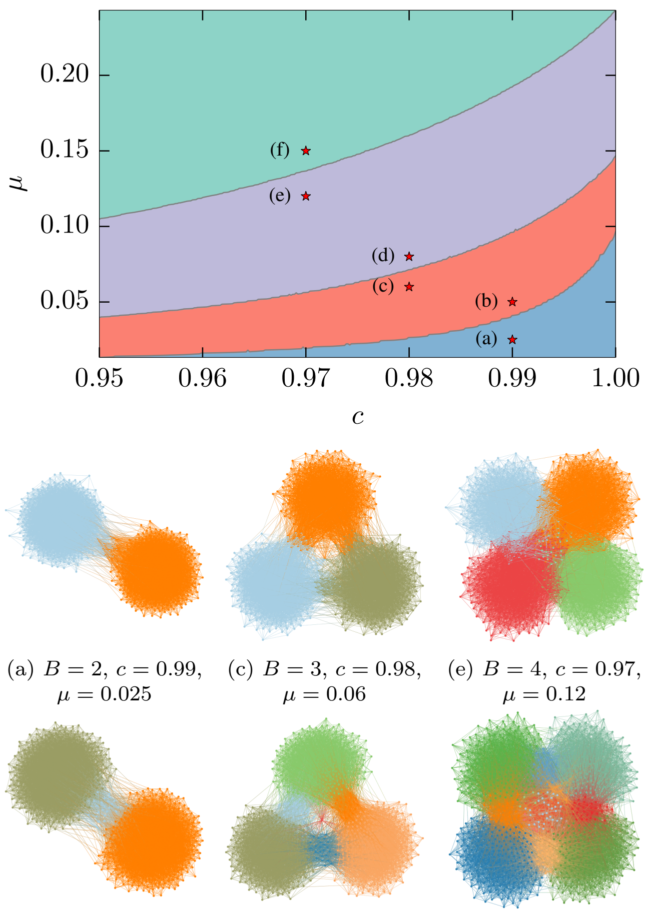
where

$$e_{\vec{b}_1\vec{b}_2} = \sum_{rs} b_1^r b_2^s \frac{e_{rs}}{n_r n_s} n_{\vec{b}_1} n_{\vec{b}_2} \tag{23}$$

is the expected number of edges between mixtures $\vec{b}_1$ and $\vec{b}_2$. By exchanging the sums and using Jensen's inequality we observe directly that

$$\mathcal{S}'_t \leq E - \frac{1}{2}\sum_{rs} e_{rs}\ln\left(\frac{e_{rs}}{n_r n_s}\right), \tag{24}$$

with the right-hand side being the entropy of original overlapping model $\mathcal{S}_t$, and with the equality holding only if the original model happens to be non-overlapping to begin with. Thus, the non-overlapping model will invariably possess a lower entropy. Nevertheless, the overlapping hypothesis may still be preferred if the number of groups $B$ is sufficiently smaller than the number of individual $\vec{b}$ mixtures, so that the total description length is shorter. It should be observed, however, that since one model is contained inside the other, the difference in the description length can be interpreted simply as the difference in the prior probabilities for the model parameters. As the amount of available data increases, the effect of the priors should "wash out", and the description length should be increasingly dominated by the model entropy alone. In these cases one should expect the non-overlapping model to be preferred, regardless of the specific model which was used to generate the data. However, differently than models which generate independent data points, the "amount of available data" for network models is a finer issue. In the case of the stochastic block model it involves the simultaneous scaling of the number of edges $E$, the number of nodes $N$ and the number of groups $B$.



(a) $B = 2$, $c = 0.99$, $\mu = 0.025$

(c) $B = 3$, $c = 0.98$, $\mu = 0.06$

(e) $B = 4$, $c = 0.97$, $\mu = 0.12$

(b) $B = 3$, $c = 0.99$, $\mu = 0.05$

(d) $B = 7$, $c = 0.98$, $\mu = 0.08$

(f) $B = 15$, $c = 0.97$, $\mu = 0.15$

Figure 6. Top: Parameter regions for the model considered in the main text, with $N = 10^3$ and $\langle k\rangle = 2E/N = 20$. Each curve corresponds to one value of $B$, and separate a region above where the non-overlapping model is preferred from a region below where the overlapping model is chosen. Bottom: Networks and their preferred partitions, corresponding to parameter values indicated in the top panel.

As a case example, here we consider a simple overlapping assortative model, with $e_{rs} = 2E[\delta_{rs}c/B + (1-\delta_{rs})(1-c)/B(B-1)]$, with $c \in [0,1]$ controlling the degree of assortativity. The mixtures are parameterized as $n_{\vec{b}} = C\prod_r \mu^{b_r}$, with $C$ being a normalization constant, and $\mu \in ]0,1]$ controlling the degree of overlap. For $\mu \to 0$ we obtain asymptotically a non-overlapping partition with $n_r = N/B$, and for $\mu = 1$ all mixtures $\vec{b}$ have the same size. We compare the difference in description length between this model and its equivalent parametrization with each mixture as a separate group. As can be seen in Fig. 6, for any given value of $c$, there is a value of $\mu$ above which the non-overlapping model is preferred. In this parameter region, the group intersections are sufficiently well populated with nodes, so that their representation as individual groups is chosen. For values of $\mu$ below this value, the intersections are significantly smaller than the non-overlapping portion. In this case, the data is better explained as larger groups of almost non-overlapping nodes, with few nodes at the intersections. The boundary separating the two regions recedes upwards as the number of groups $B$ is increased, meaning that a larger number of distinct intersections can compensate for a smaller number of non-overlapping nodes. It should also be pointed out that the boundaries move downwards as the number of nodes and edges is increased, such that the average degree in the network remains the same (not shown), so that it is not only the relative sizes of the intersections which are the relevant properties, but also their absolute sizes. The same occurs if the average degree increases and everything else remains constant. Hence in the limit of sufficient data, either with the number of nodes inside each group and intersection becoming sufficiently large, or each part becoming sufficiently dense, the non-overlapping model is the one which will be selected. For empirical networks, this may not be the scaling condition which is more representative, since the most appropriate number of groups and degree of overlap may in fact follow any arbitrary scaling, and hence the overlapping model may still be selected, even for very large or very dense networks. Nevertheless, this example seems to suggest that the non-overlapping model is general enough to accommodate structures generated by the overlapping model in these limiting cases, and may serve as a partial explanation to why the overlapping model is seldom selected in the empirical systems analyzed in Sec. IV.

## VI. INFERENCE ALGORITHM

The inference procedure consists in finding the labeling of the half-edges of the graph such that the description length is minimized. Such global optimization problems are often NP-hard, and require heuristics to be solvable in practical time. One possibility is to use Markov chain Monte Carlo (MCMC), which consists in modifying the block membership of each half-edge in a random fashion, and accept or reject each move with a probability given as a function of the description length difference $\Delta\Sigma$. By choosing the acceptance probabilities in the appropriate manner, i.e. by enforcing ergodicity and detailed balance, one can guarantee that the labellings will be sampled with the correct probability after a sufficiently long equilibration time is reached. However, naive formulations of the Markov chain will lead to very long equilibration times, which become unpractical for large networks. Here we adapt the algorithm developed in Ref. [74] for the non-overlapping case which implements a fast Markov chain. It consists in the move proposal of each half-edge incident on node $i$ of type $r$ to type $s$ with a probability given by

$$p(r \to s|t) = \frac{e_{ts} + \epsilon}{e_t + \epsilon B}, \tag{25}$$

where $t$ is the block label the half-edge opposing a randomly chosen half-edge incident to the same node as the half-edge being moved, and $\epsilon \geq 0$ is a free parameter. Eq. 25 means that we attempt to guess the label of a given half-edge by inspecting the group membership the neighbors of the node to which it belongs, and using the currently inferred model parameters to choose the most likely group to which it should be moved. It should be emphasized that this move proposal does not result in a preference for either assortative or dissortative networks, since it depends only on the matrix $\{e_{rs}\}$ currently inferred. For any choice of $\epsilon > 0$, this move proposal preserves ergodicity, but not detailed balance. This last characteristic can be enforced via the Metropolis-Hastings criterion [75, 76] by accepting each move with a probability $a$ given by

$$a = \min\left\{e^{-\beta\Delta\Sigma}\frac{\sum_t p_t^i p(s \to r|t)}{\sum_t p_t^i p(r \to s|t)}, 1\right\}, \tag{26}$$

where $p_t^i$ is the fraction of opposing half-edges of node $i$ which belong to block $t$, and $p(s \to r|t)$ is computed after the proposed $r \to s$ move (i.e. with the new values of $e_{tr}$), whereas $p(r \to s|t)$ is computed before. The parameter $\beta$ in Eq. 26 is an inverse temperature, which can be used to sample partitions according to their description length ($\beta = 1$) or to find the ground state ($\beta \to \infty$). As explained in Ref. [74], this move proposal as well as the computation of $a$ can be done efficiently, with minimal book-keeping, so that a sweep of the network (where each half-edge move is attempted once) is done in time $O(E)$, independent of the number of groups $B$. This is true even in the overlapping case, since updating Eqs. 1, 2, 5 and 10 after each half-edge move can be done in time $O(1)$.

As discussed in Ref. [74], although the MCMC method above succeeds in equilibrating faster than a naive Markov chain, it still suffers from a strong dependence on how close one starts from the global minimum. Usually, starting from a random partition of the half-edges leads to metastable states where the Markov chain seems to have equilibrated, but in fact the network structure has only been partially discovered, and will move from such configurations only after a very long time. This is a problem common to many inference procedures based on local moves such as expectation maximization [14] and belief propagation [77]. In Ref. [74] a multilevel agglomerative heuristic was proposed, which significantly alleviates this problem. It consists in equilibrating the chain for a larger number of groups, and then merging the groups using the same algorithm used for the block membership moves. This method, however, cannot be used unmodified in the overlapping case, since the strict merging of groups will not properly explore the landscape of possible overlap-

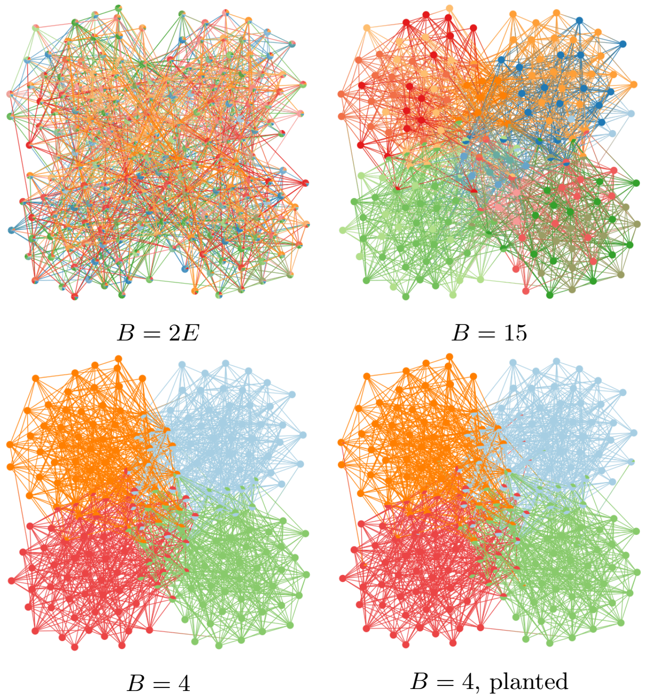


Figure 7. Typical outcome of the greedy multilevel agglomerative algorithm described in the text, for a network sampled from the overlapping model with $B = 4$. The different panels show the progression of the algorithm from $B = 2E$ to $B = 4$. The panel on the lower right shows the planted partition used to generate the network.

ping partitions. We therefore modify the approach as follows. Before groups are merged, the half-edges belonging to each one of them are split into subgroups corresponding to the different group memberships at the opposing sides. These subgroups are then treated as separate groups, and are merged together until the desired number of groups is achieved. All the details of the algorithm beyond this modification are performed exactly as described in Ref. [74]. Since this algorithm usually does a good job in finding the a partition very close to the final one, it also tends to perform very well when the algorithm is turned into a greedy heuristic, by starting with $B = 2E$ and each half-edge in its own group, and by making $\beta \to \infty$. An example of a typical outcome of the greedy algorithm is show in Fig. 7. The greedy version is very fast, with an overall complexity of $O(E \ln^2 E)$, which makes it usable for very large networks. Note that this complexity is independent on the number of groups, $B$. This is a strong contrast to other methods proposed for the same problem, such as the stochastic optimization algorithm of Gopalan et al [78], and the expectation maximization algorithm of Ball et al [14], both of which have a complexity of $O(EB)$ per sweep, although they only consider strictly assortative models, and applying the same techniques to the more general models considered here would lead to a $O(EB^2)$ complexity, similar to belief propagation algorithms for non-overlapping models [22, 77]. Although these approaches can be very efficient if the number of groups is very small, they quickly become prohibitive if the most appropriate number of groups scales as some function of the system size (which seems to be generally the case when model selection is applied, see table I and Ref. [7]), which is not an issue with the algorithm described above. It should also be noted that none of the other algorithms mentioned [14, 22, 77, 78] are designed to overcome metastable solutions, like the multilevel approach presented here.

For most networks analyzed in this work, the fast heuristic version of the algorithm was used, together with the algorithm described in Ref. [7] to infer the upper layers of the hierarchy (which includes the determination of the number of groups $B$ at the lowest level, in addition to the entire hierarchy, in a non-parametric fashion)[2].

## VII. CONCLUSION

We presented a method of inferring overlapping and degree-corrected versions of the stochastic block model based on the minimum description length principle (MDL) that avoids overfitting and allows for the comparison between model classes. Based on a Bayesian interpretation of MDL, we derived a posterior odds ratio test that yields a degree of confidence with which models can be selected or discarded. In applying this method to a variety of empirical networks, we obtained that for the majority of them the non-overlapping degree-corrected model variant is the one which best fits the data.

Although overlapping structures are often considered to be more reasonable explanations for some networks, we showed that in many representative cases the non-overlapping model can accommodate the same structure while providing a more parsimonious description of the data. We expect this fact to bear on tasks which require high quality fits, such as the prediction of missing or spurious links [6, 8], or other generalizations of the data.

The models considered in this work generate unlabeled networks, without any other properties associated with the nodes or edges. However, it is often the case that either the nodes or edges have weights [18] or are of different types [17, 20], or have temporal information [21]. This sort of additional data may corroborate the evidence supporting the generation via a specific type of model (e.g. with overlaps) and tip the scale towards it. The approach presented here is generalizable to these cases as well, by augmenting the model to generate covariates associated with the edges and nodes. Furthermore, one should be able to perform a similar comparison with models which belong to very different classes, such as latent space models [80], or others.

[2] A complete implementation of the algorithm is freely available as part of the `graph-tool` library [79] at `http://graph-tool.skewed.de`.

## Appendix A: Directed graphs

The same approach of the main text can be carried over to directed graphs with no difficulties. In this case the edge counts are in general asymmetric, $e_{rs} \neq e_{sr}$, which leads to the entropy for the non-degree-corrected model [31]

$$S_t \simeq E - \sum_{rs} e_{rs} \ln\left(\frac{e_{rs}}{n_r n_s}\right). \tag{A1}$$

For the degree-corrected case, there are two degree sequences for the labelled out- and in-degrees, $\{k^{+}{}_{i}^{r}\}$ and $\{k^{-}{}_{i}^{r}\}$, respectively. Applying the same argument as for the undirected case, the entropy becomes [31]

$$S_d \simeq -E - \sum_{rs} e_{rs} \ln\left(\frac{e_{rs}}{e^{+}_{r} e^{-}_{s}}\right) - \sum_{ir} \ln k^{+}{}_{i}^{r}! - \sum_{ir} \ln k^{-}{}_{i}^{r}!, \tag{A2}$$

where $e^{+}_{r} = \sum_s e_{rs}$ and $e^{-}_{r} = \sum_s e_{sr}$.

The description length for the overlapping partition is identical to the undirected case, with $\mathcal{L}_p$ given by Eq. 5. For the labeled degree sequence, we have instead

$$\mathcal{L}_\kappa = \sum_r \ln\left(\!\!\binom{m_r}{e^{+}_{r}}\!\!\right) + \ln\left(\!\!\binom{m_r}{e^{-}_{r}}\!\!\right) + \sum_{\vec{b}} \min\left(\mathcal{L}^{(1)}_{\vec{b}}, \mathcal{L}^{(2)}_{\vec{b}}\right). \tag{A3}$$

with

$$\mathcal{L}^{(1)}_{\vec{b}} = \sum_r \ln\left(\!\!\binom{n_{\vec{b}}}{e^{+}{}^{r}_{\vec{b}}}\!\!\right) + \ln\left(\!\!\binom{n_{\vec{b}}}{e^{-}{}^{r}_{\vec{b}}}\!\!\right). \tag{A4}$$

and

$$\mathcal{L}^{(2)}_{\vec{b}} = \sum_r b_r \left(\ln \Xi^{r+}_{\vec{b}} + \ln \Xi^{r-}_{\vec{b}}\right) + \ln n_{\vec{b}}! - \sum_{\vec{k}} \ln n^{\vec{b}}_{\vec{k}^{+},\vec{k}^{-}}!, \tag{A5}$$

where $\ln \Xi^{r+}_{\vec{b}}$ and $\ln \Xi^{r-}_{\vec{b}}$ are computed as in Eq. 9 but using $e^{+}{}^{r}_{\vec{b}} = \sum_{\vec{k}^{+},\vec{k}^{-}} k^{+}_{r} n^{\vec{b}}_{\vec{k}^{+},\vec{k}^{-}}$ and $e^{r-}_{\vec{b}} = \sum_{\vec{k}^{+},\vec{k}^{-}} k^{-}_{r} n^{\vec{b}}_{\vec{k}^{+},\vec{k}^{-}}$, respectively, which give the total number of out- and in-edges incident on the mixture $\vec{b}$. In the previous equations the counts $n^{\vec{b}}_{\vec{k}^{+},\vec{k}^{-}}$ refer to the joint distribution of labelled in- and out-degrees, so that each vector $\vec{k}^{+/-}$ describes the in- and out-degrees labelled according to degree membership, i.e. $\vec{k}^{+}_{i} = \{k^{+}{}_{i}^{r}\}$ and $\vec{k}^{-}_{i} = \{k^{-}{}_{i}^{r}\}$.

## Appendix B: Poisson Models

### 1. Non-degree-corrected

This approximation of the formulation with “hard” constraints of the multiple membership model discussed in the main text is closely related to a Poisson variant of the model with “soft” constraints, where each half-edge of the graph is labeled with a latent variable specifying which group memberships were responsible for its existence, and the number of edges of type $(r,s)$ between nodes $i$ and $j$, $A^{rs}_{ij}$, is independently sampled according to a Poisson distribution, so that the likelihood becomes

$$P(G|\{\vec{b}_i\},\{p_{rs}\}) = \prod_{i>j}\prod_{r\geq s} p_{rs}^{A^{rs}_{ij}} e^{-p_{rs} b^r_i b^s_j} / A^{rs}_{ij}!, \tag{B1}$$

where $p_{rs}$ is the average number of edges of type $(r,s)$ between nodes which belong to each group. The log-likelihood can be written as

$$\ln P = \frac{1}{2}\sum_{rs} e_{rs} \ln p_{rs} - n_r n_s p_{rs} - \sum_{i>j}\sum_{r\geq s} \ln A^{rs}_{ij}!. \tag{B2}$$

Maximizing $\ln P$ w.r.t. $p_{rs}$, we obtain $\hat{p}_{rs} = e_{rs}/n_r n_s$, and hence

$$\ln \hat{P} = -E + \frac{1}{2}\sum_{rs} e_{rs} \ln\left(\frac{e_{rs}}{n_r n_s}\right) - \sum_{i>j}\sum_{r\geq s} \ln A^{rs}_{ij}!. \tag{B3}$$

For simple graphs with $A^{rs}_{ij} \in \{0,1\}$, the last term in the above equation is equal zero, and we have that the approximation of the likelihood of the model with “hard” constraints in the sparse case is identical to the exact maximum likelihood of the Poisson model with “soft” constraints.

This model is similar to the popular mixed membership stochastic block model (MMSBM) [13], however it differs in the important aspect that it generates strictly denser overlaps. In the MMSBM, the existence of an edge $A_{ij}$ is sampled from a Bernoulli distribution with parameter $\lambda_{ij} = \sum_{rs} \theta^r_i \theta^s_j p_{rs}$, where $\theta^r_i$ is the probability that node $i$ belongs to group $r$, such that $\sum_r \theta^r_i = 1$, and $p_{rs} \in [0,1]$ is the probability that two nodes belonging to groups $r$ and $s$ are connected. Although for sparse graphs the difference between Poisson and Bernoulli models tend to disappear, with this parametrization the density of the overlaps are mixed with normalized weights. More specifically, for a node $i$ which belongs simultaneously to groups $r$ and $s$, its expected degree is equal to the weighted average of the unmixed degrees, $\left<k\right>_i = \theta^r_i \left<k\right>_r + \theta^s_i \left<k\right>_s$, where $\left<k\right>_r = \sum_s p_{rs} \sum_i \theta^s_i$ is the expected degree of a node that belongs only to group $r$. Thus, in the MMSBM the nodes in the mixture have an intermediate density between the sparser and the denser groups. In contrast, in the model considered in the main text, as well as the Poisson model above, we have simply $\left<k\right>_i = \left<k\right>_r + \left<k\right>_s$, and therefore the overlaps are always strictly denser than the pure groups. In this respect, it is equivalent to other formulations of the MMSBM, e.g. Refs. [81, 82].

### 2. Degree-corrected

A connection to a version of the model with “soft” constraints can also be made. We may consider each labelled entry $A^{rs}_{ij}$ in the adjacency matrix to be Poisson distributed with an average given by $\theta_i^r\theta_j^s\lambda_{rs}$,

$$P(G|\{\vec{b}_i\},\{\lambda_{rs}\},\{\theta_r\}) = \prod_{i>j}\prod_{r\geq s}(\theta_i^r\theta_j^s\lambda_{rs})^{A_{ij}^{rs}}e^{-\theta_i^r\theta_j^s\lambda_{rs}}/A_{ij}^{rs}!, \tag{B4}$$

where $A_{ij}^{rs}$ is the number of edges of type $(r,s)$ between nodes $i$ and $j$, and $\theta_i^r$ is the propensity with which a node receives an edge of type $r$. The log-likelihood can be written as

$$\ln P = \frac{1}{2}\sum_{rs}e_{rs}\ln\lambda_{rs} + \sum_{ir}k_i^r\ln\theta_i^r - \sum_{r\geq s}\lambda_{rs}\sum_{i>j}\theta_i^r\theta_j^s - \sum_{i>j}\sum_{r\geq s}\ln A_{ij}^{rs}!. \tag{B5}$$

Maximizing $\ln P$ w.r.t. $\{\lambda_{rs}\}$ and $\{\theta_i^r\}$, we obtain $\hat{\lambda}_{rs} = e_{rs}/e_re_s$ and $\hat{\theta}_i^r = k_i^r$, and hence

$$\ln\hat{P} = -E + \frac{1}{2}\sum_{rs}e_{rs}\ln\left(\frac{e_{rs}}{e_re_s}\right) + \sum_{ir}k_i^r\ln k_i^r - \sum_{i>j}\sum_{r\geq s}\ln A_{ij}^{rs}!. \tag{B6}$$

Again, for simple graphs with $A_{ij}^{rs}\in\{0,1\}$, the last term in the above equation is equal zero, however even in that case the likelihood is not identical to the version with "hard" constraints considered above, as is the case for the single membership version as well [31]. Both likelihoods only become the same in the limit $k_i^r \gg 1$ such that $\ln k_i^r! \simeq k_i^r\ln k_i^r - k_i^r$. Nevertheless, for the purpose of this paper, which is classification of empirical networks, the differences between these models can be overlooked.

There is a direct connection between this model and the one proposed by Ball et al [14]. In the not strictly assortative version of their model, the number of edges $A_{ij}$ is distributed according to a Poisson with average $\lambda_{ij} = \sum_{rs}\eta_i^r\eta_j^s\omega_{rs}$, where $\eta_i^r$ is the propensity with which node $i$ receives edges of type $r$ and $\omega_{rs}$ regulates the number of edges across groups. The total likelihood of that model is

$$P(G|\{\vec{b}_i\},\{\omega_{rs}\},\{\eta_r\}) = \prod_{i>j}\lambda_{ij}^{A_{ij}}e^{-\lambda_{ij}}/A_{ij}!. \tag{B7}$$

Since the sum of independent Poisson random variables is also distributed according to a Poisson, if we generate a graph with the model of Eq. B4 and observe only the total unlabelled edge counts $A_{ij} = \sum_{rs}A_{ij}^{rs}$, they are distributed exactly like Eq. B7, for the same choice of parameters $\theta_i^r = \eta_i^r$ and $\lambda_{rs} = \omega_{rs}$. Hence, the model of the main text is an equivalent formulation of the one in Ref. [14] where one keeps track of the latent variables specifying the exact type of each half-edge, instead of their marginal probability. This has the advantage that the maximum likelihood estimates for the model parameters $\lambda_{rs}$ and $\theta_i^r$ can be obtained directly by differentiation, and do not require iterations of an EM algorithm as in Ref. [14]. On the other hand we are left with the determination of labels in the half-edges, which is done with the method already described in Sec. VI.

## Appendix C: Maximum entropy ensemble of counts with constrained average

Suppose we want to compute the number of all possible non-negative integer counts $\{n_k\}$, subject to a normalization constraint $\sum_{k=0}^{\infty}n_k = N$ and a fixed average $\sum_{k=0}^{\infty}kn_k = E$. This can be obtained approximately, by relaxing the constraints so that they hold only on average. The maximum entropy ensemble given these constraints is the one with the probabilities $P(\{n_k\}) = e^{-H(\{n_k\})}/Z$, with $H(\{n_k\}) = \lambda\sum_k n_k + \mu\sum_k kn_k$, where $\lambda$ and $\mu$ are the Lagrange multipliers which keep the constraints in place. This is mathematically analogous to a simple Bose gas with energy levels given by $k$. The partition function is given by

$$Z = \sum_{\{n_k\}}e^{-\lambda\sum_k n_k - \mu\sum_k kn_k} = \prod_k Z_k, \tag{C1}$$

with

$$Z_k = \left[1 - e^{-\lambda-\mu k}\right]^{-1}. \tag{C2}$$

The average counts are given by $\langle n_k\rangle = -\partial Z_k/\partial\lambda = [\exp(\lambda+\mu k)-1]^{-1}$, and the parameters $\lambda$ and $\mu$ are determined via the imposed constraints,

$$\sum_{k=0}^{\infty}[\exp(\lambda+\mu k)-1]^{-1} = N, \tag{C3}$$

$$\sum_{k=0}^{\infty}k\,[\exp(\lambda+\mu k)-1]^{-1} = E. \tag{C4}$$

Further analytical progress can be made by replacing the sums with integrals, and using the Polylogarithm function and its connection with the Bose–Einstein distribution, $\mathrm{Li}_s(z) = \Gamma(s)^{-1}\int_0^{\infty}\frac{t^{s-1}}{e^t/z-1}\mathrm{d}t$,

$$\int_0^{\infty}dk\,[\exp(\lambda+\mu k)-1]^{-1} = \frac{\mathrm{Li}_1(e^{-\lambda})}{\mu} = N, \tag{C5}$$

$$\int_0^{\infty}dkk\,[\exp(\lambda+\mu k)-1]^{-1} = \frac{\mathrm{Li}_2(e^{-\lambda})}{\mu^2} = E. \tag{C6}$$

Eq. C5 can be inverted as $e^{-\lambda} = 1-\exp(-N/\mu)$, but Eq. C6 cannot be solved for $\lambda$ in closed form. However, by assuming a sufficiently "high temperature" regime where $\mu\sim O(1)$, we have that the fugacity simplifies in the thermodynamic limit, $e^{-\lambda}\to 1$ for $N\gg 1$, and hence we obtain $\mu\simeq\sqrt{\mathrm{Li}_2(1)/E}$. Using Eqs. C5 and C6, we can write the entropy of the ensemble $\ln\Xi = -\sum_k[\partial\ln Z_k/\partial\lambda + \partial\ln Z_k/\partial\mu + \ln Z_k]$, as

$$\ln\Xi = \lambda N + 2\mu E, \tag{C7}$$

and for the regime $e^{-\lambda} \to 1$, we have

$$\ln \Xi \simeq 2\sqrt{\zeta(2)E}, \tag{C8}$$

where the identity $\mathrm{Li}_2(1) = \zeta(2)$ was used. Although Eq. C8 becomes asymptotically exact in the thermodynamic limit with $E \sim N$ and $N \gg 1$, the exact solution can also be obtained with arbitrary precision simply by iterating Eqs. C5 and C6 as $\hat{\lambda}(t+1) = 1-\exp(-N/\mu(t))$, $\mu(t+1) = \sqrt{E/\,\mathrm{Li}_2(\hat{\lambda}(t))}$, where $\hat{\lambda} \equiv e^{-\lambda}$, with the starting points $\hat{\lambda}(0) = 1$, $\mu(0) = \sqrt{\mathrm{Li}_2(1)/E}$, until sufficient convergence is reached, and the results are substituted in Eq. C7. (We actually use this more precise procedure when computing Eq. 8 in the main text, throughout the analysis.)

[1] M. E. J. Newman, Nat Phys **8**, 25 (2011).
[2] S. Fortunato, Physics Reports **486**, 75 (2010).
[3] P. Holme, Physical Review E **72**, 046111 (2005).
[4] M. P. Rombach, M. A. Porter, J. H. Fowler, and P. J. Mucha, arXiv:1202.2684 (2012).
[5] D. B. Larremore, A. Clauset, and A. Z. Jacobs, Physical Review E **90**, 012805 (2014).
[6] A. Clauset, C. Moore, and M. E. J. Newman, Nature **453**, 98 (2008).
[7] T. P. Peixoto, Physical Review X **4**, 011047 (2014).
[8] R. Guimerà and M. Sales-Pardo, Proceedings of the National Academy of Sciences **106**, 22073 (2009).
[9] S. V. Buldyrev, R. Parshani, G. Paul, H. E. Stanley, and S. Havlin, Nature **464**, 1025 (2010).
[10] A. Apolloni, C. Poletto, J. J. Ramasco, P. Jensen, and V. Colizza, Theoretical Biology and Medical Modelling **11**, 3 (2014).
[11] R. Guimerà and L. A. Nunes Amaral, Nature **433**, 895 (2005).
[12] B. Karrer and M. E. J. Newman, Physical Review E **83**, 016107 (2011).
[13] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing, J. Mach. Learn. Res. **9**, 1981 (2008).
[14] B. Ball, B. Karrer, and M. E. J. Newman, Physical Review E **84**, 036103 (2011).
[15] G. Palla, L. Lovász, and T. Vicsek, Proceedings of the National Academy of Sciences **107**, 7640 (2010).
[16] J. Leskovec, D. Chakrabarti, J. Kleinberg, C. Faloutsos, and Z. Ghahramani, J. Mach. Learn. Res. **11**, 985 (2010).
[17] M. Mariadassou, S. Robin, and C. Vacher, The Annals of Applied Statistics **4**, 715 (2010), mathematical Reviews number (MathSciNet): MR2758646.
[18] C. Aicher, A. Z. Jacobs, and A. Clauset, Journal of Complex Networks , cnu026 (2014).
[19] B. Ball and M. Newman, Network Science **1**, 16 (2013).
[20] M. Kivelä, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter, Journal of Complex Networks **2**, 203 (2014).
[21] W. Fu, L. Song, and E. P. Xing, in *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09 (ACM, New York, NY, USA, 2009) pp. 329–336.
[22] X. Yan, C. Shalizi, J. E. Jensen, F. Krzakala, C. Moore, L. Zdeborová, P. Zhang, and Y. Zhu, Journal of Statistical Mechanics: Theory and Experiment **2014**, P05007 (2014).
[23] M. E. J. Newman, Proceedings of the National Academy of Sciences **103**, 8577 (2006).
[24] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann, Nature **466**, 761 (2010).
[25] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, Nature **435**, 814 (2005).
[26] M. E. J. Newman, EPL (Europhysics Letters) **103**, 28003 (2013).
[27] M. E. J. Newman, Physical Review E **88**, 042822 (2013).
[28] R. R. Nadakuditi and M. E. J. Newman, Physical Review Letters **108**, 188701 (2012).
[29] F. Krzakala, C. Moore, E. Mossel, J. Neeman, A. Sly, L. Zdeborová, and P. Zhang, Proceedings of the National Academy of Sciences , 201312486 (2013).
[30] G. Bianconi, Physical Review E **79**, 036114 (2009).
[31] T. P. Peixoto, Physical Review E **85**, 056122 (2012).
[32] P. D. Grünwald, *The Minimum Description Length Principle* (The MIT Press, 2007).
[33] J. Rissanen, *Information and Complexity in Statistical Modeling*, 1st ed. (Springer, 2010).
[34] M. Rosvall and C. T. Bergstrom, Proceedings of the National Academy of Sciences **104**, 7327 (2007).
[35] P. W. Holland, K. B. Laskey, and S. Leinhardt, Social Networks **5**, 109 (1983).
[36] S. E. Fienberg, M. M. Meyer, and S. S. Wasserman, Journal of the American Statistical Association **80**, 51 (1985).
[37] K. Faust and S. Wasserman, Social Networks **14**, 5 (1992).
[38] C. J. Anderson, S. Wasserman, and K. Faust, Social Networks **14**, 137 (1992).
[39] P. Latouche, E. Birmelé, and C. Ambroise, Electronic Journal of Statistics **8**, 762 (2014).
[40] T. P. Peixoto, Physical Review Letters **110**, 148701 (2013).
[41] S. H. Jeffreys, *The Theory of Probability* (Oxford University Press, 1998).
[42] L. A. Adamic and N. Glance, in *Proceedings of the 3rd international workshop on Link discovery*, LinkKDD '05 (ACM, New York, NY, USA, 2005) pp. 36–43.
[43] D. Holten, IEEE Transactions on Visualization and Computer Graphics **12**, 741 (2006).
[44] J. Mcauley and J. Leskovec, ACM Trans. Knowl. Discov. Data **8**, 4:1 (2014).
[45] D. E. Knuth, *The Stanford GraphBase: A Platform for Combinatorial Computing*, 1st ed. (Addison-Wesley Professional, New York, N.Y. : Reading, Mass, 1993).
[46] M. Girvan and M. E. J. Newman, Proceedings of the National Academy of Sciences **99**, 7821 (2002).
[47] T. S. Evans, FigShare (2012), 10.6084/m9.figshare.93179.
[48] W. W. Zachary, Journal of Anthropological Research **33**, 452 (1977).
[49] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney, arXiv:0810.1355 (2008).

[50] B. Klimt and Y. Yang, in *CEAS* (2004).

[51] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, Behavioral Ecology and Sociobiology **54**, 396 (2003).

[52] O. Richters and T. P. Peixoto, PLoS ONE **6**, e18384 (2011).

[53] E. Cho, S. A. Myers, and J. Leskovec, in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11 (ACM, New York, NY, USA, 2011) pp. 1082–1090.

[54] D. J. Watts and S. H. Strogatz, Nature **393**, 409 (1998).

[55] J. Leskovec, J. Kleinberg, and C. Faloutsos, ACM Trans. Knowl. Discov. Data **1** (2007), 10.1145/1217299.1217301.

[56] M. E. J. Newman, Physical Review E **74**, 036104 (2006).

[57] M. Richardson, R. Agrawal, and P. Domingos, in *The Semantic Web - ISWC 2003*, Lecture Notes in Computer Science No. 2870, edited by D. Fensel, K. Sycara, and J. Mylopoulos (Springer Berlin Heidelberg, 2003) pp. 351–368.

[58] K. I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A. L. Barabási, Proceedings of the National Academy of Sciences **104**, 8685 (2007).

[59] H. Yu, P. Braun, M. A. Yıldırım, I. Lemmens, K. Venkatesan, J. Sahalie, T. Hirozane-Kishikawa, F. Gebreab, N. Li, N. Simonis, T. Hao, J.-F. Rual, A. Dricot, A. Vazquez, R. R. Murray, C. Simon, L. Tardivo, S. Tam, N. Svrzikapa, C. Fan, A.-S. d. Smet, A. Motyl, M. E. Hudson, J. Park, X. Xin, M. E. Cusick, T. Moore, C. Boone, M. Snyder, F. P. Roth, A.-L. Barabási, J. Tavernier, D. E. Hill, and M. Vidal, Science **322**, 104 (2008).

[60] T. Reguly, A. Breitkreutz, L. Boucher, B.-J. Breitkreutz, G. C. Hon, C. L. Myers, A. Parsons, H. Friesen, R. Oughtred, A. Tong, C. Stark, Y. Ho, D. Botstein, B. Andrews, C. Boone, O. G. Troyanskya, T. Ideker, K. Dolinski, N. N. Batada, and M. Tyers, Journal of Biology **5**, 11 (2006).

[61] S. R. Collins, P. Kemmeren, X.-C. Zhao, J. F. Greenblatt, F. Spencer, F. C. P. Holstege, J. S. Weissman, and N. J. Krogan, Molecular & cellular proteomics: MCP **6**, 439 (2007).

[62] H. Salgado, M. Peralta-Gil, S. Gama-Castro, A. Santos-Zavaleta, L. Muñiz-Rascado, J. S. García-Sotelo, V. Weiss, H. Solano-Lira, I. Martínez-Flores, A. Medina-Rivera, G. Salgado-Osorio, S. Alquicira-Hernández, K. Alquicira-Hernández, A. López-Fuentes, L. Porrón-Sotelo, A. M. Huerta, C. Bonavides-Martínez, Y. I. Balderas-Martínez, L. Pannier, M. Olvera, A. Labastida, V. Jiménez-Jacinto, L. Vega-Alvarado, V. Del Moral-Chávez, A. Hernández-Alvarez, E. Morett, and J. Collado-Vides, Nucleic Acids Research **41**, D203 (2013).

[63] J. McAuley and J. Leskovec, in *Computer Vision – ECCV 2012*, Lecture Notes in Computer Science No. 7575, edited by A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid (Springer Berlin Heidelberg, 2012) pp. 828–841.

[64] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney, Internet Mathematics **6**, 29 (2009).

[65] J. Yang and J. Leskovec, in *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, MDS '12 (ACM, New York, NY, USA, 2012) pp. 3:1–3:8.

[66] R. Albert, H. Jeong, and A.-L. Barabási, Nature **401**, 130 (1999).

[67] J. Leskovec, D. Huttenlocher, and J. Kleinberg, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10 (ACM, New York, NY, USA, 2010) pp. 1361–1370.

[68] J. Leskovec, D. Huttenlocher, and J. Kleinberg, in *Proceedings of the 19th International Conference on World Wide Web*, WWW '10 (ACM, New York, NY, USA, 2010) pp. 641–650.

[69] T. S. K. Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D. S. Somanathan, A. Sebastian, S. Rani, S. Ray, C. J. H. Kishore, S. Kanth, M. Ahmed, M. K. Kashyap, R. Mohmood, Y. L. Ramachandra, V. Krishna, B. A. Rahiman, S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady, and A. Pandey, Nucleic Acids Research **37**, D767 (2009).

[70] M. E. J. Newman, Physical Review E **69**, 066133 (2004).

[71] A. Clauset, M. E. J. Newman, and C. Moore, Physical Review E **70**, 066111 (2004).

[72] A. Arenas, L. Danon, A. Díaz-Guilera, P. M. Gleiser, and R. Guimerá, The European Physical Journal B - Condensed Matter and Complex Systems **38**, 373 (2004).

[73] A. Lancichinetti, S. Fortunato, and J. Kertész, New Journal of Physics **11**, 033015 (2009).

[74] T. P. Peixoto, Physical Review E **89**, 012804 (2014).

[75] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, The Journal of Chemical Physics **21**, 1087 (1953).

[76] W. K. Hastings, Biometrika **57**, 97 (1970).

[77] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, Physical Review E **84**, 066106 (2011).

[78] P. K. Gopalan and D. M. Blei, Proceedings of the National Academy of Sciences **110**, 14534 (2013).

[79] T. P. Peixoto, figshare (2014), 10.6084/m9.figshare.1164194.

[80] P. D. Hoff, A. E. Raftery, and M. S. Handcock, Journal of the American Statistical Association **97**, 1090 (2002).

[81] J. Parkkinen, J. Sinkkonen, A. Gyenge, and S. Kaski, in *Proceedings of the 7th International Workshop on Mining and Learning with Graphs (MLG 2009), Leuven* (2009).

[82] J. Yang and J. Leskovec, in *2012 IEEE 12th International Conference on Data Mining (ICDM)* (2012) pp. 1170–1175.